Available online at www.sciencedirect.com
ScienceDirect
Journal of Statistical Planning and Inference 138 (2008) 3022–3038
journal of statistical planning and inference
www.elsevier.com/locate/jspi

# On the coefficient of determination for mixed regression models

Ola Hössjer

*Department of Mathematics, Stockholm University, S-106 91 Stockholm, Sweden*

Received 8 February 2007; received in revised form 3 October 2007; accepted 20 November 2007
Available online 4 December 2007

**Abstract**

For mixed regression models, we define a variance decomposition including three terms, explained individual variance, unexplained individual variance and noise variance. In contrast to traditional variance decomposition, it is thus the *unexplained*, not the explained, variance that is split. It gives rise to a coefficient of individual determination (CID) defined as the estimated fraction of explained individual variance. We argue that in many applications CID is a valuable complement to $R^2$, since it excludes noise variance (which can never be explained) and thus has one as a natural upper bound.

A general theory for coefficients determination is presented, including various choices of regression models, weight functions and parameter estimates. In particular we focus on models where CID is computable, such as univariate mixed Poisson and logistic regression models, as well as multivariate mixed linear regression models. Large sample properties and confidence intervals are derived and finally, the theory is exemplified using Poisson regression on a Swedish motor traffic insurance data set.


*Keywords:* Coefficient of determination; Explained variance; Individual variance; Mixed regression models; Noise variance; Variance decomposition

## 1. Introduction

The coefficient of determination, $R^2$, is a well known quantity in univariate multiple linear regression, quantifying the proportion of variation of the response variables $Y_i$ explained by the non-intercept covariates. Writing $E(Y_i|x_i) = m_i = \beta x_i^{\mathrm{T}}$, where $x_i$ is the $1 \times p$ vector of covariates for individual $i$ and $\beta$ the $1 \times p$ vector of regression coefficients,

$$R^2 = \frac{\sum_{i=1}^n (\hat{m}_i - \hat{m})^2}{\sum_{i=1}^n (Y_i - \hat{m})^2}, \tag{1}$$

where $\hat{m}_i = \hat{\beta} x_i^{\mathrm{T}}$, $\hat{\beta}$ is the least square estimate of $\beta$ and $\hat{m} = \sum_i Y_i/n = \sum_i \hat{m}_i/n$.

The purpose of this paper is to define coefficients of determination for univariate mixed regression models including possible nonlinearities and heteroscedasticity. The main idea is to decompose the variance of *untransformed* data into three parts: variance explained by the covariates, individual variance not explained by the covariates due to random effects and remaining noise variance. This differs from the usual approach where the *explained* variance is divided into various components, whereas we split the *unexplained* variance. As a result, we get two different coefficients of

*E-mail address:* ola@math.su.se.

0378-3758/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.jspi.2007.11.010

determination. The first one, $R^2$, is the estimated proportion of variance explained by the covariates, in agreement with (1). The second one, the coefficient of individual determination (CID), is the estimated proportion of *individual* variance explained by the covariates. We argue that in many applications CID is a valuable complement to $R^2$, having one as a natural upper bound when all individual variation are explained.

Our main focus is a class of univariate regression models for which noise variance is a deterministic function of the regression parameters and covariates. This class includes Poisson regression and binomial regression with overdispersion, as well as linear regression with known noise variance. In these cases, both $R^2$ and CID can be computed. On the other hand, CID is not computable for linear regression with unknown noise variance. The reason is that unexplained individual variance and noise variance cannot be separated. This is in contrast to multivariate linear regression, where CID is computable even with unknown noise variance, by using the correlation structure of the residuals from a (weighted) least squares fit.

There are alternatives to decomposing untransformed variance, such as (i) computing $R^2$ as the proportion of explained variance for transformed data and (ii) defining $R^2$ generally within a likelihood framework.

The transformation approach (i) has several advantages. For models without random components, one may use a variance stabilizing transformation and estimate parameters by least squares, see for instance Cochran (1940) and Carroll and Ruppert (1988). When random effects are included, as in the generalized linear mixed model (Breslow and Clayton, 1993), one may approximate transformed data by a linear mixed model, see for instance Lee and Chaubey (1996). The likelihood approach (ii) is appealing because of its general interpretation of $1 - R^2$ as the proportion of unexplained variation, see Maddala (1983), Cox and Snell (1989, pp. 208–209), Magee (1990) and Nagelkerke (1991).

On the other hand, in many cases, the untransformed response variable is the quantity of major interest to the experimenter. It may represent a count of accidents, a proportion of successful experiments or an economic cost. Then a coefficient of determination based on untransformed data is more useful than (i)–(ii).

The paper is organized as follows: In Sections 2 and 3 we define the class of regression models and coefficients of determination. Weighting of observations is considered in Section 4 and parameter estimation in Section 5. In Section 6 we give several examples of models and in Section 7 we consider multivariate extensions. Large sample properties of $R^2$ and CID are derived in Section 8 and a Swedish car accident data set is analyzed in Section 9. Section 10 contains a discussion and, finally, proofs and technical results are collected in the appendix.

## 2. Variance decomposition of the response variable

Consider a collection $(x_1, t_1, Y_1), \ldots, (x_n, t_n, Y_n)$ of observations, with a scalar response variable, $Y_i$, a $1 \times p$ vector of explanatory variables, $x_i$, and a known constant for the $i$th individual or cell, $t_i$. For instance, $t_i$ may represent the time of exposure or area of data collection in Poisson regression or the number of trials in binomial regression.

We assume that $\{Y_i\}$ are conditionally independent given $\{x_i\}$ and $\{t_i\}$ with a mean function

$$m(\beta x^{\mathrm{T}}, t) = E(Y|x, t) \tag{2}$$

depending on a $1 \times p$ vector of unknown regression parameters $\beta$.

We will focus on mixed regression models, containing a hidden liability $L$ such that, $E(Y|L, x, t) = M(\eta, t)$, with $\eta = \beta x^{\mathrm{T}}$. We regard $L$ as quantity that is only partially explained by the covariates. It summarizes the individual characteristic of each observation, whereas the distribution of $Y|L, x, t$ is due noise that bears no information about the individual. By integrating out $L$ in the definition of $M$ we get $E(M(\eta, t)) = m(\eta, t)$ and

$$\begin{aligned} v(\eta, \xi, t) &= \mathrm{Var}(Y|x, t) \\ &= E(\mathrm{Var}(Y|L, x, t)) + \mathrm{Var}(M(\eta, t)) \\ &=: v_1(\eta, t) + v_2(\eta, \xi, t). \end{aligned} \tag{3}$$

In (3), $\xi$ is an unknown scalar variance parameter such that $\xi = 0$ implies a deterministic liability, i.e. $v_2(\eta, 0, t) = 0$. In this case, $\mathrm{Var}(Y|x, t) = v_1(\beta x^{\mathrm{T}}, t)$, which is assumed to be a known function of the regression parameters. This crucial property makes $v_1$ identifiable, since $\beta$ can be estimated from data.

Let $w(\eta, \xi, t)$ be a given weight function. Consider an individual $I \in \{1, \ldots, n\}$ drawn at random from the sample with probabilities

$$P(I = i) = w_i \Big/ \sum_j w_j,$$

where $w_i = w(\eta_i, \xi, t_i)$ and $\eta_i = \beta x_i^{\mathrm{T}}$. The drawn individual has mean response

$$m = E(Y_I | \{x_i, t_i\}_{i=1}^n) = \sum_i w_i m_i \Big/ \sum_i w_i, \tag{4}$$

where $m_i = m(\eta_i, t_i)$, and variance $\sigma^2 = \mathrm{Var}(Y_I | \{x_i, t_i\}_{i=1}^n)$. The variance can be decomposed into three terms

$$\begin{aligned}\sigma^2 &= \sum_i w_i E((Y_i - m)^2 | x_i, t_i) \Big/ \sum_i w_i \\ &= \left( \sum_i w_i (m_i - m)^2 + \sum_i w_i (v_i - v_{1i}) + \sum_i w_i v_{1i} \right) \Big/ \sum_i w_i \\ &=: \sigma_1^2 + \sigma_2^2 + \sigma_3^2, \end{aligned} \tag{5}$$

where $v_i = v(\eta_i, \xi, t_i)$ and $v_{1i} = v_1(\eta_i, t_i)$. The first term, $\sigma_1^2$, is the variance component explained by the covariates $x$, the second term, $\sigma_2^2$, the variance of the liability, and the third term, $\sigma_3^2$, the remaining variance. We interpret $\sigma_2^2$ as variance due to unobserved individual characteristics, whereas $\sigma_3^2$ represents noise variance. Their sum, $\sigma_{\mathrm{unexp}}^2 = \sigma_2^2 + \sigma_3^2$, is the variance not explained by the covariates.

