On association in regression: the coefficient of determination revisited

A. van der Linde* & G. Tutz**

10.6.2004

Abstract

Universal coefficients of determination are investigated which quantify the strength of the relation between a vector of dependent variables Y and a vector of independent covariates X. They are defined as measures of dependence between Y and X through $\theta(x)$, with $\theta(x)$ parameterizing the conditional distribution of Y given $X = x$. If $\theta(x)$ involves unknown coefficients γ the definition is conditional on γ, and in practice γ, respectively the coefficient of determination has to be estimated. The estimates of quantities we propose generalize R^2 in classical linear regression and are also related to other definitions previously suggested. Our definitions apply to generalized regression models with arbitrary link functions as well as multivariate and nonparametric regression. The definition and use of the proposed coefficients of determination is illustrated for several regression problems with simulated and real data sets.

Keywords: association, mutual information, regression, coefficient of determination, R^2, correlation curve.

---

* University of Bremen, Germany
FB3: Institute of Statistics, PO Box 330 440, 28334 Bremen, email: avdl@math.uni-bremen.de (corresponding author)

** Ludwig-Maximilian-University Munich, Germany
Institute of Statistics, Ludwigstr. 33, 80539 Munich, email: tutz@stat.uni-muenchen.de

1 Introduction

As part of a regression analysis statisticians are interested in describing the strength of the relation between the dependent variable Y and the independent variables X. The measure commonly used to this end is the coefficient of determination, usually denoted by R^2. For jointly Gaussian variables X and Y, R^2 is related to the decomposition of the total variance of Y into the 'between variance' $var_X\, E(Y|X)$ (explained by regression) and the 'within variance' $E_X var(Y|X)$ (error) and describes the proportion of the total variance explained by regression. The decomposition of variance holds without any distributional assumption and might therefore be used as a starting point for generalizations. In particular local measures $R^2(x)$ summarized in a 'correlation curve' were suggested ([9]: Doksum et al., 1994) referring to this idea. For one-parameter exponential families where the conditional distribution of Y given $X = x$ is specified by generalized (linear) regression models different definitions of R^2 were proposed ([26]: Magee, 1990; [28]: Nagelkerke, 1991). These are based on a full likelihood ratio rather than just the variance thus taking into account that the separation of the conditional mean $E(Y|x)$ and the conditional variance $var(Y|x)$ is a special feature of the Gaussian distribution.

Although the attempts to define more generally a coefficient of determination indicate that there is a 'natural appeal' ([26]: Magee, 1990) of such a measure its definition and use is not really clear. Often a coefficient of determination is specified descriptively rather than theoretically as a quantity of interest. Hence features of the parameter of interest and properties of estimates continue being mixed up.

The ambiguity of continuing attempts to provide a coefficient of determination while its use and interpretation even in classical examples remains a subject of debate indicates a lack of clarity about the underlying concept 'strength of relation between Y and X' that is to be captured and quantified. In this paper we describe simple information theoretic ideas that help to disentangle the various notions associated with a coefficient of determination, and we introduce and review universal definitions of such global as well as local (conditional on x) coefficients as measures of dependence. Although measures of association and dependence between variables X and Y have been extensively discussed, especially in psychometrics (e.g. [7]: Cramer and Nicewander, 1979; [33]: Särndal, 1974) they have hardly been evaluated taking into account a regression model ([20]: Kent, 1983). Based on measures of association the determination of Y by $X(= x)$, can be interpreted in two ways: determination of the conditional density $p(y|x)$, capturing the discriminatory power of $X(= x)$ or determination of the (range of) values of Y given $X(= x)$, capturing the explanatory power of $X(= x)$. Of course, these ideas are related as different conditional densities correspond to different ranges of values of Y. The first notion will be the focus of the paper, but the second notion, usually formalized referring to entropies of distributions of Y will be discussed as well.

We specify a parameter θ that determines the distribution of a response Y as a function of the independent variables X and coefficients γ, and hence in our approach a regression model induces a family of parameterized densities $p(y|\theta(x,\gamma))$ for Y which vary depending on x. We claim that a coefficient of determination should quantify the strength of the relation between X and Y in a regression model (given the regression coefficients) measuring the variation in the family of distributions induced by X. Different amounts of variation may be obtained depending on which covariates are included in the model ; indeed a major field of application of coefficients of determination has been the problem of variable selection (e.g. [11]: Draper and Smith, 1981, ch.6.1; [25]: McKay, 1977). Also, the choice of the link function in a generalized regression model or the choice of a smoothing parameter in nonparametric regression has an impact on the strength of the relation between covariates X and Y. Thus intuitively our approach is based on the idea that the flexibility of a (regression) model is reflected by the range of conditional densities. This range is quantified as the average distance to a reference density, e.g. the marginal density $p_\theta(y)$ which is estimable only under joint sampling. Alternatively the reference density is often chosen corresponding to a regression model with regression coefficients equal to zero. Under conditional sampling the average w.r.t. X is to be taken according to the experimental design.

A coefficient of determination is not only a descriptive measure of a given regression model, but mainly of interest in model comparison. The coefficient used to this end depends on the set of models under consideration. Two comparisons are usually distinguished: (1) given a joint distribution of Y and all covariates the comparison of sub-models e.g. based on subsets of covariates, (2) the comparison of models specified by the same type of conditional density $p(y|x)$ with different parameters $\theta(x,\gamma)$, e.g. corresponding to different subsets or different functions of the covariates. The reference density in the coefficient of determination has to be chosen accordingly, and hence several coefficients of determination have been proposed in the literature.
In order to express emphasis on x and for notational convenience we drop γ writing simply $\theta(x) := \theta(x,\gamma)$ unless we explicitly refer to γ. In practice γ needs to be estimated from a given data set.

A local measure of association given $X = x$ may be defined based on the rate of change of densities $p(y|\theta(x))$ if discrimination of conditional densities is in focus and X is continuous. Alternatively the reduction of uncertainty by x yields a local coefficient of determination in terms of explanation.

We introduce and investigate universal measures of dependence for regression models yielding coefficients of determination for all regression models, especially for univariate as well as multivariate response variables, for non- and semi-parametric regression models, for generalized regression models with arbitrary link function or for regression models with heterogeneous variances.

The paper is organized as follows. In section 2 we start out with a formal

definition of a coefficient of determination based on a joint distribution of X and Y. We derive representations in exponential families, and we also illustrate for classical examples. In section 3 we refine the proposed definition considering approximations as well as alternatives, particularly alternatives based on different reference densities. We further discuss determination curves as generalization of correlation curves quantifying the local strength of relation between X and Y, and we introduce alternatives. In section 4 we summarize some properties of coefficients of determination, especially monotonicity in the number of covariates, and we discuss the use of coefficients of determination in model comparison. Section 5 briefly addresses estimation of a coefficients of determination, and in section 6 we exemplify our approach for several regression models applied to simulated and real data. Section 7 concludes with a discussion.
Readers who are interested in a conceptual outline only are referred to sections 2.1-2.2, the discussion in section 7 and perhaps section 3. Examples with technical details are collected in sections 2.3 and section 6. Sections 4 and 5 are both sketchy, yet addressing rather subtle results and may be skipped by casual readers.

2 Mutual information in regression

2.1 Motivation and definitions

Assume a random vector Y is observed with its distribution given by the density $p(y|\theta)$. In a regression problem we try to explain Y by covariates X modelling θ as a function of X and parameters γ, part of which are regression coefficients. Typically $\gamma = (\omega, \beta)$, where β denotes the regression coefficients and ω comprises an intercept α and possibly a scale factor ϕ. Thus, more precisely, $\theta = \theta(x, \gamma)$. In order to determine the strength of the relation between Y and X we propose to use a measure of dependence between Y and X, keeping γ fixed. We may either assume that observations (y, x) are realizations of (Y, X) following a joint distribution of (Y, X) or that we observe Y conditional on x, and that the values of X are specified according to an experimental design $d = \binom{x_1 \ldots x_q}{n(x_1) \ldots n(x_q)}$ which can be interpreted as a discrete probability measure such that $P(X = x_i) = p(x_i) = n(x_i)/n$ where $n = \sum_i n(x_i)$. We set $p_\theta(x, y) = p(y|\theta(x))p(x)$ and $p_\theta(y) = \int p_\theta(x, y)dx$.

In the sequel we use the following notation for (sequential) expectations: random variables w.r.t. which expectations are taken are written in capital letters, and for the first (and possibly only) expectation no subscript is appended. For sequential expectations a subscript indicates which variable it refers to. For example, in $cov_Y(E(\zeta(X)|Y), t(Y))$ the first expectation refers to X conditional on Y, and the covariance is a covariance of functions of Y.