To quantify how large a proportion of the variance is explained by the covariates, we use either

$$\rho = \sigma_1^2 / \sigma^2 \tag{6}$$

or

$$\rho_{\mathrm{ind}} = \sigma_1^2 / \sigma_{\mathrm{ind}}^2, \tag{7}$$

where $\sigma_{\mathrm{ind}}^2 = \sigma_1^2 + \sigma_2^2$ is the total variance due to individual variation. Hence, $\rho$ gives the fraction of the *total* variance, whereas $\rho_{\mathrm{ind}}$ gives the fraction of the *individual* variance explained by the covariates. For this reason, $\rho_{\mathrm{ind}}$ is often more interesting, with 1 as a natural upper bound corresponding to all individual variation being explained by $\{x_i\}$ and $\{t_i\}$. On the other hand, $0 \leqslant \rho \leqslant 1 - \sigma_3^2/\sigma^2$, since the noise variance component $\sigma_3^2$ cannot be explained by the covariates.

## 3. Coefficients of determination

With parameter estimates $\hat{\theta} = (\hat{\beta}, \hat{\xi})$, we let $\hat{w}_i = w(\hat{\eta}_i, \hat{\xi}, t_i)$, $\hat{\eta}_i = x_i \hat{\beta}^{\mathrm{T}}$, $\hat{m}_i = m(\hat{\eta}_i, t_i)$, $\hat{v}_i = v(\hat{\eta}_i, \hat{\xi}, t_i)$ and $\hat{v}_{1i} = v_i(\hat{\eta}_i, t_i)$. Then the mean response of a randomly picked individual is estimated by

$$\hat{m} = \sum_i \hat{w}_i \hat{m}_i \Big/ \sum_i \hat{w}_i$$

and the estimated variance decomposition is

$$\hat{\sigma}^2 = \hat{\sigma}_1^2 + \hat{\sigma}_2^2 + \hat{\sigma}_3^2, \tag{8}$$

where $\hat{\sigma}_1^2 = \sum_i \hat{w}_i (\hat{m}_i - \hat{m})^2 \big/ \sum_i \hat{w}_i$ and $\hat{\sigma}_3^2 = \sum_i \hat{w}_i \hat{v}_{1i} \big/ \sum_i \hat{w}_i$. There are at least two ways of defining $\hat{\sigma}_2^2$, either

$$\hat{\sigma}_2^2 = \hat{\sigma}_{\mathrm{unexp}}^2 - \hat{\sigma}_3^2 = \sum_i \hat{w}_i (Y_i - \hat{m}_i)^2 \Big/ \sum_i \hat{w}_i - \hat{\sigma}_3^2 \tag{9}$$

or

$$\hat{\sigma}_2^2 = \sum_i \hat{w}_i(\hat{v}_i - \hat{v}_{1i}) \Big/ \sum_i \hat{w}_i. \tag{10}$$

Whereas (10) requires specification of a correct variance function $v_2$, (9) only uses estimates of $\beta$, not $\xi$ (as long as $\hat{w}_i$ does not involve $\hat{\xi}$). For this reason, we will use (9) in the sequel. In either case, $\hat{\sigma}^2$ need not equal $\sum_i \hat{w}_i(Y_i - \hat{m})^2 \big/ \sum_i \hat{w}_i$.

Replacing the theoretical variance components by their estimates in (6) and (7), we get two versions of the coefficient of determination

$$R^2 = \hat{\rho} = \frac{\hat{\sigma}_1^2}{\hat{\sigma}^2} = \frac{\sum_i \hat{w}_i(\hat{m}_i - \hat{m})^2}{\sum_i \hat{w}_i((\hat{m}_i - \hat{m})^2 + (Y_i - \hat{m}_i)^2)} \tag{11}$$

and

$$\mathrm{CID} = \hat{\rho}_{\mathrm{ind}} = \frac{\hat{\sigma}_1^2}{\hat{\sigma}_{\mathrm{ind}}^2} = \frac{\sum_i \hat{w}_i(\hat{m}_i - \hat{m})^2}{\sum_i \hat{w}_i((\hat{m}_i - \hat{m})^2 + (Y_i - \hat{m}_i)^2 - \hat{v}_{1i})}. \tag{12}$$

In order to compute CID, we utilize that all $v_{1i}$ are functions of no other unknown quantities than $\beta$, which is estimable, so that all $\hat{v}_{1i}$ can be computed. This is not always guaranteed, as in the case for univariate linear regression (see Example 3 of Section 6).

## 4. Choice of weight functions

The weight function $w$ may be chosen in several ways. Some important special cases are:

(i) *Uniform weights*: Let $w(\eta, \xi, t) \equiv 1$, so that $w_i = \hat{w}_i = 1$. In this case each individual contributes an equal amount to the variance decomposition.

(ii) *Inverse variance*: If $w(\eta, \xi, t) = v(\eta, \xi, t)^{-1}$, each individual contributes an amount inversely proportional to its variance. Then the estimated unexplained variation

$$\hat{\sigma}_{\mathrm{unexp}}^2 = \hat{\sigma}_2^2 + \hat{\sigma}_3^2 = \sum_i \frac{(Y_i - \hat{m}_i)^2}{n\hat{v}_i}$$

equals Pearson's unscaled $\chi^2$ statistic (McCullagh and Nelder, 1989) apart from the term $n$ of the denominator. For linear regression (see Example 3), $R^2$ based on inverse variance weighting is equivalent to the coefficient of determination for generalized least squares (Buse, 1973, 1979).

(iii) *Inverse non-dispersed variance*: A simplified form of inverse variance weighting, $w(\eta, \xi, t) = v_1(\eta, t)^{-1}$, employs the variance function of the non-dispersed model $\xi = 0$. Then $\hat{w}_i$ does not involve $\hat{\xi}$, and no explicit model of overdispersion is needed to define $R^2$ and CID.

(iv) *Deterministic weights*: If the weight function $w(\eta, \xi, t) = a(t)$ is a known function, we need not estimate the weights, i.e. $\hat{w}_i = w_i$. A typical case is when the non-dispersed variance factorizes as $v_1(\eta, t) = b(\eta)/a(t)$. The deterministic weight function is then a simplified form of (iii), obtained by dropping the term $b(\eta)$. Typical choices are models for Poisson rates and binomial proportions (see Section 6), with $a(t) = t$.

## 5. Examples of parameter estimates

Regarding choice of parameter estimates, we mention three possibilities:

(i) *Weighted nonlinear least squares*: In case the weights $w_i$ do not depend on unknown parameters, we may first estimate $\beta$ by

$$\hat{\beta} = \arg\min_{\beta} \sum_i w_i(Y_i - m(\beta x_i^{\mathrm{T}}, t_i))^2 \tag{13}$$

and then $\xi$ from the residuals $Y_i - \hat{m}_i$ by solving

$$\sum_i w_i v(\hat{\eta}_i, \hat{\xi}, t_i) = \sum_i w_i (Y_i - \hat{m}_i)^2. \tag{14}$$

(ii) *Maximum likelihood*: Let $f(y|\eta, \xi, t)$ denote the density of $Y$ given $\eta$, $\xi$ and $t$, and let us estimate all parameters simultaneously from

$$(\hat{\beta}, \hat{\xi}) = \arg\max_{\beta,\xi} l(\beta, \xi),$$

where $l(\beta, \xi) = \sum_i \log f(Y_i|\beta x_i^{\mathrm{T}}, \xi, t_i)$ is the log likelihood function. A simplified version is to estimate the regression parameters by ML from the deterministic liability model

$$\hat{\beta} = \arg\max_{\beta} l(\beta, 0) \tag{15}$$

and then estimate $\xi$ separately from the residuals, using e.g. (14). Even though (15) uses a misspecified model when $\xi \neq 0$, $\hat{\beta}$ is still consistent as long as the estimating equation is unbiased at the true parameter value, see e.g. White (1982) and Cox (1983).

(iii) *Extended quasi-likelihood*: Estimate $\theta$ by solving the $p + 1$ estimating equations $U_\beta(\hat{\theta}) = 0$ and $U_\xi(\hat{\theta}) = 0$ simultaneously, where

$$U_\beta(\theta) = \sum_i x_i m_i' \frac{Y_i - m_i}{v_i}, \tag{16}$$

$m_i' = \partial m(\eta_i, t_i)/\partial \eta_i$ and

$$U_\xi(\theta) = \sum_i \left( \frac{(Y_i - m_i)^2}{v_i} - 1 \right).$$

This is an extension of the quasi-likelihood estimate $U_\beta(\hat{\beta}, \xi) = 0$ of $\beta$, defined for fixed $\xi$, see Wedderburn (1974), McCullagh (1983) and Moore (1986). A simplified version of extended QL is to estimate $\beta$ separately from the deterministic liability model, replacing $v_i$ by $v_{1i}$ in (16) and then estimating $\xi$.