A well-known measure of dependence between two random vectors is the distance between their joint density and the product of their marginal densities representing independence. Using the *Kullback–Leibler (KL-) discrepancy* or *directed divergence* which for two densities $p(z), q(z)$ is given by

$$I_{KL}(p(z), q(z)) = \int p(z) \log \frac{p(z)}{q(z)} dz$$

one obtains

$$I(\theta(X), Y) := I_{KL}(p_\theta(x, y), p(x)p_\theta(y)). \tag{1}$$

The directed divergence is also called the *mutual information* between Y and X. The KL-discrepancy between two densities has been extensively studied by Kullback ([22]: 1968) as a key quantity in establishing an information-theoretic approach to statistics. $I(\theta(X), Y)$ is symmetric in X and Y, but not a symmetric distance between the densities $p_\theta(x, y)$ and $p(x)p_\theta(y)$. Using the symmetric discrepancy or *divergence*

$$J_{KL}(p(z), q(z)) = I_{KL}(p(z), q(z)) + I_{KL}(q(z), p(z)) \tag{2}$$

one obtains

$$\begin{aligned} J(\theta(X), Y) := &\quad J_{KL}(p_\theta(x, y), p(x)p_\theta(y)) \\ = &\int (p_\theta(x, y) - p(x)p_\theta(y)) \log \frac{p_\theta(x, y)}{p(x)p_\theta(y)}\, dx\, dy. \end{aligned} \tag{3}$$

Note that $J(\theta(X), Y) = J(X, Y)$. We use the more complex notation though, in order to emphasize the model in use, and representations of $J(\theta(X), Y)$ may depend on the parameterization. For our purposes the representation

$$\begin{aligned} J(\theta(X), Y) &= E \int (p(y|\theta(X)) - p_\theta(y)) \log \frac{p(y|\theta(X))}{p_\theta(y)}\, dy \\ &= E\, J_{KL}(p(y|\theta(X)), p_\theta(y)) \end{aligned} \tag{4}$$

is intuitively most appealing. $J(\theta(X), Y)$ describes the range of densities for Y induced by regression models $\theta(x)$ w.r.t. to the marginal reference density $p_\theta(y)$ which does not depend on x. $J(\theta(X), Y)$ measures how strongly the densities induced by $\theta(X)$ deviate from the marginal density. Thus it captures the discriminatory power of x within the regression model. $J(\theta(X), Y)$ takes values in $\mathbb{R}^+$. A value of 0 indicates independence of Y and X, while a high value indicates variability of $p(y|\theta(x))$ as a function of x. We suggest to refer to a scaled version of $J(\theta(X), Y)$ as a coefficient of determination.

Definition 1:
Define

$$R_J^2 = \frac{J(\theta(X), Y)}{1 + J(\theta(X), Y)} \in [0, 1] \tag{5}$$

to be the *coefficient of determination of Y by X through θ* based on the divergence. $\lhd$

One can immediately read off the definition that if the conditional density of Y given x actually does not depend on x, e.g. $\theta(x) \equiv c_0$, $J(\theta(X), Y)$ equals zero and $R_J^2 = 0$.

2.2 Representations of the divergence

2.2.1 The integrated log–odds ratio function

Recently, Osius ([29]: 2000) demonstrated that under very mild regularity conditions the joint distribution of two random elements is characterized by their marginal distributions and the odds ratio function. Hence the association between the two random elements is captured in the odds ratio function, and a measure of dependence might be regarded as a functional of a representative of the odds ratio function specified by reference values. In our context focusing on random vectors Y and X, a representative of the *log–odds ratio function* with reference values x_0, y_0 is given by

$$\Psi^0(\theta(x), y) = \log \frac{p(y|\theta(x))}{p(y_0|\theta(x))} - \log \frac{p(y|\theta(x_0))}{p(y_0|\theta(x_0))}. \tag{6}$$

Then the following result holds ((([24]: van der Linde, 2004, appendix).

Result 1:

$$J(\theta(X), Y) = E_{XY} \Psi^0(\theta(X), Y) - E_X\, E_Y \Psi^0(\theta(X), Y) \tag{7}$$

for all reference values x_0, y_0. $\square$

Silvey's ([34]: 1964) measure of association is derived from a similar intuition, but it does not operate on the log-scale.

2.2.2 Exponential families

Prevailing in regression problems is the assumption that conditional on x respectively on $\theta(x)$ the density of Y belongs to a k–parameter exponential family. We write

$$p(y|\zeta(x)) = \exp\{\zeta(x)^T t(y) - M(\zeta(x))\},$$

where $\zeta(x) \in \mathbb{R}^k$ is the canonical or natural parameter, $t(y)$ is a sufficient statistic for $\zeta(x)$ and M is a normalizing function. (We assume that functions of y only are incorporated in the dominating measure.) The natural parameter space is defined by the condition of integrability of $\exp\{\zeta(x)^T t(Y)\}$ (given x). It is well known that the derivatives of M w.r.t. $\zeta(x)$ provide first and second order moments of $t(Y)$,

$$\nabla M(\zeta(x)) = \left(\frac{\partial}{\partial\zeta_1(x)}M(\zeta(x))\ldots\frac{\partial}{\partial\zeta_k(x)}M(\zeta(x))\right)^T = E(t(Y)|\zeta(x)) =: \tau(x), \tag{8}$$

$$H_M(\zeta(x)) := ((\frac{\partial^2}{\partial\zeta_i(x)\,\partial\zeta_j(x)}M(\zeta(x))))_{i,j=1\ldots k} = cov(t(Y)|\zeta(x)). \tag{9}$$

We can parameterize an exponential family by the natural parameter $\zeta(x)$, the mean $\tau(x)$ or a predictor $\eta(x)$. Depending on the parameterization special representations of the measures of dependence are obtained. We present such results keeping the general notation $\theta(x)$ on the left hand side of an equation and inserting the parameter actually chosen on the right hand size. Thus in exponential families $\theta(x) \in \{\zeta(x), \tau(x), \eta(x)\}$, and each parameter identifies the conditional density of Y given x. Since for exponential families the log-odds ratio function is bi–affine in $\zeta(x)$ and $t(y)$, i.e.

$$\Psi^0(\zeta(x), y) = (\zeta(x) - \zeta(x_0))^T(t(y) - t(y_0)), \tag{10}$$

we obtain the following result.

Result 2:

$$\begin{aligned} J(\theta(X), Y) &= tr\{cov_Y(E(\zeta(X)|Y), t(Y))\} & (11)\\ &= tr\{cov_X(\zeta(X), E(t(Y)|\zeta(X)))\}. \quad \square & (12) \end{aligned}$$

Thus $J(\theta(X), Y)$ measures the covariance of one (transformed) variable and the projection of the other one onto the measurable space that the former spans. Since $J(\theta(X), Y)$ does not depend on the reference values x_0 and y_0, it is often convenient to choose $\zeta(x_0) = \overline{\zeta} := E\zeta(X)$, provided $\overline{\zeta}$ belongs to the natural parameter space. Then one has

$$\begin{aligned} J(\theta(X), Y) &= E_Y[(E(\zeta(X)|Y) - \overline{\zeta})^T t(Y)] & (13)\\ &= E_X[(\zeta(X) - \overline{\zeta})^T E(t(Y)|\zeta(X))]. & (14) \end{aligned}$$

Moreover, it can be shown that in (4) the marginal reference density can be replaced by $p(y|\overline{\zeta})$, that is by a reference density which is also a member of the exponential family. It is mainly due to this property that we emphasize the divergence in comparison to the directed divergence. In particular, if the reference density belongs to the same family of parameterized densities $\{p(y|\theta(x))\}$, expansions in the parameter can be used to obtain approximations of $J(\theta(X), Y)$ and R^2_J. Hence the following representation is a key result providing a link to other definitions of a coefficient of determination in the literature.

RESULT 3:

$$J(\theta(X), Y) = EJ_{KL}(p(y|\zeta(X)), p(y|\overline{\zeta})). \qquad \Box \tag{15}$$

For example, in generalized linear regression, assuming that $p(y|\theta(x))$ belongs to a one-parameter exponential family ($k = 1$), a linear predictor

$$\eta(x) = \alpha + a(x)^T\beta$$

is defined and the regression model is specified by a one–to–one link function g relating $\tau(x) = E(t(Y)|\zeta(x))$ to the predictor by

$$g(\tau(x)) = \eta(x).$$

In this case $\zeta(x) = \zeta(x, \gamma)$ with $\gamma = (\alpha, \beta)$, $\beta \in \mathbb{R}^p$. As a special case the canonical link g_0 maps the expected value to the natural parameter

$$g_0(\tau(x)) = \zeta(x) = \alpha + a(x)^T\beta.$$

If the vectors $a(x)^T$ are centered such that $E\, a(X)^T = 0$, $\overline{\zeta} = \alpha$. In this case the reference density represents a regression model where the conditional distribution of Y does not depend on x.