Of (i)–(iii), the maximum likelihood estimates are the most efficient but in general also most sensitive to model misspecification. The quasi-likelihood estimate of $\beta$ often coincides with the ML-estimate when $\xi = 0$.

## 6. Examples of models

**Example 1** (*Overdispersed Poisson regression*). Assume that a certain characteristic (car accidents, disease incidences, etc.) occurs according to a Poisson process with individual specific rate $L = \Lambda$. If this process is observed during $t$ time units, the observed rate of incidences is

$$Y|\Lambda, t \in \mathrm{Po}(\Lambda t)/t. \tag{17}$$

The log link corresponds to a mean rate parameter $E(\Lambda|x) = \lambda(x) = \exp(\eta)$, where $\eta = \beta x^{\mathrm{T}}$. We assume $\mathrm{Var}(\Lambda|x) = \xi\lambda(x)^a$ for some (known) constant $a > 0$, so that

$$\begin{aligned} m(\eta, t) &= \exp(\eta), \\ v(\eta, \xi, t) &= \exp(\eta) t^{-1} + \xi \exp(a\eta). \end{aligned} \tag{18}$$

Parameter estimates and data analysis for this model have been carried out by several authors, e.g. Pocock et al. (1981), Hinde (1982), Breslow (1984), Lawless (1987) and references therein. An important special case of (17) is when $\Lambda$ has a gamma distribution. Then, the mixed Poisson distribution of $tY|x, t$ is negative binomial. The choice $a = 2$ in (18)

is appealing since $\xi$ is the squared coefficient of variation of $\Lambda|x$ for all $x$. Alternatively, $a = 1$ is frequently used in connection with generalized linear models.

**Example 2** (*Overdispersed logistic regression*). Let $Y$ represent the relative number of successes in $t$ trials with an individual specific success probability $L = \Pi$, i.e.

$$Y|\Pi, t \in \text{Bin}(t, \Pi)/t.$$

Using a logistic link function, $E(\Pi|x) = \pi(x) = 1/(1 + \exp(-\eta))$. If the variance function of $\Pi$ satisfies $\text{Var}(\Pi|x) = \xi\pi(x)(1 - \pi(x))$ we get

$$m(\eta, t) = 1/(1 + \exp(-\eta)),$$

$$v(\eta, \xi, t) = t^{-1}(1 + \xi(t - 1)) \exp(-\eta)/(1 + \exp(-\eta))^2.$$

Parameter estimation for this model has been considered, for instance, by Williams (1982). An important special case, the beta-binomial model, occurs when $\Pi$ has a beta distribution, see e.g. Williams (1975) and Crowder (1978). A review of analysis methods for data with extra-binomial variation is provided by Haseman and Kupper (1978).

**Example 3** (*Univariate linear regression*). Consider a homoscedastic linear model

$$Y = \beta x^{\mathrm{T}} + \delta + \varepsilon \tag{19}$$

with $L = \delta$, an individual specific zero mean random variable, and $\varepsilon$, an independent zero mean noise term. Put $\xi = \text{Var}(\delta)$ and $\tau = \text{Var}(\varepsilon)$. If $\tau$ is a known constant, (19) is a special case of (2)–(3) with $t \equiv 1$ and

$$m(\beta x^{\mathrm{T}}) = \beta x^{\mathrm{T}},$$

$$v_1 = \tau,$$

$$v_2(\xi) = \xi.$$

In the more realistic case of unknown $\tau$, the parameter vector is

$$\theta = (\beta, \tau, \xi). \tag{20}$$

Since the random terms $\delta$ and $\varepsilon$ in (19) cannot be separated, we cannot estimate $\sigma_2^2 = \tau$ and $\sigma_3^2 = \xi$ from data, but only $\sigma^2_{\text{unexp}} = \tau + \xi$. Hence, the last two variance components of (5) are not estimable and only $R^2$, not CID, is computable.

When weighted least squares (13) with known weights ($\hat{w}_i = w_i$) are used

$$\hat{m} = \sum_i w_i Y_i \Big/ \sum_i w_i,$$

$$\hat{\sigma}^2 = \hat{\sigma}_1^2 + \hat{\sigma}^2_{\text{unexp}} = \sum_i w_i (Y_i - \hat{m})^2 \Big/ \sum_i w_i$$

by well known orthogonality properties of WLS. In particular, for uniform weights, $R^2$ agrees with (1).

## 7. Multivariate response variables

Consider the multivariate extension of (19) where $Y$ is a $1 \times q$ vector for some $q > 1$. Then $L = \delta$ is a $1 \times q$ vector of individual effects, $\varepsilon$ a $1 \times q$ vector of error terms, $\beta$ a $1 \times p$ vector, and $x$ a $q \times p$ matrix. This yields

$$E(Y|x) = \beta x^{\mathrm{T}},$$

$$\text{Cov}(Y|x, t) = \xi C(t) + \tau I_q = V(t, \xi, \tau), \tag{21}$$

where $\xi$ is the variance of the random effects components, $\tau$ the variance of the error components, $C(t) = \text{Corr}(\delta|t)$ the correlation matrix of the random effects, and $I_q$ the identity matrix of order $q$. See Lynch and Walsh (1998) for some examples.

Consider a sample of $n$ cells, let $m_i = \beta x_i^{\mathrm{T}}$ $C_i = C(t_i)$ and $W_i = W(\beta x_i^{\mathrm{T}}, \tau, \xi, t_i)$ be a symmetric and positive definite weight matrix. Then we may generalize (4)–(5) to

$$m = \sum_i m_i W_i 1_q^{\mathrm{T}} \Big/ \sum_i 1_q W_i 1_q^{\mathrm{T}},$$

$$\sigma_1^2 = \sum_i (m_i - m) W_i (m_i - m)^{\mathrm{T}} \Big/ \sum_i \mathrm{tr}(W_i),$$

$$\sigma_2^2 = \xi \sum_i \mathrm{tr}(C_i W_i) \Big/ \sum_i \mathrm{tr}(W_i),$$

$$\sigma_3^2 = \tau \sum_i \mathrm{tr}(I_q W_i) \Big/ \sum_i \mathrm{tr}(W_i) = \tau$$

and then define $\rho$ and $\rho_{\mathrm{ind}}$ through (6) and (7). Typical choices of $W_i$ are uniform weighting ($W_i = I_q$) and inverse variance weighting ($W_i = V_i^{-1}$), where $V_i = V(t_i, \xi, \tau)$. For simplicity of exposition, we assume that $W_i$ is known. When $W_i$ is diagonal, we notice that $\sigma_2^2 = \xi$, since $C_i$ has ones along the diagonal.

We estimate the regression parameters by weighted least squares

$$\hat{\beta} = \arg\min_{\beta} \sum_i (Y_i - \beta x_i^{\mathrm{T}}) W_i (Y_i - \beta x_i^{\mathrm{T}})^{\mathrm{T}}$$

and put $\hat{m}_i = \hat{\beta} x_i^{\mathrm{T}}$. The most important difference between the multivariate and univariate cases is that both $\tau$ and $\xi$ can be estimated, using the covariance structure of data, see Searle et al. (1992) for details. This makes not only $R^2 = \hat{\sigma}_1^2/\hat{\sigma}^2$ well defined, but also $\mathrm{CID} = \hat{\sigma}_1^2/\hat{\sigma}_{\mathrm{unexp}}^2$, with

$$\hat{m} = \sum_i \hat{m}_i W_i 1_q^{\mathrm{T}} \Big/ \sum_i 1_q W_i 1_q^{\mathrm{T}},$$

$$\hat{\sigma}_1^2 = \sum_i (\hat{m}_i - \hat{m}) W_i (\hat{m}_i - \hat{m})^{\mathrm{T}} \Big/ \sum_i \mathrm{tr}(W_i),$$

$$\hat{\sigma}_2^2 = \hat{\xi} \sum_i \mathrm{tr}(C_i W_i) \Big/ \sum_i \mathrm{tr}(W_i),$$

$$\hat{\sigma}_3^2 = \hat{\tau}.$$

With inverse variance weighting, $R^2$ agrees with Buse's (1973, 1979) generalized least squares extension of (1). In the psychology literature, $R^2$ has been considered for general weight matrices by Jöreskog and Sörbom (1981) and Tanaka and Huba (1989).