2.3 Examples

2.3.1 Gaussian response

All definitions of a coefficient of determination originate from the case of Normal distributions with homogeneous covariance. Assume

$$Y|\theta(x) \sim N(\mu(x), \Sigma), \quad Y \in \mathbb{R}^r.$$

If Σ is known, $p(y|\theta(x))$ can be regarded as an exponential family with canonical parameter $\zeta(x) = \Sigma^{-1}\mu(x)$ and $t(y) = y$. One obtains

$$J(\theta(X), Y) = tr\{\Sigma^{-1} cov\, \mu(X)\}.$$

The resulting R_J^2 is related to the decomposition of covariance

$$cov\ t(Y) = E_X cov(t(Y)|\zeta(X)) + cov_X E(t(Y)|\zeta(X)) \tag{16}$$

which we abbreviate as

$$T = W + B \tag{17}$$

alluding to total, within and between covariance.
For the Normal distribution one has $W = \Sigma$, $B = cov\, \mu(X)$ yielding

$$J(\theta(X), Y) = tr\{W^{-1}B\} \tag{18}$$

and

$$R_J^2 = \frac{tr\{W^{-1}B\}}{1 + tr\{W^{-1}B\}}. \tag{19}$$

A special case of a conditional Normal response Y with homogeneous covariance is given if (X, Y) are jointly Gaussian,

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim N(\begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \begin{pmatrix} \Sigma_X & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_Y \end{pmatrix}).$$

Then $Y|x \sim N(\mu(x), \Sigma)$ where

$$\begin{aligned} \mu(x) &= E(Y|x) = \mu_Y + \Sigma_{YX}\Sigma_X^{-1}(x - \mu_X), \\ \Sigma &= \Sigma_Y - \Sigma_{YX}\Sigma_X^{-1}\Sigma_{XY}. \end{aligned}$$

The corresponding decomposition (17) is given by

$$\Sigma_Y = \Sigma + \Sigma_{YX}\Sigma_X^{-1}\Sigma_{XY}$$

where $\Sigma_{YX}\Sigma_X^{-1}\Sigma_{XY} = B = E((\mu(X) - \mu_Y)(\mu(X) - \mu_Y)^T)$. In this case $Y \sim N(\mu_Y, \Sigma_Y)$ and

$$J(\theta(X), Y) = tr(\Sigma^{-1}B) = \sum_{i=1}^{r} \frac{\rho_i^2}{1 - \rho_i^2} \tag{20}$$

(see [22]: Kullback, 1968, p.203), where ρ_i^2 denotes the i–the squared canonical correlation coefficient. If Y is univariate ($r = 1$), this reduces to

$$J(\theta(X), Y) = \frac{\rho^2}{1 - \rho^2}, \tag{21}$$

where ρ^2 is the multiple correlation coefficient of Y with X. Thus

$$R_J^2 = \rho^2, \tag{22}$$

with ρ^2 being further reduced to $corr(X, Y)$ if also X is univariate. Hence R_J^2 does generalize the conventional quantities of interest in the case where X and Y are jointly Gaussian. Actually (21) motivates the transformation (5) of J to R_J^2 introduced in definition 1.

2.3.2 Multinomial response

Assume $\widetilde{Y}$ takes values in $\{0,\dots,k\}$ with probabilities π_i. $\widetilde{Y}$ may be represented by a k-dimensional vector of binary (dummy) variables $\widetilde{Y}_i$ with $\pi_i = P(\widetilde{Y}_i = 1)$. Thus identify $\widetilde{Y} = (\widetilde{Y}_1,\dots,\widetilde{Y}_k)^T$. Modelling π_i as $\pi_i(x)$ depending on values of covariates X and summarizing the results of $n(x)$ trials in $Y = (Y_1 \dots Y_k)$, Y conditionally follows a multinomial distribution

$$Y|\pi(x) \sim M(n(x), \pi(x))$$

where $\pi(x) = (\pi_1(x),\dots,\pi_k(x))^T$. The multinomial distributions form a k–parameter exponential family with

$$\begin{aligned}\zeta(x) &= (\zeta_1(x),\dots,\zeta_k(x)),\\ \zeta_i(x) &= \log\frac{\pi_i(x)}{\pi_0(x)},\end{aligned}$$

where $\pi_0(x) = 1 - \sum_{j=1}^k \pi_i(x)$. Furthermore one has

$$\begin{aligned}&t(y) = y,\\ &E(Y_i|\zeta(x)) = n(x)\pi_i(x) = \tau_i(x), \quad var(Y_i|\zeta(x)) = n(x)\pi_i(x)(1-\pi_i(x)),\\ &cov(Y_i, Y_j|\zeta(x)) = -n(x)\pi_i(x)\pi_j(x) \quad for\, i \neq j.\end{aligned}$$

¿From (12) with $\log\pi(x) = (\log\pi_1(x),\dots,\log\pi_k(x))^T$ and 1_k denoting the k–dimensional vector of ones,

$$\begin{aligned}J(\theta(X), Y) &= tr\{cov_X(\zeta(X), E(t(Y)|\zeta(X)))\}\\ &= tr\{cov(\log\pi(X) - \log\pi_0(X)1_k, n(X)\pi(X))\}. \qquad (23)\end{aligned}$$

For binomial distributions with all $n(x) = m$, i.e. $Y|\pi(x) \sim B(m, \pi(x))$, the expression simplifies to

$$J(\theta(X), Y) = m\; cov(logit\,\pi(X), \pi(X)) \qquad (24)$$

where *logit* $\pi(x) = \log[\pi(x)/(1-\pi(x))]$.

Model 1 (linear logistic regression, canonical link)

If $\zeta(x) = \eta(x) = \alpha + \beta^T a(x)$ where now $\alpha \in \mathbb{R}^k$, $\beta = (\beta_1 \dots \beta_k) \in \mathbb{R}^{p\times k}$ and $E(a(X)) = 0$ as before we obtain from (14)

$$\begin{aligned}J(\theta(X), Y) &= E((\zeta(X) - \overline{\zeta})^T\tau(X))\\ &= E(a(X)^T\beta\tau(X))\end{aligned}$$

with

$$\tau_i(x) = n(x)\frac{\exp\{\alpha_i + \beta_i^T a(x)\}}{1 + \sum_{j=1}^k \exp\{\alpha_j + \beta_j^T a(x)\}}.$$

For univariate Y and X with $n(x) = 1$, $a(x) = x$, this reduces to

$$\begin{aligned} J(\theta(X), Y) &= cov(\alpha + \beta X, \pi(X)) \\ &= \beta\, cov(X, \pi(X)). \end{aligned} \tag{25}$$

For small $\sigma_X^2 = varX$ one may further use the approximation $\pi(x) \approx \pi(x_0) + \pi'(x_0)(x - x_0)$ at $x_0 = 0$ which yields $J(\theta(X), Y) \approx \pi(x_0)(1 - \pi(x_0))\beta^2\sigma_X^2$, and therefore

$$R_J^2 \approx \frac{\beta^2\sigma_X^2}{[\pi(x_0)(1 - \pi(x_0))]^{-1} + \beta^2\sigma_X^2}. \tag{26}$$

For $\sigma_X \to 0$ one obtains $R_J^2 \to 0$ for all β. On the other hand, if σ_X is fixed, $\beta \to \infty$ yields $R_J^2 \to 1$ and $\beta \to 0$ yields $R_J^2 \to 0$. Also for large variances σ_X^2, R_J^2 may take the extreme values 0 and 1. Thus for the logistic model R_J^2 naturally depends only on the coefficients of the regression function and the variability of X.

In contrast, for Normally distributed responses (cp. (19))

$$R_J^2 = \frac{\beta^2\sigma_X^2}{\sigma^2 + \beta^2\sigma_X^2} = \frac{\beta^2\sigma_X^2/\sigma^2}{1 + \beta^2\sigma_X^2/\sigma^2} \tag{27}$$

where $\sigma^2 = var(Y|x)$. Thus R_J^2 depends on β, the variability of X *and* σ^2. R_J^2 becomes 1 if $\beta \to \infty$ or the ratio σ_X^2/σ^2 increases. For fixed σ_X^2 the coefficient of determination depends on β^2/σ^2.