## 8. Large sample properties

In this section, we consider large sample properties of $R^2$ and CID when $q = 1$. We assume that the estimate of $\beta$ admits an asymptotically linear expansion

$$\hat{\beta} = \beta + \frac{1}{n} \sum_i \mathrm{IF}(Z_i) + \mathrm{o_p}(n^{-1/2}) \tag{22}$$

with $Z_i = (x_i, t_i, Y_i)$ and $\mathrm{IF}(z) = \mathrm{IF}(z; \theta)$ the vector valued $(1 \times p)$ influence function of $\hat{\beta}$, satisfying $E(\mathrm{IF}(s, Y)) = 0$ and $E|\mathrm{IF}(s, Y)|^2 < \infty$ for all $s = (x, t)$. See Hampel (1974) and Hampel et al. (1986) for more details.

Let $s_i = (x_i, t_i)$ be the $i$th design point. We assume that the empirical distribution of $\{s_i\}$ converges weakly

$$P_n = \frac{1}{n}\sum_i \delta_{s_i} \xrightarrow{\mathscr{L}} P \tag{23}$$

to some limiting measure $P$ as $n \to \infty$, where $\delta_s$ is a point mass at $s$.

We will assume that $w(\eta, t)$ is not a function of $\xi$. Given $s = (x, t)$, let $m(s) = m(\beta x^{\mathrm{T}}, t)$, $v(s) = v(\beta x^{\mathrm{T}}, \xi, t)$, $v_1(s) = v_1(\beta x^{\mathrm{T}}, t)$ and $w(s) = w(\beta x^{\mathrm{T}}, t)$. We further assume that

$$\begin{aligned} w := \sum_i w_i/n \to \bar{w} &:= \int w(s)\,\mathrm{d}P(s), \\ m \to \bar{m} &:= \int m(s)\,\mathrm{d}P(s)/\bar{w}, \\ \sigma_1^2 \to \bar{\sigma}_1^2 &:= \int w(s)(m(s)-\bar{m})^2\,\mathrm{d}P(s)/\bar{w}, \\ \sigma_2^2 \to \bar{\sigma}_2^2 &:= \int w(s)(v(s)-v_1(s))\,\mathrm{d}P(s)/\bar{w}, \\ \sigma_3^2 \to \bar{\sigma}_3^2 &:= \int w(s)v_1(s)\,\mathrm{d}P(s)/\bar{w} \end{aligned} \tag{24}$$

as $n \to \infty$. This gives the population coefficients of determination

$$\bar{\rho} = \bar{\sigma}_1^2/\bar{\sigma}^2,$$

$$\bar{\rho}_{\mathrm{ind}} = \bar{\sigma}_1^2/\bar{\sigma}_{\mathrm{ind}}^2,$$

where $\bar{\sigma}^2 = \bar{\sigma}_1^2 + \bar{\sigma}_2^2 + \bar{\sigma}_3^2$ and $\bar{\sigma}_{\mathrm{ind}}^2 = \bar{\sigma}_1^2 + \bar{\sigma}_2^2$. One option is to analyze the asymptotic distribution of $R^2$ and CID, viewed as estimators of $\bar{\rho}^2$ and $\bar{\rho}_{\mathrm{ind}}^2$. However, we think it is more relevant, given the sample at hand, to view $R^2$ and CID as estimators of $\rho$ and $\rho_{\mathrm{ind}}$. Therefore, asymptotic normality of $R^2$ and CID are stated as follows:

**Theorem 1.** *Assume that $m(\eta, t)$, $v_1(\eta, t)$ and $w(\eta, t)$ are continuously differentiable functions of $\eta$, that $\hat{\beta}$ admits an asymptotic expansion* (22)*, that the sequence of design points $\{s_i\}$ is such that* (23)*,* (24) *and* (A.10) *in the appendix hold, that the Lindeberg Condition is satisfied in* (A.12)*, as well as* (A.14) *and* (A.15)*. Then*

$$\sqrt{n}(R^2 - \rho) \xrightarrow{\mathscr{L}} \mathrm{N}(0, \sigma_R^2) \tag{25}$$

*and*

$$\sqrt{n}(\mathrm{CID} - \rho_{\mathrm{ind}}) \xrightarrow{\mathscr{L}} \mathrm{N}(0, \sigma_{\mathrm{CID}}^2) \tag{26}$$

*as $n \to \infty$, where the asymptotic variances $\sigma_R^2$ and $\sigma_{\mathrm{CID}}^2$ are defined in the proof of Theorem* 1 *in the appendix.*

Based on Theorem 1, we also compute standard errors (that is, estimates of $\sigma_R/\sqrt{n}$ and $\sigma_{\mathrm{CID}}/\sqrt{n}$). See the appendix for details.

## 9. An overdispersed poisson example

We illustrate the theory with a car accident data set from If P&C Insurance Company, using the overdispersed Poisson model of Example 1. We give a brief description of the data set here. More details can be found in Hössjer et al. (2006) and Järnmalm (2006). Car accidents are reported for a total of $n = 439\,283$ individuals having a uninterrupted three year period between January 1, 2002 and December 31, 2005. We measure time in units of three year intervals. Hence, $t_i \equiv 1$, although we allow for general time durations in the theoretical computations below. As rating factors we use previous number of years of insurance at the company (4), geographic zone (19), age of car (6), premium class, defined

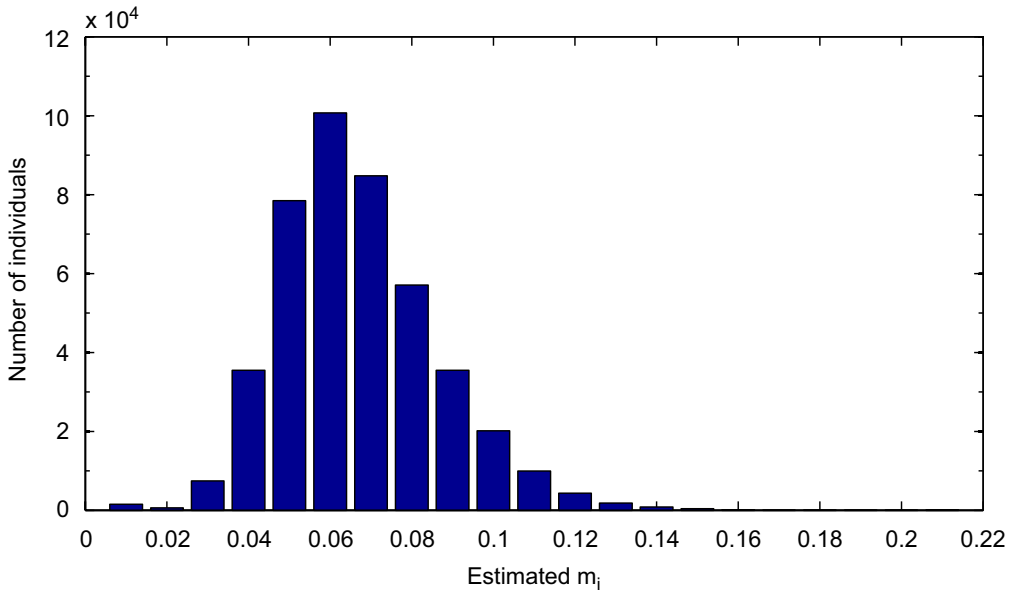


Fig. 1. Histogram of all $\hat{m}_i$, $i = 1, \ldots, n$ for the car accident data set.

by the type of car (10), annual driving distance (5) and sex/age (26). The number of classes within each rating factor is given in brackets. Each rating factor contains a reference class with a regression coefficient of zero (not included in $\beta$). All other classes have a separate regression coefficient $\beta_r$, with $x_{ir} = 1$ if individual $i$ belongs to the given class and $x_{ir} = 0$ otherwise. Including intercept ($\beta_1$, $x_{i1} \equiv 1$), the total number of regression coefficients is

$$p = 1 + \sum_{j=1}^{6} (k_j - 1) = 65,$$

where $k_j$ is the number of classes within the $j$th rating factor.

The regression parameters are estimated using ML from a generalized linear model without dispersion variance, (15). A histogram of all estimated claim frequencies $\hat{m}_i$ is displayed in Fig. 1. They range from 0.01 to 0.20 although $0.045 \leqslant \hat{m}_i \leqslant 0.075$ for more than 50% of all individuals. The estimated coefficients of determination are computed using a deterministic weight function $w_i = t_i$ of the kind described in part (iv) of Section 4. This yields

$$R^2 = 0.0053,$$

$$\mathrm{CID} = 0.1120.$$

Hence, only 0.5% of the total variance is explained by the covariates, and $R^2$ is much smaller than CID, since the individual unexplained variance $\hat{\sigma}_2^2$ is only 4% of the noise variance $\hat{\sigma}_3^2$. The reason for this is that time durations in motor traffic insurance are very short in relation to claim frequencies $m_i$. The relative amount of overdispersion is thus small, but yet highly significant (see Hössjer et al., 2006).