Model 2 (linear logistic regression, probit link)

For notational convenience consider the Binomial case ($k = 1$) with

$$Y|\pi(x) \sim B(n(x), \pi(x))$$

and assume

$$\pi(x) = \Phi(\eta(x)),$$

where Φ denotes the cumulative distribution function of a standard univariate Gaussian distribution. Then

$$\zeta(x) = logit\, \pi(x) = logit\, (\Phi(\eta(x))),$$

$$\overline{\zeta} = E\zeta(X) \neq logit\, \Phi(E\eta(X)) = logit\, \Phi(\alpha)$$

However, there exists a predictor η_0 such that $J_{KL}(p(y|\zeta(x)), p(y|\overline{\zeta}))$ equals $J_{KL}(p(y|\zeta(x)), p(y|logit(\Phi(\eta_0))))$.

More complex models induced by sophisticated link functions are specified by Fahrmeir and Tutz ([12]: 1994, ch.3).

2.3.3 Binary response Y with Gaussian covariate X

A simple distributional model for (Y, X) illustrates how $J(\theta(X), X)$ or R_J^2 capture the association between a dichotomous variable $Y \in \{0, 1\}$ and a continuous variable X. Let

$$X|Y = 1 \sim N(\mu_1, \sigma^2), \quad X|Y = 0 \sim N(\mu_0, \sigma^2)$$

and the marginal distribution be fixed by $p(1) = P(Y = 1)$, $p(0) = 1 - p(1)$ with $p(1) \in (0, 1)$. For simplicity $E(X) = 0$ is assumed which is equivalent to $\mu_0 = -\mu_1 p(1)/p(0)$.

The regression of X on Y has the form $E(X|y) = \beta_{0y} + \beta_y y$ with $\beta_{0y} = -\mu_1 p(1)/p(0)$, $\beta_y = \mu_1/p(0)$. Treating the Normal distributions as one-parameter family with fixed σ^2, the reference density $p(x|\overline{\zeta})$ is normal with $N(0, \sigma^2)$.

The logit regression model which reversely regresses Y on X is known to be linear. For

$$logit\, \pi(x) = \beta_{0x} + \beta_x x$$

one obtains $\beta_{0x} = (\mu_0^2 - \mu_1^2)/(2\sigma^2) + log(p(1)/p(0))$ and $\beta_x = \mu_1/(p(0)\sigma^2)$.

Application of (11) shows that

$$J(\theta(X), Y) = \frac{\mu_1^2}{\sigma^2} \frac{p(1)}{p(0)} \mu_1^2, \quad R_J^2 = \frac{p(1)\mu_1^2}{p(0)\sigma^2 + p(1)\mu_1^2}.$$

The extreme case $R_J^2 = 0\,(J(\theta(X), Y) = 0)$ is obtained for $\mu_1^2 = 0$ meaning independence of X and Y. The case $R_J^2 = 1\,(J(\theta(X), Y) \to \infty)$ is obtained for $\sigma^2 = 0$ or $\mu_1 \to \infty$. If $\mu_1 \to \infty$ the regression parameters β_x as well as β_y tend to infinity meaning that the conditional distributions $p(y|x)$ as well as $p(x|y)$ are maximally separated. For $\sigma^2 = 0$ the regression parameter of the logit model β_x tends to infinity whereas the separation of the distributions $p(x|1)$ and $p(x|0)$ does not show up in the regression parameters.

$J(\theta(X), Y)$ and R_J^2 are symmetric measures of dependence between X and Y. In the light of regression models they capture how strongly a conditional distribution $p(y|x)$ varies with x. Here, in terms of β_x the coefficient R_J^2 is given by

$$R_J^2 = \frac{p(1)p(0)\beta_x^2}{1/\sigma^2 + p(1)p(0)\beta_x^2}.$$

Reversely, when regressing X on Y the distance between $p(x|1)$ and $p(x|0)$ is reflected what is seen from

$$R_J^2 = \frac{p(1)p(0)\beta_y^2}{\sigma^2 + p(1)p(0)\beta_y^2}.$$

R_J^2 may also be interpreted as the proportion of explained variance of X: $\beta_y = \mu_1/p(0)$ immediately yields

$$R_J^2 = \frac{\mu_1^2 p(1)/p(0)}{\sigma^2 + \mu_1^2 p(1)/p(0)},$$

and $varX = \sigma^2 + \mu_1^2 p(1)/p(0)$. While this interpretation holds for the regression of X or Y it does *not* apply to the decomposition of $varY$ when regressing Y on X. Yet the divergence quantifies the explanatory value of one of the variables for the other variable because of the symmetry in X and Y.

3 Measures of determination

3.1 $I(\theta(X), Y)$ as alternative to $J(\theta(X), Y)$

Similar concepts might be suggested measuring the distance of $p(y|\theta(x))$ to a reference density differently. Especially the mutual information $I(\theta(X), Y)$ (given in (1)) as an average of the directed divergences

$$I_{KL}(p(y|\theta(x)), p_\theta(y)) = E(\log \frac{p(Y|\theta(x))}{p_\theta(y)} |\theta(x))$$

may be an appealing alternative to $J(\theta(X), Y)$ ([36]: Soofi et al., 2000; [19]: Joe, 1989; [20]: Kent, 1983). Analogously to the decomposition of variance $I(\theta(X), Y)$ describes the decomposition of entropy:

$$I(\theta(X), Y) = H(Y) - H(Y|\theta(X)), \tag{28}$$

where $H(Y) = -E(\log p_\theta(y))$ and $H(Y|\theta(X)) = E_X[-E\{\log p(Y|\theta(X))|\theta(X)\}]$ (see e.g. [6]: Cover and Thomas, 1991, p.19).

In case of a bivariate Gaussian distribution of (X, Y) with correlation coefficient ρ

$$I(\theta(X), Y) = -\frac{1}{2}\log(1-\rho^2), \tag{29}$$

([22]: Kullback, 1968, p.203). Then $R_I^2 = \rho^2$ may be derived from the directed divergence by

$$R_I^2 = 1 - \exp(-2I(\theta(X), Y)). \tag{30}$$

The relation also holds for the multiple correlation coefficient if X is multidimensional. The transformation (30) has in fact been suggested in terms of estimates to define a general coefficient of determination ([20]: Kent, 1983; [26]: Magee, 1990; [28]: Nagelkerke, 1991), and R_I^2 can be interpreted as a scaling transformation of $I(\theta(X), Y)$. See also Theil ([40]: 1987).

The main advantage of the symmetrized mutual information $J(\theta(X), Y)$ over the mutual information $I(\theta(X), Y)$ appears to be result 1 yielding result 3, the representation within an exponential family. $EI_{KL}(p(y|\zeta(X)), p(y|\bar{\zeta}))$ and $I(\theta(X), Y)$ usually differ.

3.2 Reference densities

The various reference densities we introduced partly explain the variety of suggestions generalizing the coefficient of determination. $J(\theta(X), Y)$ is defined for a joint density of X and Y but only a partial feature is explored in a regression model. In particular, the (multivariate) distribution of X is often left unspecified. Hence (the type of) the marginal density $p_\theta(y)$ is not determined unless Y is categorical. Also, even if the distribution of X is known, the marginal densities $p_\theta(y) = Ep(y|\theta(X))$ are different for different models. In order to allow for a comparison across models with the same type of sampling distribution therefore a reference density $p(y|\theta_0)$ is used where $p(y|\theta_0)$ is assumed to belong to the same family of densities as $p(y|\theta(x))$ and θ_0 is the specific parameter. $\theta_0 = \theta(x, \gamma_0)$ with $\gamma_0 = (\omega_0, \beta_0)$ and $\beta_0 = 0$ corresponds to a regression model with constant regression function.

DEFINITION 2:
Define

$$R^2_{J,\theta_0} = \frac{EJ_{KL}(p(y|\theta(X)), p(y|\theta_0))}{1 + EJ_{KL}(p(y|\theta(X)), p(y|\theta_0))} \tag{31}$$

to be the *coefficient of determination of* Y *by* X *through* θ *based on* the symmetric discrepancy to the density specified by *a specific parameter* θ_0. $\lhd$

Similarly the KL-discrepancy $I_{KL}(p(y|\theta(X)), p(y|\theta_0))$ might be used to define R^2_{I,θ_0} analogously to (30). The average discrepancy might intuitively have been aimed at by those authors who suggest to empirically define

$$R^2_{LR} = 1 - \exp(-\frac{2}{n} LR) \tag{32}$$

where LR denotes the log–likelihood ratio test statistic to compare the model of interest to a 'null model' which is understood to be that with regression coefficients set to zero (see [28]: Nagelkerke, 1991). A second order Taylor expansion of $\log p(y|\zeta_0)$ in $\zeta(X)$ yields ([22]: Kullback, 1968, pp.26)

$$EJ_{KL}(p(y|\zeta(X)), p(y|\zeta_0)) \approx 2EI_{KL}(p(y|\zeta(X)), p(y|\zeta_0)), \tag{33}$$

and to this extent the coefficients of determination R^2_{J,θ_0} and R^2_{I,θ_0} work similarly.

Based on $I(\theta(X), Y) = H(Y) - H(Y|\theta(X))$ an alternative to using a reference other than $p_\theta(y)$ is to consider the difference of entropies $H(Y|\theta_0) - H(Y|\theta(X)$. Model comparison then refers to $H(Y|\theta(X))$ and might be extended to models which are not related being members of a common family of sampling distributions.

3.3 Approximations in exponential families

A second order Taylor expansion of $\log p(y|\zeta(x))$ in $\overline{\zeta}$ links the symmetrized mutual information $J(\theta(X), Y) = EJ_{KL}(p(y|\zeta(X)), p(y|\overline{\zeta}))$ in exponential families to the matrix $H_M(\overline{\zeta})$. Kullback ([22]: 1968, pp.26f) shows the following result.

Result 4:

$$J(\theta(X), Y) \approx tr\{H_M(\overline{\zeta}) cov(\zeta(X))\}. \qquad \square \tag{34}$$

A similar result is obtained using a reference density $p(y|\zeta_0)$ and expanding in ζ_0. This result relates our definition of R_J^2 to the one based on Wald's test for testing $H_0 : \beta = 0$ against $H_1 : \beta \neq 0$ or $H_0 : \zeta(x) = \zeta_0$ against $H_1 : \zeta(x) \neq \zeta_0$ for all x. For a fixed x the component $w(x)$ of a test statistic with $\widehat{\zeta}$ estimating ζ is

$$w(x) = (\widehat{\zeta}(x) - \zeta_0)^T H_M(\zeta_0)(\widehat{\zeta}(x) - \zeta_0).$$

Then averaging over x yields

$$w = E[tr\{H_M(\zeta_0)(\widehat{\zeta}(X) - \zeta_0)(\widehat{\zeta}(X) - \zeta_0)^T\}].$$

An empirical coefficient of determination of the type

$$R_w^2 = \frac{\widehat{w}}{1 + \widehat{w}} \tag{35}$$

was suggested and investigated for several regression models, listed by Magee ([26]: 1990).

A link to the interpretation of the coefficient of determination as the proportion of the total variance explained by regression may be based on the approximation of $J(\theta(X), Y)$ using the mean parameterization.

Result 5:

$$J(\theta(X), Y) \approx tr\{(cov(t(Y)|\overline{\tau}))^{-1} cov_X E(t(Y)|\tau(X))\} \tag{36}$$

where $\bar{\tau} := E\tau(X)$. $\square$

The right hand side of (36) is estimated by the test statistic of the score test which Magee ([26]: 1990) referred to in order to generalize the (estimated) coefficient of determination.

3.4 Decomposition of covariance

The decomposition of covariance (17) yields (18),

$$J(\theta(X), Y) = tr\{W^{-1}B\},$$

in the special case of Gaussian distributions of $Y|\theta(x)$ with homogeneous variance, and (18) also holds approximately according to result 5 if $W \approx cov(t(Y)|\overline{\tau})$. Hence

$$\begin{aligned} J(\theta(X), Y) &\approx tr[(T-B)^{-1}TT^{-1}B] \\ &= tr[(I-T^{-1}B)^{-1}T^{-1}B]. \end{aligned} \tag{37}$$

This equation may be read as an analogue to the defining transformation (5),

$$J(\theta(X), Y) = \frac{R_J^2}{1-R_J^2}.$$

Thus R_J^2 is associated to the *matrix* $T^{-1}B$ though not directly defined by a *functional* of it like $tr(T^{-1}B)$ or $\det(T^{-1}B)$ as previously suggested in the literature (see [27]: Mardia et al., 1995, ch.6.5.4, pp170f).

For example, Hooper ([17]: 1959) refers to the decomposition $I = T^{-1}W + T^{-1}B$ to obtain

$$R_H^2 = \frac{1}{k} tr(T^{-1}B) \tag{38}$$

Glahn ([13]: 1969) uses the decomposition applying the determinant to suggest

$$R_G^2 = |B|/|T|. \tag{39}$$

This measure coincides with γ_1 in Cramer's and Nicewanders review ([7]: 1979) of multivariate measures of association under the assumption of a joint multivariate Gaussian distribution.

Based on the decomposition of covariance various measures have been derived which focus on ratios of (re-scaled) between variances to (re-scaled) total variances. For instance the re-scaled variance ratio (called γ_3 by Cramer and Nicewander ([7]: 1979)

$$\frac{tr\{W^{-1}B\}}{tr\{W^{-1}T\}} = \frac{tr\{W^{-1}B\}}{k + tr\{W^{-1}B\}} = \frac{\frac{1}{k}J(\theta(X), Y)}{1 + \frac{1}{k}J(\theta(X), Y)} \tag{40}$$

is built similarly to our R_J^2. For a univariate response it coincides with R_J^2.

An issue we finally address is the classification of measures of multivariate dependence into (symmetric) measures of association and (asymmetric) measures of redundancy. The measures discussed by Cramer and Nicewander ([7]: 1979) under the assumption of joint Normality are all - like ours - measures of dependence between X and Y. They can be rewritten in information theoretic terms as transformations of $I(X,Y)$, $J(X,Y)$ or based on differences of the entropies $H(Y)$, $H(Y|X)$, $H(E(Y|X))$. They are all functions of the canonical correlation coefficients, and hence X and Y may be interchanged. In contrast the ratio of between to total variance of the standardized response vector Y

$$RI = \frac{tr\{C_{YX}C_X^{-1}C_{XY}\}}{tr\{C_Y\}}, \tag{41}$$

where C denotes a correlation matrix, is known as redundancy index in the literature ([39]: Stewart and Love, 1968; [14]: Gleason, 1976). The redundancy index is not symmetric in X and Y. In case of joint Gaussianity of X and Y it is equal to the average squared multiple correlation coefficient of components of Y given X thus quantifying linear predictability of Y (componentwise, not truly multivariate) given X. It has been felt by many authors that determination in regression should be measured in such an asymmetric way.

Relating to that discussion we justify our study of a symmetric measure of dependence arguing that essentially a measure of redundancy is a scaled measure of dependence. In information theory the redundancy of a variable Z is defined as

$$red(Z) = 1 - \frac{H(Z)}{H_{\max}}, \tag{42}$$

where $H_{\max}$ denotes a context dependent maximum achievable entropy. Lower entropy of Z indicates structure, and the more structure is present the higher the redundancy of Z. (Cp. e.g. the discussion in ([18]: Jessop, 1995, pp.51f). In regression we are interested in reducing the uncertainty about Y by X, measured by $H(Y|\theta(X))$, that is in maximizing the redundancy of $Y|X$. This amounts to being interested in

$$red(Y|\theta(X)) = 1 - \frac{H(Y|\theta(X))}{H(Y)} = \frac{I(\theta(X),Y)}{H(Y)}, \tag{43}$$

the mutual information between X and Y scaled w.r.t. the entropy of Y.

For a bivariate Gaussian distribution of X and Y we obtain with $e = \exp(1)$

$$red(Y|\theta(X)) = -\frac{\log(1-\rho^2)}{\log(2\pi e\sigma_Y^2)} \tag{44}$$

which tends to 0 if $\rho^2 \to 0$, and tends to ∞ if $\rho^2 \to 1$. More generally for a joint Normal distribution

$$I(\theta(X),Y) = \frac{1}{2}\log\frac{|T|}{|W|}.$$

If Y is standardized yielding $T_s = C_Y$ and W_s, say,

$$red(Y|\theta(X)) = \frac{\log|C_Y| - \log|W_s|}{\log(2\pi e)^k|C_Y|}, \tag{45}$$

(where $k = \dim Y$). Thus, although $red(Y|\theta(X))$ and RI do not exactly coincide, the motivating idea is the same. $red(Y|\theta(X))$ is a truly multivariate (w.r.t. Y) index and generalizes to other but Gaussian distributions, particularly discrete distributions. In conclusion we suggest to consider measures of association and dependence as primary, and measures of redundancy as secondary, derived measures, and in this paper we discuss the decomposition of variance as a special Gaussian representation of association. Furthermore, although measures of dependence like $J(\theta(X), Y)$ are symmetric, the specification of the conditional density $p(y|\theta(x))$ is based on a regression model expressing ideas about a directed influence of X on Y.