Even without noise variance, only 11% of the total individual variation is explained by the covariates. Three possible reasons for the low CID are (1) model error in capturing the true relation $x \to m(x)$ between the *given covariates* and the claim frequency, (2) missing covariates from the *given rating factors* (the rating factors have too few classes, missing interactions) or (3) missing rating factors, such as individual driving characteristics or road conditions. Of these, the latter is probably the most significant one.

In order to compute standard errors we need to specify the constant $a$ of the variance function in (18). An initial data analysis in Hössjer et al. (2006) gives $a = 1.3$, although the assumed $a$ affects standard errors very marginally (due to the small amount of overdispersion). Given $a$, $\xi$ is estimated using (14), which for our variance function becomes

$$\hat{\xi} = \sum_i t_i((Y_i - \hat{m}_i)^2 - \hat{m}_i/t_i) \Big/ \sum_i t_i \hat{m}_i^a = 0.0979. \tag{27}$$

The standard error of $R^2$ and CID is computed from (A.16) to (A.17), with $\hat{h}(Z_i)$ as in (A.26) and the standard error of $\hat{\xi}$ from (A.31). This yields the 95% Wald confidence intervals

$$I_\rho = (0.0049, 0.0058),$$

$$I_{\rho_{\text{ind}}} = (0.0967, 0.1274),$$

$$I_\xi = (0.0849, 0.1108).$$

We notice that $\rho_{\text{ind}}$ has a much wider confidence interval than $\rho$. This is due to the fact that unexplained individual variance is much more difficult to estimate than total unexplained variance in this data set, which has a small amount of overdispersion.

## 10. Discussion

In this paper we have defined a new coefficient of determination, CID, which quantifies the proportion of individual variation in data explained by the covariates within a general univariate mixed effects regression model. We have primarily focused on univariate regression models for which the noise variance $v_{1i}$ of $Y_i$ is a function of the regression parameters and the individual unexplained variance $v_{2i}$ of $Y_i$ contains an extra liability parameter $\xi$. This makes $\rho_{\text{ind}}$ well defined and identifiable, so that CID is computable.

Several extensions of the work are possible. For instance, when $p/n$ is non-negligible, $R^2$ typically has an upward bias and should be adjusted for reduced degrees of freedom. For uniform weights, one possibility, which generalizes the adjusted $R^2$ in linear regression (Fisher, 1924), is

$$R^2 = 1 - \frac{(n/(n-p))\hat{\sigma}^2_{\text{unexp}}}{(n/(n-1))\hat{\sigma}^2}.$$

The analogous adjustment of CID would be

$$\text{CID} = 1 - \frac{(n/(n-p))\hat{\sigma}^2_2}{(n/(n-1))\hat{\sigma}^2_{\text{ind}}}.$$

We have defined $R^2$ and CID using (9) rather than (10). In this way, both coefficients of determination depend on $\hat{\beta}$, but not $\hat{\xi}$. This is appropriate, since only $\hat{\beta}$ is robust towards misspecification of the variance function for a wide range of regression models. The standard errors of $R^2$ and CID, on the other hand, depend on both $\hat{\beta}$ and $\hat{\xi}$. When the amount of overdispersion is large, it is of interest to use robust estimates of the variance functions in order to obtain robust standard errors. See for instance Liang and Zeger (1986), Breslow (1990) and Moore and Tsiatis (1991) for more discussion on this subject.

For confidence intervals, we have used asymptotic normality of $R^2$ and CID. An alternative is to apply the normal approximation to a nearly variance stabilizing transformation of $R^2$, such as $R^2/(1-R^2)$, see e.g. Muirhead (1985). The same approach could also be used for CID. Another possibility would be to use bootstrap.

*Unconditional* versions of the variance decomposition (5) and its estimated counterpart (8) are obtained if $m$ and $\hat{m}$ are replaced by

$$m_{\text{uncond}} = E(Y_I|t_1,\dots,t_n) = \sum_i w_i E(Y_i|t_i) \Big/ \sum_i w_i,$$

$$\hat{m}_{\text{uncond}} = \sum_i \hat{w}_i Y_i \Big/ \sum_i \hat{w}_i.$$

A conceptual advantage of this approach is that the resulting variance $\sigma^2_{\text{uncond}}$ as well as its estimate $\hat{\sigma}^2_{\text{uncond}}$ are both independent of the chosen model, i.e. the set of included covariates. On the other hand, a random model for the covariates $x_i$ is implicitly required.

The two unexplained variance components, $\sigma_2^2$ and $\sigma_3^2$, are unaffected by removing conditioning on $\{x_i\}$, whereas the explained variance component $\sigma_1^2$ is changed to

$$\sigma_{1,\text{uncond}}^2 = \sigma_1^2 + (m_{\text{uncond}} - m)^2.$$

Similarly, for the estimated variance decomposition, only the explained variance component is changed from $\hat{\sigma}_1^2$ to

$$\hat{\sigma}_{1,\text{uncond}}^2 = \hat{\sigma}_1^2 + (\hat{m}_{\text{uncond}} - \hat{m})^2.$$

In practice, there is a very small difference between the two estimated variance decompositions. In fact, they often agree, i.e. $\hat{m} = \hat{m}_{\text{uncond}}$. This happens, for instance, for univariate linear regression and weighted least squares estimation of $\beta$ (although $\sigma_2^2$ and $\sigma_3^2$ cannot be separated) and Poisson regression when $w_i = t_i = 1$ and $\beta$ is estimated by non-dispersed ML (15).

## Acknowledgment

This work was supported by the Swedish Research Council, contract number 621-2005-2810. I wish to thank Bengt Eriksson at If Inc. for suggesting to look at unexplained individual variation of Poisson traffic data, for sharing the car accident's data set and for help with data analysis. Also thanks to Kajsa Järnmalm for help with data analysis and valuable comments on an earlier version of this manuscript as well as to an anonymous referee.

## Appendix A.

**Proof of Theorem 1.** We start by defining the asymptotic variances in (25)–(26) as

$$\sigma_R^2 = G\Sigma G^{\mathrm{T}}$$

and

$$\sigma_{\text{CID}}^2 = G_{\text{ind}}\Sigma G_{\text{ind}}^{\mathrm{T}},$$

where

$$G = \bar{\sigma}^{-2}(1 - \bar{\rho}, -\bar{\rho}, 0),$$

$$G_{\text{ind}} = \bar{\sigma}_{\text{ind}}^{-2}(1 - \bar{\rho}_{\text{ind}}, -\bar{\rho}_{\text{ind}}, \bar{\rho}_{\text{ind}}), \tag{A.1}$$

$$\Sigma = \int E(h(s, Y)^{\mathrm{T}} h(s, Y))\, \mathrm{d}P(s), \tag{A.2}$$

$$h(z) = \mathrm{IF}(z)D + \Delta(z), \tag{A.3}$$

$z = (s, y) = (x, t, y)$, $\Delta(z) = (0, w(s)((y - m(s))^2 - v(s))/\bar{w}, 0)$ and $D$ is the $p \times 3$ matrix defined in (A.11).

It is convenient to start the proof by introducing $S(\beta) = (S_1(\beta), S_2(\beta), S_3(\beta))$, where

$$S_1(\beta) = \sum_i w_i(m_i - m)^2,$$

$$S_2(\beta) = \sum_i w_i(Y_i - m_i)^2,$$

$$S_3(\beta) = \sum_i w_i v_{1i}. \tag{A.4}$$

Moreover, let $\hat{S} = (S_1(\hat{\beta}), S_2(\hat{\beta}), S_3(\hat{\beta}))$ and $S = (S_1(\beta), \sum_i w_i v_i, S_3(\beta))$. Then

$$\rho = g(S),$$

$$\rho_{\text{ind}} = g_{\text{ind}}(S) \tag{A.5}$$

and

$$R^2 = g(\hat{S}),$$
$$\mathrm{CID} = g_{\mathrm{ind}}(\hat{S}), \tag{A.6}$$

where $g(S_1, S_2, S_3) = S_1/(S_1 + S_2)$ and $g_{\mathrm{ind}}(S_1, S_2, S_3) = S_1/(S_1 + S_2 - S_3)$. To find the asymptotic distributions of $R^2$ and CID, we will use Taylor expansion of $g$ and $g_{\mathrm{ind}}$ around $S$. To this end, we first derive asymptotic approximations of $g'(S)$ and $g'_{\mathrm{ind}}(S)$. It follows from (24) that