3.5 Local coefficients of determination

Sometimes it is felt that the strength of the relation between Y and X varies with x, and a local rather than a global coefficient of determination is wanted. For example, Doksum et al. ([9]: 1994) resume an example given by Härdle ([15]: 1990) where Y= expenditure for food depending on X= net income is analyzed. The higher the income the weaker the relation is, and this structure is captured by Doksum et al. ([9]: 1994) in a 'correlation curve'. We discuss several concepts of local coefficients of determination.

3.5.1 Correlation curves

The definition of a (squared) correlation curve aims at a local 'decomposition of variance'. Doksum et al. ([4]: 1993; [9]: 1994) suggested for univariate X and $t(Y)$ with a conditional distribution in an exponential family

$$\rho^2(x) = \frac{\tau'(x)^2\sigma_X^2}{var(t(Y)|\theta(x)) + \tau'(x)^2\sigma_X^2}, \tag{46}$$

where $\sigma_X^2 = varX$ and $\tau'(x)$ denotes the derivative of $\tau(x)$. Actually Doksum et al. ([9]: 1994) motivate their definition based on a linear model by

$$R_J^2 = \frac{var(\alpha + \beta X)}{var(\alpha + \beta X + \epsilon)} = \frac{\beta^2\sigma_X^2}{var\epsilon + \beta^2\sigma_X^2},$$

where β is replaced by $\tau'(x)$ and $var\epsilon$ by the heterogeneous variances $\sigma^2(x) = var(t(Y)|\theta(x))$ depending on x. If $\tau'(x)$ and $\sigma^2(x)$ are constant, the correlation coefficient ρ^2 is re-obtained in bivariate Gaussian regression. For multivariate $X \in \mathbb{R}^s$ (46) is generalized to

$$\rho^2(x) = \frac{(\nabla\tau(x))^T\Sigma_X\nabla\tau(x)}{\sigma^2(x) + (\nabla\tau(x))^T\Sigma_X\nabla\tau(x)} \tag{47}$$

3.5.2 Determination curves

A local measure of the strength of relation between X and Y may be based on the *rate of change*. Following Blyth ([3]: 1994) consider local discrepancies

$$J(x, \Delta x) = J_{KL}(p(y|\theta(x)), p(y|\theta(x + \Delta x)) \tag{48}$$

for *continuous* X. For the limiting local discrepancy one has based on a second order Taylor expansion

$$\begin{aligned} J_0(x) &= \lim_{\delta \to 0} \frac{J(x, \Delta x)}{\delta^2} & (49) \\ &= \sum_{i,j} I_{i,j}(x) = 1_s^T I(x) 1_s, & (50) \end{aligned}$$

where $I(x)$ is the $s \times s$-dimensional Fisher information matrix defined by

$$I_{i,j}(x) = E[\frac{\partial}{\partial x_i} \log p(Y|\theta(x)) \frac{\partial}{\partial x_j} \log p(Y|\theta(x))|\theta(x)].$$

If $p(y|\theta(x))$ belongs to a $k-$parameter exponential family,

$$I(x) = \nabla_{(k)}\zeta(x)(cov(t(Y)|\zeta(x))(\nabla_{(k)}\zeta(x))^T, \tag{51}$$

where $\nabla_{(k)}\zeta(x)$ is the $s \times k$ matrix $((\frac{\partial \zeta_\kappa}{\partial x_i}))_{\kappa=1\ldots k, i=1\ldots s}$. The Fisher information is well known to be an indicator of the sensitivity of (the conditional distribution of) Y to changes in the parameter $\theta(x)$. See the discussion by Rao ([30]:1973, pp.331f). More explicitly, for example, if X is univariate ($s = 1$),

$$I(x) = E[(\frac{d}{dx} \log p(Y|\theta(x)))^2|\theta(x)],$$

and hence under regularity conditions

$$I(x) = var[\frac{d}{dx} \log p(Y|\theta(x))|\theta(x)]. \tag{52}$$

$I(x)$ describes the rate of change (on the log scale) of the density induced by a change of the parameter (depending on x) and thus refers to the interpretation of association in terms of discrimination.

For illustration consider again the example of logistic regression discussed in section 2.3.2 (model 1 and $n(x) = 1$). Then with $\pi(x) = \tau(x)$

$$\tau(x) = \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)},$$

$$\tau'(x) = \frac{\beta \exp(\alpha + \beta x)}{(1 + \exp(\alpha + \beta x))^2}$$

and $var(Y|\zeta(x)) = \tau(x)(1-\tau(x))$,

$$J_0(x) = \frac{\tau'(x)^2}{\tau(x)(1-\tau(x))} = \beta\tau'(x)$$

(cp. (26)). Hence $J_0(x) = I(x)$ converges to 0 for x tending to $\pm\infty$ and thus reflects the asymptotes of the sigmoid logistic curve $\tau(x)$.

As Blyth ([3]: 1994) pointed out, in a one-parameter exponential family with univariate X the correlation curve appears to be a one-to-one transformation of $J_0(x)$. If $p(y|\theta(x))$ is from a one-parameter exponential family (where X may be again multivariate), $J_0(x)$ takes the form of a signal-to-noise ratio

$$J_0(x) = \frac{(\nabla\tau(x))^T\nabla\tau(x))}{var(t(Y)|\theta(x))}. \tag{53}$$

If additionally $s = 1$ the transformation of $J_0(x)$,

$$R^2_{\sigma_X J_0}(x) = \frac{\sigma_X^2 J_0(x)}{\sigma_X^2 J_0(x) + 1} \tag{54}$$

yields the correlation curve $\rho^2(x)$. The scaling by σ_X^2 in (54) is chosen in order to recover ρ^2 for bivariate Gaussian (X, Y), where $\tau'(x)$ and $\sigma^2(x)$ and hence $J_0(x)$ are constant functions.

In terms of $I(x)$ the correlation curve $\rho^2(x)$ for $s = k = 1$ is also given by

$$\rho^2(x) = \frac{\sigma_X^2 I(x)}{\sigma_X^2 I(x) + 1} \tag{55}$$

which can be generalized to $s \geqq 1$, $k \geqq 1$.

Definition 3:
Define

$$R_J^2(x) = \frac{tr\{I_\theta(x)\Sigma_X\}}{tr\{I_\theta(x)\Sigma_X\} + 1} \tag{56}$$

to be the value of the *determination curve for Y in x through θ*, where $I_\theta(x)$ denotes the Fischer information matrix derived from $p(y|\theta(x))$. $\lhd$

According to (51) in a $k-$parameter exponential family

$$\begin{aligned} tr\{I(x)\Sigma_X\} &= tr\{[cov(t(Y)|\tau(x))]^{-1}(\nabla\tau(x))^T\Sigma_X\nabla\tau(x)\} \\ &= tr\{[W(x)]^{-1}B(x)\} \qquad \text{(say).} \end{aligned}$$

Hence (47) is obtained as a special case where $s \geqq 1 = k$, and the family of $N(\tau(x), \sigma^2(x))$-distributions is treated as a one-parameter exponential family. The curves for a Gaussian response, where the density is assumed to belong to a two-parameter exponential family, will be discussed for a numerical example in section 6.3.

3.5.3 Certainty curves

If X is not continuous but e.g. nominal, concepts of local variation and differentiation do not apply. Yet the influence of X at x can be assessed as the reduction of uncertainty about Y measured in terms of entropies. Set $H(Y|\theta(x)) = -E_{Y|\theta(x)}(\log p(y|\theta(x)))$.

DEFINITION 4:
Define

$$R_I^2(x) = 1 - \exp\{-2[H(Y) - H(Y|\theta(x))]\} \tag{57}$$

to be the *coefficient of certainty about Y given x through θ*. The resulting curve as a function of x is called the *certainty curve.* $\triangleleft$

Definition 4 may also be applied to continuous X with differential entropies if these exist, i.e. if the integrals are finite. For example, if (X, Y) are bivariate Gaussian, $H(Y) = (1/2)\log(2\pi e\sigma_Y^2)$, $H(Y|\theta(x)) = (1/2)\log(2\pi e\sigma^2_{Y|\theta(x)})$, where $\sigma^2_{Y|\theta(x)} = \sigma^2$ does not depend on x. Hence the reduction of uncertainty about Y is the same for all x, and the certainty curve is constant taking the value of the correlation coefficient.