$$S/(n\bar{w}) \xrightarrow{\mathrm{p}} \bar{S} = (\bar{\sigma}_1^2, \bar{\sigma}_2^2 + \bar{\sigma}_3^2, \bar{\sigma}_3^2) \tag{A.7}$$

as $n \to \infty$. This implies, by differentiating $g$ and $g_{\mathrm{ind}}$,

$$ng'(S) \xrightarrow{\mathrm{p}} G/\bar{w},$$
$$ng'_{\mathrm{ind}}(S) \xrightarrow{\mathrm{p}} G_{\mathrm{ind}}/\bar{w} \tag{A.8}$$

with $G$ and $G_{\mathrm{ind}}$ as in (A.1). Combining (A.5), (A.6) and (A.8), a Taylor expansion of $g$ and $g_{\mathrm{ind}}$ around $S$ yields

$$R^2 = \rho + (\hat{S} - S)G^{\mathrm{T}}/(\bar{w}n) + \mathrm{o_p}(n^{-1/2}),$$
$$\mathrm{CID} = \rho_{\mathrm{ind}} + (\hat{S} - S)G_{\mathrm{ind}}^{\mathrm{T}}/(\bar{w}n) + \mathrm{o_p}(n^{-1/2}). \tag{A.9}$$

To finalize the proof, we will derive the asymptotic distribution of $(\hat{S} - S)/\bar{w}$. To this end, define the $p \times 3$ matrix $S' = S'(\beta) = \partial S(\beta)/\partial\beta$ with transposed columns $S'_i$ given by

$$S'_1 = 2\sum_i x_i m'_i w_i (m_i - m) + \sum_i x_i w'_i (m_i - m)^2,$$
$$S'_2 = -2\sum_i x_i m'_i w_i (Y_i - m_i) + \sum_i x_i w'_i (Y_i - m_i)^2,$$
$$S'_3 = \sum_i x_i (w_i v'_{1i} + w'_i v_{1i}),$$

where $w'_i = \partial w(\eta_i, t_i)/\partial\eta_i$ and $v'_{1i} = \partial v_1(\eta_i, t_i)/\partial\eta_i$. It follows from Lemma 1 that

$$S'/(\bar{w}n) \xrightarrow{\mathrm{p}} D \tag{A.10}$$

as $n \to \infty$, where $D = (D_1^{\mathrm{T}}, D_2^{\mathrm{T}}, D_3^{\mathrm{T}})$ is a $p \times 3$ matrix, with

$$\bar{w}D_1 = 2\int x m'(s) w(s)(m(s) - \bar{m})\,\mathrm{d}P(s) + \int x w'(s)(m(s) - \bar{m})^2\,\mathrm{d}P(s),$$
$$\bar{w}D_2 = \int x w'(s) v(s)\,\mathrm{d}P(s),$$
$$\bar{w}D_3 = \int x(w(s)v'_1(s) + w'(s)v_1(s))\,\mathrm{d}P(s), \tag{A.11}$$

$s = (x, t)$, $m'(s) = \partial m(\eta, t)/\partial\eta|_{\eta=\beta x^{\mathrm{T}}}$, $v'_1(s) = \partial v_1(\eta, t)/\partial\eta|_{\eta=\beta x^{\mathrm{T}}}$ and $w'(s) = \partial w(\eta, t)/\partial\eta|_{\eta=\beta x^{\mathrm{T}}}$.

A Taylor expansion of $S(\cdot)$ around $\beta$ gives

$$\begin{aligned}\hat{S} &= S + (\hat{\beta} - \beta)S' + (S(\beta) - S) + \mathrm{o_p}(n^{1/2}) \\ &= S + n\bar{w}(\hat{\beta} - \beta)D + \bar{w}\sum_i \Delta(Z_i) + \mathrm{o_p}(n^{1/2}) \\ &= S + \bar{w}\sum_i h(Z_i) + \mathrm{o_p}(n^{1/2}),\end{aligned} \tag{A.12}$$

where $\Delta(Z_i) = (0, w_i((Y_i - m_i)^2 - v_i)/\bar{w}, 0)$. In the last step of (A.12) we used (22) and (A.3). By assumption, the sequence of design measures, $P_n$, is such that the Lindeberg Condition is satisfied in the last line of (A.12). Hence, the Central Limit Theorem implies

$$(\hat{S} - S)/(\bar{w}\sqrt{n}) \xrightarrow{\mathcal{L}} \mathrm{N}(0, \Sigma) \tag{A.13}$$

as $n \to \infty$. The theorem follows by combining (A.13) with (A.9). □

**Lemma 1.** *Assume that*

$$\lim_{a\to\infty} \sup_n \int_{s;|k_j(s)|\geqslant a} |k_j(s)|\,\mathrm{d}P_n(s) = 0 \tag{A.14}$$

*holds for each of the functions* $k_1(s) = 2xm'(s)w(s)(m(s) - \bar{m}) + xw'(s)(m(s) - \bar{m})^2$, $k_2(s) = xw'(s)v(s)$ *and* $k_3(s) = x(w(s)v_1'(s) + w'(s)v_1(s))$. *Assume further that*

$$\int k_j(s)\,\mathrm{d}P_n(s) = \mathrm{o}(n) \tag{A.15}$$

*holds for* $k_4(s) = (xx^{\mathrm{T}}m'(s)w(s))^2 v(s)$ *and* $k_5(s) = (xx^{\mathrm{T}}w'(s))^2\tau(s)$, *where* $\tau(s) = E((Y - m(\beta x^{\mathrm{T}}, t))^2 - v(\beta x^{\mathrm{T}}, \xi, t))^2$. *Then* (A.10) *follows*.

**Proof.** By definition of the functions $k_j$, we have

$$S_1'/n = \int k_1(s)\,\mathrm{d}P_n(s),$$

$$S_2'/n = W + \int k_2(s)\,\mathrm{d}P_n(s),$$

$$S_3'/n = \int k_3(s)\,\mathrm{d}P_n(s),$$

$$\bar{w}D_j = \int k_j(s)\,\mathrm{d}P(s), \quad j = 1, 2, 3,$$

where $W$ is a random variable with first two moments satisfying $E(W) = 0$ and $E(WW^{\mathrm{T}}) \leqslant n^{-1}(8\int k_4(s)\,\mathrm{d}P_n(s) + 2\int k_5(s)\,\mathrm{d}P_n(s))$. Chebyshev's Inequality and (A.15) then imply $W \xrightarrow{\mathrm{p}} 0$ as $n \to \infty$.

It thus remains to show that $\int k_j(s)\,\mathrm{d}P_n(s) \to \int k(s)\,\mathrm{d}P(s)$ as $n \to \infty$ for $j = 1, 2, 3$. This is equivalent to showing that $E(W_{jn}) \to E(W_j)$, where the random variables $W_{jn}$ and $W_j$ are defined as follows: $W_{jn} = k_j(x_I, t_I)$, with $I$ having a uniform distribution on $\{1, \ldots, n\}$ and $W_j = k_j(x, t)$, with $(x, t) \sim P$. The regularity conditions of Theorem 1 imply that each $k_j$ is a continuous function. Hence, (23) and the Continuous Mapping Theorem imply that $W_{jn} \xrightarrow{\mathcal{L}} W_j$ for $j = 1, 2, 3$ and (A.14) is a uniform integrability condition for $\{W_{jn}\}_n$ assuring that $E(W_{jn}) \to E(W_j)$. □

*Standard errors of* $R^2$ *and CID*: We use Theorem 1 to define the squared standard errors

$$d^2 := n^{-1}\hat{G}\hat{\Sigma}\hat{G}^{\mathrm{T}},$$
$$d_{\mathrm{ind}}^2 := n^{-1}\hat{G}_{\mathrm{ind}}\hat{\Sigma}\hat{G}_{\mathrm{ind}}^{\mathrm{T}} \tag{A.16}$$

of $R^2$ and CID, with

$$\hat{G} = \hat{\sigma}^{-2}(1 - R^2, -R^2, 0),$$
$$\hat{G}_{\mathrm{ind}} = \hat{\sigma}_{\mathrm{ind}}^{-2}(1 - \mathrm{CID}, -\mathrm{CID}, \mathrm{CID}),$$
$$\hat{\Sigma} = \sum_i \hat{h}(Z_i)^{\mathrm{T}}\hat{h}(Z_i)/n. \tag{A.17}$$

Here $\hat{h}(Z_i) = \widehat{\mathrm{IF}}(Z_i)\hat{D} + \hat{\Delta}(Z_i)$ is an estimate of $h(Z_i)$, $\hat{w} = \sum_i \hat{w}_i/n$, $\hat{\Delta}(Z_i) = (0, \hat{w}_i((Y_i - \hat{m}_i)^2 - \hat{v}_i)/\hat{w}, 0)$, and $\hat{D} = (\hat{D}_1^{\mathrm{T}}, \hat{D}_2^{\mathrm{T}}, \hat{D}_3^{\mathrm{T}}) = S'(\hat{\beta})/\sum_i \hat{w}_i$, i.e.

$$\hat{D}_1 = \left(2\sum_i x_i \hat{m}_i' \hat{w}_i(\hat{m}_i - \hat{m}) + \sum_i x_i \hat{w}_i'(\hat{m}_i - \hat{m})^2\right) \Big/ \sum_i \hat{w}_i,$$

$$\hat{D}_2 = \left(-2\sum_i x_i \hat{m}_i' \hat{w}_i(Y_i - \hat{m}_i) + \sum_i x_i \hat{w}_i'(Y_i - \hat{m}_i)^2\right) \Big/ \sum_i \hat{w}_i,$$

$$\hat{D}_3 = \sum_i x_i(\hat{w}_i \hat{v}_{1i}' + \hat{w}_i' \hat{v}_{1i}) \Big/ \sum_i \hat{w}_i, \tag{A.18}$$

where $\hat{m}_i' = \partial m(\hat{\eta}_i, t_i)/\partial\hat{\eta}_i$, $\hat{v}_{1i}' = \partial v_1(\hat{\eta}_i, t_i)/\partial\hat{\eta}_i$ and $\hat{w}_i' = \partial w(\hat{\eta}_i, t_i)/\partial\hat{\eta}_i$.

It remains to compute the estimated influence function $\widehat{\mathrm{IF}}$, usually referred to as the sensitivity function. Assume first that $\beta$ is estimated separately from $\xi$, using $p$ estimating equations

$$\sum_i \psi(Z_i; \hat{\beta}) = 0 \tag{A.19}$$

given some $1 \times p$-valued function $\psi$. This includes nonlinear least squares, with $\psi(z; \beta) = xm'(\eta, t)(Y - m(\eta, t))$ and $m'(\eta, t) = \partial m(\eta, t)/\partial\eta$. Other examples are ML-estimators (15), with $\psi(z; \beta) = x\partial \log f(y|\eta, 0, t)/\partial\eta$ and QL-estimators $U_\beta(\hat{\beta}, 0) = 0$, with $\psi(z; \beta) = xm'(\eta, t)(y - m(\eta, t))/v(\eta, 0, t)$.

The influence function for estimators (A.19) is given by

$$\mathrm{IF}(z) = -\psi(z; \beta)B^{-1},$$

where the $p \times p$ matrix $B = \int E(\psi'(s, Y; \beta))\,\mathrm{d}P(s)$ and $\psi'(z; \beta) = \partial\psi(z; \beta)/\partial\beta$, see Hampel et al. (1986). The sensitivity function is

$$\widehat{\mathrm{IF}}(z) = -\psi(z; \hat{\beta})\hat{B}^{-1}, \tag{A.20}$$

where $\hat{B} = \sum_i \psi'(Z_i; \hat{\beta})/n$.

When $\beta$ and $\xi$ are estimated simultaneously from $p + 1$ estimating equations

$$\sum_i \psi(Z_i, \hat{\theta}) = 0,$$

the influence function for $\hat{\beta}$ is

$$\mathrm{IF}(z) = -(\psi(z; \theta)B^{-1})_{1:p},$$

where $B = \int E(\psi'(s, Y; \theta))\,\mathrm{d}P(s)$ is a $(p + 1) \times (p + 1)$ matrix, $\psi'(z; \theta) = \partial\psi(z; \theta)/\partial\theta$ and $a_{1:p}$ contains the first $p$ components of the row vector $a$. To compute the sensitivity function $\widehat{\mathrm{IF}}$, we proceed analogously as in (A.20).

*Standard errors of $R^2$ and CID for Poisson model with non-dispersed ML-estimates*: In order to compute $d^2$ and $d^2_{\mathrm{ind}}$ in (A.16), we need an explicit expression for $\hat{h}(Z_i)$ in (A.17). The probability distribution for the non-dispersed Poisson model is

$$f(y|\eta, 0, t) = \mathrm{e}^{-t\mathrm{e}^{\eta}}(t\mathrm{e}^{\eta})^{ty}/(ty)!. \tag{A.21}$$

Taking the logarithm of (A.21), putting $\eta = \beta x^{\mathrm{T}}$ and differentiating twice with respect to $\beta$, we obtain

$$\psi(z; \beta) = tx(y - m(\eta)),$$

$$\psi'(x; \beta) = -tx^{\mathrm{T}}xm(\eta). \tag{A.22}$$

Since $m(\eta) = \mathrm{e}^{\eta}$ and $v(\eta, 0) = \mathrm{e}^{\eta}/t$, we find that $\hat{m}_i = \hat{m}_i' = \exp(x_i\hat{\beta}^{\mathrm{T}})$ and $\hat{v}_{1i}' = \exp(x_i\hat{\beta}^{\mathrm{T}})/t_i$, and from (A.22) we obtain $\hat{B} = -\sum_i t_i x_i^{\mathrm{T}} x_i \hat{m}_i/n$.

With weights $w_i = \hat{w}_i = t_i$, the transposed columns of $\hat{D}$ are

$$\hat{D}_1 = 2\sum_i t_i x_i \hat{m}_i(\hat{m}_i - \hat{m}) \Big/ \sum_i t_i,$$

$$\hat{D}_2 = -2\sum_i t_i x_i \hat{m}_i(Y_i - \hat{m}_i) \Big/ \sum_i t_i,$$

$$\hat{D}_3 = \sum_i x_i \hat{m}_i \Big/ \sum_i t_i \tag{A.23}$$

and $\hat{\Delta}(Z_i) = (0, t_i((Y_i - \hat{m}_i)^2 - \hat{v}_i)/t, 0)$, where $\hat{v}_i = \hat{m}_i/t_i + \hat{\xi}\hat{m}_i^a$ and $t = \sum_i t_i/n$. Putting things together we find that

$$\hat{h}(Z_i) = \left(a_i(Y_i - \hat{m}_i), b_i(Y_i - \hat{m}_i) + \frac{t_i}{t}((Y_i - \hat{m}_i)^2 - \hat{v}_i), c_i(Y_i - \hat{m}_i)\right), \tag{A.24}$$

where

$$a_i = -t_i x_i \hat{B}^{-1}\hat{D}_1^{\mathrm{T}} \overset{t_i=1}{=} -x_i \hat{B}^{-1}\hat{D}_1^{\mathrm{T}},$$

$$b_i = -t_i x_i \hat{B}^{-1}\hat{D}_2^{\mathrm{T}} \overset{t_i=1}{=} -x_i \hat{B}^{-1}\hat{D}_2^{\mathrm{T}},$$

$$c_i = -t_i x_i \hat{B}^{-1}\hat{D}_3^{\mathrm{T}} \overset{t_i=1}{=} -x_i \hat{B}^{-1}\hat{D}_3^{\mathrm{T}} = 1, \tag{A.25}$$

where the last identity follows since the model includes intercept ($x_{i1} \equiv 1$). We can simplify further by letting $b_i = 0$ (using the fact that $\hat{D}_2 \approx 0$). Consequently, when $t_i \equiv 1$, we put

$$\hat{h}(Z_i) = (a_i(Y_i - \hat{m}_i), (Y_i - \hat{m}_i)^2 - \hat{v}_i, Y_i - \hat{m}_i), \tag{A.26}$$

which is inserted into (A.16)–(A.17) to yield the standard errors of $R^2$ and CID.