With the asymmetric scaling yielding the redundancy index one may similarly define a *redundancy curve* by

$$red(Y|\theta(x)) = 1 - \frac{H(Y|\theta(x))}{H(Y)}. \tag{58}$$

taking values in $\mathbb{R}^+$. Then $red(Y|\theta(X))$ is the average redundancy.

Locally the difference between the notions of discriminatory power and explanatory power of X related to dependence measures based on J and I respectively, becomes more distinct while globally the idea of variability within the family of densities $\{p(y|\theta(x)\}$ covers both aspects.

4 Properties and use of coefficients of determination

The main idea we elaborate in this paper is the application of association measures under the assumption of a regression model. To this aim we focus on $J(\theta(X), Y)$ and consider to some extent $I(\theta(X), Y)$. A list of postulates, of desirable properties of a measure of association, introduced by Renyi ([32]: 1959) and modified later on e.g. by Bell ([1]: 1962) has been agreed on in the literature. These requirements comprise (i) generality of the definition, (ii) symmetry in X and Y, (iii) normalization and (iv) invariance under 1:1-transformations of

X and Y. The crucial issue yielding refinements and modifications of association measures turned out to be the normalization in two respects:

- Renyi ([32]: 1959) claimed that the measure should coincide with (the absolute value of) the correlation coefficient in case X and Y are bivariate Gaussian. Bell ([1]: 1962) weakened this postulate suggesting that a measure of association should be a monotone function of the (squared) correlation coefficient in that case. In regression analysis Renyi's claim is still virulent, and hence we emphasize standardizing transformations yielding the squared correlation coefficient as a special case.

- However, an additional requirement, namely that a value of a measure of association equal to 1 should indicate functional dependence between X and Y may conflict with the orientation towards correlation. Particularly for nominal random variables therefore different ways of scaling were proposed, for example by Joe ([19]: 1997).

Apart from the properties of association measures additional features of coefficients of determination like additivity are of interest. See e.g. Soofi et al. ([36]: 2000). For correlation curves properties were proven by Bjerve and Doksum ([4]: 1993). Here we only address a few issues.

4.1 Invariance to re-parameterizations

$J_{KL}(p(y|\theta(x)), p_\theta(y))$ is invariant to one-to-one transformations of θ, but result 3, providing a representation within an exponential family with reference density $p(y|\overline{\zeta})$, holds for the natural parameter only. Further, there is a lack of interchangeability of parameterization and expectation w.r.t. X, e.g. $\bar{\tau} \neq \tau(\overline{\zeta})$ and different reference densities may not be equally useful.

4.2 Monotonicity in the number of covariates

Interpreting (directed) divergences as measures of variability of densities $p(y|\theta(x))$ one might expect this variability to increase if more informative covariates are included in the regression model. Indeed, within an encompassing model including all covariates, i.e. with θ being derived from the joint distribution of X and Y, $I(\theta(X), Y)$ is increasing with the number of covariates because of the chain rule for mutual information. Darbellay ([8]: 1998) elaborated for R_I^2 how the coefficient increases if new covariates are added. To our knowledge similarly general results are not available for $J(\theta(X), Y)$.(See also the related discussion for estimates R_w^2 by Magee ([26]: 1990).) Usually, however, (even nested) regression models are not assumed to be related by a joint distribution, and monotonicity due to additivity of information is important. Simon ([35]: 1973) investigated conditions for additivity of discrepancies in exponential families if models are nested.

Assuming $\theta_0 \prec \theta_1 \prec \theta_2$, where $\prec$ indicates nesting, in exponential families

$$I_{KL}(p(y|\theta_0), p(y|\theta_2)) = I_{KL}(p(y|\theta_0), p(y|\theta_1)) + I_{KL}(p(y|\theta_1), p(y|\theta_2)) \tag{59}$$

holds under the condition of orthogonality

$$(\theta_1 - \theta_2)^T(\tau(\theta_0) - \tau(\theta_1)) = 0. \tag{60}$$

For example, if $Y|\theta(x) \sim N(\mu(x), \Sigma)$, $\theta(x) = \mu(x)$, for $x = (x_1, x_2)$, $\theta_0 = \alpha$, $\theta_1(x) = \theta_0 + a_1(x_1)^T\beta_1$, $\theta_2(x) = \theta_1(x) + a_2(x_2)^T\beta_2$ the condition (60) amounts to the requirement

$$\beta_2^T a_2(x_2)^T a_1(x_1)\beta_1 = 0$$

which is met if $a_1(x_1)$ is orthogonal to $a_2(x_2)$.

Soofi and Retzer ([37]: 2002, p.16) summarize results on additivity of information indices for exponential families.

4.3 Use of a coefficient of determination in model comparison

Coefficients of determination as measures of dependence between X and Y are used to quantify a feature of a regression model under consideration. In the comparison of models typically two situations are distinguished:

(1) An encompassing model is given by the joint density of Y and all covariates X. Submodels are derived within the encompassing model, for example $p(y|\theta_{(s)}(x_{(s)}))$ for subsets $X_{(s)}$ of the covariates. Often then the minimally comprehensive model is found that is still close enough to the most comprehensive model (in terms of its 'modelling potential'). An assessment of the potential of a regression model is especially of interest in variable selection where a parsimonious model is looked for. Indeed, variable selection has been a main field of application of coefficients of determination. R_I^2 is appropriate in this case, and monotonicity in the number of covariates applies.

(2) Models corresponding to $p(y|\theta(x))$ are not assumed to be related within an encompassing model, but the conditional densities are assumed to belong to a specified family of distributions. In this case $p_\theta(y)$ varies with the regression model, and $p(y|\theta_0)$ is chosen as a reference density instead. Hence the comparison of models is based on R^2_{I,θ_0} or R^2_{J,θ_0}, and monotonicity may hold for nested models. The discrepancies corresponding to R^2_{I,θ_0} or R^2_{J,θ_0} are also natural quantities to test associated hypotheses $H_0 : \theta(x) = \theta_0$ against $H_1 : \theta(x) \neq \theta_0$ about regression parameters. These tests aim at the significance of a difference in the explanatory power of models specified by $\theta(x)$ or θ_0.

Because of the monotonicity properties coefficients of determination are measures of the explanatory potential of a model rather than adequate criteria for model choice aiming at prediction of future observations which typically compromise between data fit and model complexity. Model comparisons based on coefficients of determination apply to models related by specifications according to either (1) or (2). They are useful to find a parsimonious relative explanatorily powerful model within a set of related models. Monotonicity of a coefficient of determination supports strategies of forward selection or backward elimination in variable selection.

Often an internal scaling of coefficients of determination is desired in model comparison w.r.t. the highest value attained. Although for instance R^2_{J,θ_0} is already a scaling transformation of $EJ_{KL}(p(y|\theta(X)), p(y|\theta_0))$ taking values between 0 and 1, across models $R^2_{J,\theta_0} = 1$ may not be attainable within the set of models to be compared. In practice therefore often estimates $\widehat{R}^2_J$ are internally scaled w.r.t. e.g. the most comprehensive model ([11]: Draper and Smith, 1981, p.42; [28]: Nagelkerke, 1991) and used as relative measures for the assessment of models in an informal way.

5 Estimation of a coefficient of determination

We already occasionally referred to estimates of coefficients of determination but we address the topic of estimation in a more systematic and explicit way in this section. As often coefficients of determination (and measures of association) are defined descriptively, in fact estimates of quantities of interest are suggested, and frequently 'corrections' or 'modifications of the definition' of such coefficients are meant to improve properties of estimators. For example, Särndal ([33]: 1974) in his review discusses at length 'corrections' that yield unbiased estimators of association measures. We investigate estimates in special examples in section 6.

5.1 Maximum Likelihood-Estimation

An immediate approach in order to obtain estimates is to substitute parameters θ by their maximum likelihood (ml) estimates $\widehat{\theta}$. For example, for $I(\theta(X), Y) = H(Y) - H(Y|\theta(X))$ ranking of submodels according to R^2_I reduces to ranking of the quantities $H(Y|\theta(X))$ estimated by $\widehat{H}(Y|\theta(X)) = H(Y|\widehat{\theta}(X))$. Within an encompassing model these estimates are monotone for submodels with an increasing number of covariates.

A crucial result is obtained for the ml-estimates of $EI_{KL}(p(y|\theta(X)), p(y|\theta_0))$. Under conditions frequently met in generalized regression (cp. [35]: Simon, 1973) $I_{KL}(p(y|\hat{\theta}(x_i)), p(y|\hat{\theta}_0))$ coincides with the log-likelihood ratio $\log p(y_{ij}|\widehat{\zeta}(x_i)) - \log p(y_{ij}|\widehat{\zeta}_0)$. (59) carries over to likelihood ratios (e.g. [31]: Rao and Toutenburg, 1995, p.48). Model comparison with estimated quantities then turns out to be based on the log-likelihood with estimated parameters. Thus ml-estimation renders an empirical coefficient of determination which is simply a difference of measures of goodness of fit. This is the reason why estimated coefficients of determination are used for comparing models w.r.t. their explanatory power for a given data set but have been abandoned as criteria for model choice aimimg at prediction.