*Asymptotic normality and standard error for* $\hat{\xi}$ *in* (27): We use a procedure analogous to the proof of Theorem 1. Basically we need to redefine the functions $S(\cdot)$ and $g(\cdot)$. We redefine the first component of $S(\beta)=(S_1(\beta), S_2(\beta), S_3(\beta))$ in (A.4), choose weights $w_i = t_i$ and variance function $v_{1i} = m_i/t_i$. This yields

$$S_1(\beta) = \sum_i t_i m_i^a,$$

$$S_2(\beta) = \sum_i t_i (Y_i - m_i)^2,$$

$$S_3(\beta) = \sum_i m_i.$$

With $\hat{S} = S(\hat{\beta})$ and $S = (S_1(\beta), \sum_i t_i v_i, S_3(\beta))$, we find that

$$\xi = g(S),$$

$$\hat{\xi} = g(\hat{S}),$$

if $g(S_1, S_2, S_3) = (S_2 - S_3)/S_1$. The Taylor expansion analogous to (A.9) is

$$\hat{\xi} = \xi + (\hat{S} - S)G^{\mathrm{T}}/(\bar{t}n) + \mathrm{o_p}(n^{-1/2})$$

with $\bar{t} = \int t \,\mathrm{d}P(s)$ and

$$G = \bar{t}(-\xi, 1, -1) \Big/ \int t m^a(x) \,\mathrm{d}P(s). \tag{A.27}$$

Expansion (A.12) still holds for the redefined $\hat{S}$ and $S$, with $\bar{w} = \bar{t}$, $D = (D_1^{\mathrm{T}}, D_2^{\mathrm{T}}, D_3^{\mathrm{T}})$,

$$\begin{aligned}
&\bar{t} D_1 = a \int txm'(x) m(x)^{a-1} \, \mathrm{d}P(s) = a \int txm^a(x) \, \mathrm{d}P(s), \\
&D_2 = (0, \ldots, 0), \\
&\bar{t} D_3 = \int txm'(x) \, \mathrm{d}P(s) = \int txm(x) \, \mathrm{d}P(s)
\end{aligned} \tag{A.28}$$

and

$$h(z) = \mathrm{IF}(z) D + (0, t((y - m(x))^2 - v(x))/\bar{t}, 0). \tag{A.29}$$

Altogether, we obtain

$$\sqrt{n}(\hat{\xi} - \xi) \xrightarrow{\mathscr{L}} \mathrm{N}(0, G \Sigma G^{\mathrm{T}})$$

as $n \to \infty$, with $G$ and $\Sigma$ as in (A.27) and (A.2), using (A.29) and (A.28) in the definition of $\Sigma$.

The squared standard error of $\hat{\xi}$ is then

$$d_{\hat{\xi}}^2 = n^{-1} \hat{G} \hat{\Sigma} \hat{G}^{\mathrm{T}} \tag{A.30}$$

with $\hat{G} = (-\hat{\xi}, 1, -1) \sum_i t_i / \sum_i t_i \hat{m}_i^a$, $\hat{\Sigma}$ as defined in (A.17), $\hat{h}(Z_i)$ as in (A.24) and $a_i$, $b_i$ and $c_i$ as in (A.25). In (A.25), $\hat{D}_2$ and $\hat{D}_3$ are given by (A.23), whereas

$$\hat{D}_1 = a \sum_i t_i x_i \hat{m}_i^a \Big/ \sum_i t_i.$$

Substantial simplification is possible when $a = 1$ and $t_i = 1$. Then the first and third components of $S$, $\hat{S}$, $D$ and $\hat{D}$ are equal, and $a_i = c_i = 1$ (since intercept is included in the model). Moreover, we put $b_i = 0$ (since $\hat{D}_2 \approx 0$) and use $\sum_i t_i (Y_i - \hat{m}_i) = 0$ (which follows from the likelihood equations). Altogether, this implies

$$\begin{aligned}
d_{\hat{\xi}}^2 = \left( \sum_i Y_i \right)^{-2} \Bigg( &(\hat{\xi} + 1)^2 \sum_i (Y_i - \hat{m}_i)^2 - 2(\hat{\xi} + 1) \sum_i ((Y_i - \hat{m}_i)^3 - \hat{v}_i (Y_i - \hat{m}_i)) \\
&+ \sum_i ((Y_i - \hat{m}_i)^2 - \hat{v}_i)^2 \Bigg).
\end{aligned} \tag{A.31}$$

## References


Breslow, N.E., 1984. Extra-Poisson variation in log-linear models. Appl. Statist. 33 (1), 38–44.

Breslow, N.E., 1990. Tests of hypotheses in overdispersed Poisson regression and other quasilikelihood models. J. Amer. Statist. Assoc. 85, 565–571.

Breslow, N.E., Clayton, D., 1993. Approximate inference in generalized linear mixed models. J. Amer. Statist. Assoc. 88, 9–25.

Buse, A., 1973. Goodness of fit in generalized squares estimation. Amer. Statist. 27, 106–108.

Buse, A., 1979. Goodness of fit for seemingly unrelated regression model: a generalization. J. Econometrics 10, 109–113.

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

Cochran, W.G., 1940. The analysis of variance when experimental errors follow the Poisson or binomial law. Ann. Math. Statist. 11 (3), 335–347.

Cox, D.R., 1983. Sore remarks on overdispersion. Biometrika 70 (1), 269–274.

Cox, D.R., Snell, E.J., 1989. The Analysis of Binary Data. second ed. Chapman & Hall, London.

Crowder, M.J., 1978. Beta-binomial anova for proportions. Appl. Statist. 27, 34–37.

Fisher, R.A., 1924. The influence of rainfall on the yield of wheat at Rothamsed. Philos. Trans. R. Soc. London Ser. B 213, 89–124.

Hampel, F., 1974. The influence curve and its role in robust estimation. J. Amer. Statist. Assoc. 69, 383–393.

Hampel, F., Ronchetti, E., Rousseeuw, P., Stahel, W., 1986. Robust Statistics, The Approach based on Influence Functions. Wiley, New York.

Haseman, J.K., Kupper, L.L., 1978. The use of a correlated binomial model for the analysis of certain toxicological experiments. Biometrics 35, 281–293.

Hinde, J., 1982. Compound Poisson regression models. In: Gilchrist, R. (Ed.), GLIM 82: Proceedings of the International Conference in Generalized Linear Models. Springer, Berlin, pp. 199–121.

Hössjer, O., Eriksson, B., Järnmalm, K., Ohlsson, E., 2006. Assessing individual unexplained variation in non-life insurance. Mathematical Statistics, Stockholm University, Research Report 2006:12.

Järnmalm, K., 2006. Measures of the remaining systematic variance between individuals when divided into individual premium groups in non-life insurance. Master Thesis, Mathematical Statistics, Stockholm University, Report 2006:15 (in Swedish).

Jöreskog, K.G., Sörbom, D., 1981. Analysis of linear structural relationships by maximum likelihood and least squares methods. Research Report 81-8, Uppsala University, Sweden.

Lawless, J.F., 1987. Negative binomial and mixed Poisson regression. Canadian J. Statist. 15 (3), 209–225.

Lee, H.S., Chaubey, Y.P., 1996. MINQUE of variance components in generalized linear models with random effects. Comm. Statist. Theory Methods 25 (6), 1375–1382.

Liang, K.Y., Zeger, S.L., 1986. Longitudinal data analysis using generalized linear models. Biometrika 73, 13–22.

Lynch, M., Walsh, B., 1998. Genetics and Analysis of Quantitative Traits. Sinauer Associates Inc., Sunderland, MA, USA.

Maddala, G.S., 1983. Limited-Dependent and Qualitative Variables in Econometrics. Cambridge University Press, Cambridge.

Magee, L., 1990. $R^2$ measures based on Wald and likelihood ratio joint significance tests. Amer. Statist. 44, 250–253.

McCullagh, P., 1983. Quasi-likelihood functions. Ann. Statist. 11, 59–67.

McCullagh, P., Nelder, J.A., 1989. Generalized Linear Models. second ed. Chapman & Hall, London.

Moore, D.F., 1986. Asymptotic properties of moment estimators for overdispersed counts and proportions. Biometrika 73, 583–588.

Moore, D.F., Tsiatis, A., 1991. Robust estimation of the variance in moment methods for extra-binomial and extra-Poisson variation. Biometrics 47, 383–401.

Muirhead, R.J., 1985. Estimating a particular function of the multiple correlation coefficient. J. Amer. Statist. Assoc. 80, 923–925.

Nagelkerke, N.J.D., 1991. A note on a general definition of the coefficient of determination. Biometrika 78 (3), 691–692.

Pocock, S.J., Cook, D.G., Beresford, S.A.A., 1981. Regression of area mortality rates on explanatory variables: What weighting is appropriate? Appl. Statist. 30, 286–295.

Searle, S.R., Casella, G., McCulloch, C.E., 1992. Variance Components. Wiley, New York.

Tanaka, J.S., Huba, G.J., 1989. A general coefficient of determination for covariance structure models under arbitrary GLS estimation. British J. Math. Statist. Psych. 42, 233–239.

Wedderburn, R.W.M., 1974. Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. Biometrika 61, 439–447.

White, H., 1982. Maximum likelihood under misspecified models. Econometrica 50, 1–25.

Williams, D.A., 1975. The analysis of binary responses from toxicological experiments involving reproduction and teratogenicity. Biometrics 31, 949–952.

Williams, D.A., 1982. Extra-binomial variation in logistic linear models. Appl. Statist. 31 (2), 144–148.