In exponential families $\hat{R}^2_{J,\theta_0}$ can be derived from

$$J_{KL}(p(y|\hat{\theta}(x)), p(y|\hat{\theta}_0)) = (\widehat{\zeta}(x) - \widehat{\zeta}_0)^T(\tau(\widehat{\zeta}(x)) - \tau(\widehat{\zeta}_0)). \tag{61}$$

Similarly R^2_{J,θ_0} may be estimated inserting ml-estimates in the test statistic of the Wald- or score test (see results 4 and 5).

The ml-estimates of the discrepancies and log-likelihood ratios in exponential families are close but do not always coincide: equality requires additivity in nested model which does not always hold. Intuitively, particularly if the conditional densities $p(y|\theta(X))$ do not belong to an exponential family, the data are used twice in the log-likelihood ratio as an estimate of a KL-discrepancy: for estimating θ and for the evaluation of the integral.

5.2 Non–parametric estimation

Estimates can also be obtained if no parametric model or no function space is explicitly specified for $\theta(x)$. For example $\hat{\tau}(x)$ may be obtained from kernel estimation yielding an estimate $\hat{\theta}(x)$. Doksum and Samarov ([10]: 1995) and Lavergne and Vuong ([23]: 1998) investigated nonparametric estimates to obtain a coefficient based on a decomposition of variance.

5.3 Bayesian estimation

Within the Bayesian approach part of the model specification is a prior distribution for $\gamma = (\omega, \beta)$. Instead of plugging in a point estimate $\widehat{\theta}$ or $\widehat{\gamma}$ the parameter θ or γ may be integrated (averaged) w.r.t. a posterior distribution on the parameter space and Bayesian estimates like the posterior mean or the posterior mode may be used. Then the approach may also share features due to the additivity of information in nested models. Also, given a (joint) distribution of (X, Y, Θ) in a Bayesian approach it may be more natural to focus on the dependence of Y and X with θ integrated out than to investigate the posterior distribution of a measure of dependence given θ like $EJ_{KL}(p(y|\theta(X)), p(y|\theta_0))$ or $-H(Y|\theta(X))$. For a Bayesian approach to hypothesis testing see ([2]: Bernardo and Rueda, 2002).

6 Case studies

6.1 Standard regression models with Gaussian response

We evaluate discrepancies to a reference density as quantities of interest for a Gaussian response Y including regression models with heterogeneous variances.

Various estimates for the quantities of interest are discussed that were previously introduced as modifications of the *definition* of a coefficient of determination. Finally we investigate model comparisons for different regression functions including nested models.

Consider an outcome Y which is observed at only $q = 3$ values x_i of a covariate X and assume

$$Y(x_i) \sim N(\mu(x_i), \sigma^2), \qquad i = 1, 2, 3.$$

For illustration we simulated data according to designs $\begin{pmatrix} x_1 & x_2 & x_3 \\ n_1 & n_2 & n - n_1 - n_2 \end{pmatrix}$, such that $p_i = n_i/n$ empirically determines a distribution of X. Thus we observe $Y_j(x_i)$ for $j = 1...n_i$, $i = 1, 2, 3$. We chose $x_1 = 0.2$, $x_2 = 0.5$, $x_3 = 0.9$. First we discuss the simple linear regression model

$$\mu_1(x_i) = \alpha_1 + \beta_1(x_i - \overline{x}), \tag{62}$$

where $\overline{x} = \sum p_i x_i$, for $\alpha_1 = 1$, $\beta_1 = 1.2$ and $\sigma^2 = 0.16$.

6.1.1 Models with non-random coefficients and homogeneous variances

Quantities of interest Considering a univariate response variable Y we write σ_W^2 (σ_B^2) for the within (between) variance instead of the covariance matrices W (B) introduced in section 2.3. Assuming homogeneous variances we have $\sigma_W^2 = \sigma^2$. Having specified a set-up with conditional sampling the marginal density $p_\theta(y)$ is not estimable and cannot be used as reference density. Instead, we choose a reference density $p(y|\theta_0)$ which is Normal with mean μ_Y and variance σ_Y^2. Thus it is not claimed that $p_\theta(y)$ is Gaussian but that its first two moments might be estimated and are to be used as parameters of a Gaussian distribution.

(i) Using the reference density $p(y|\zeta_0)$ where $\zeta_0^T = (\mu_Y/\sigma_Y^2, -\sigma_Y^{-2}/2)$, we obtain

$$EJ_{KL}(p(y|\zeta(X)), p(y|\zeta_0)) = \sigma_B^2/\sigma_W^2, \tag{63}$$

and thus

$$R^2_{J,\zeta_0} = 1 - \sigma_W^2/\sigma_Y^2. \tag{64}$$

(ii)

$$EI_{KL}(p(y|\zeta(X)), p(y|\zeta_0)) = \frac{1}{2}\log(\sigma_Y^2/\sigma_W^2) = -\frac{1}{2}\log(1 - \sigma_B^2/\sigma_Y^2)$$

and

$$R^2_{I,\zeta_0} = \sigma_B^2/\sigma_Y^2 = R^2_{J,\zeta_0}.$$

We observe that given σ_Y^2 all discrepancies are decreasing functions of σ_W^2 and so are the coefficients of determination as monotone transformations.

There has been some discussion in the literature about the fact that $R^2_{J,\zeta_0} < 1$ because $\sigma^2_W > 0$. In order to reduce this effect some authors (e.g. Healy ([16]: 1984) suggested to refer in model comparison to the dependence between $\overline{Y_i} = (\sum_{j=1}^{n_i} Y_j(x_i))/n_i$, $\overline{Y_i} \sim N(\mu(x_i), \sigma^2/n_i)$ and X, where X is assumed to be uniformly distributed with $p_i = 1/q$. Thus the response variable depends on the design. In model comparisons based on the same data set this may be an option although an internal scaling would also do.

Estimation In order to estimate R^2_{J,ζ_0} estimates of variance may be inserted. The data yield an estimate $\widehat{\sigma^2_Y}$ while the underlying distribution of Y is unknown. From a set of proposed regression models then the one providing a decomposition of $\widehat{\sigma^2_Y}$ such that the dependence between X and Y is maximum, is chosen. That amounts to minimizing $\widehat{\sigma^2_W}$, interpretable as maximizing goodness of fit. A crucial issue when using variance estimates is whether the estimates satisfy $\widehat{\sigma^2_Y} = \widehat{\sigma^2_B} + \widehat{\sigma^2_W}$. For a given data set $\{y_j(x_i)|i = 1, ..., q; j = 1, ..., n_i\}$ the decomposition of the sum of squares based on least squares estimates $\widehat{\mu_i}$ and $\overline{y} = (\sum_{i,j} y_j(x_i))/n$,

$$\sum_{i,j}(y_j(x_i) - \overline{y})^2 = \sum_{i,j}(y_j(x_i) - \widehat{\mu_i})^2 + \sum_i n_i(\widehat{\mu_i} - \overline{y})^2$$

ensures the decomposition for the ml-estimates

$$\widehat{\sigma^2_Y} = \frac{1}{n}\sum_{i,j}(y_j(x_i) - \overline{y})^2, \quad \widehat{\sigma^2_W} = \frac{1}{n}\sum_{i,j}(y_j(x_i) - \widehat{\mu_i})^2, \quad \widehat{\sigma^2_B} = \frac{1}{n}\sum_i n_i(\widehat{\mu_i} - \overline{y})^2$$

yielding the conventional

$$R^2 = \widehat{\sigma^2_B}/\widehat{\sigma^2_Y} \tag{65}$$

$$= 1 - \widehat{\sigma^2_W}/\widehat{\sigma^2_Y}. \tag{66}$$

In contrast the adjusted estimates (for two unknown parameters in the regression functions of our example)

$$\begin{aligned}
\widehat{\sigma^2_{Y,adj}} &= \frac{1}{n-1}\sum_{i,j}(y_j(x_i) - \overline{y})^2, \quad \widehat{\sigma^2_{W,adj}} = \frac{1}{n-2}\sum_{i,j}(y_j(x_i) - \widehat{\mu_i})^2, \\
\widehat{\sigma^2_{B,adj}} &= \frac{1}{q-1}\sum_i^q(\widehat{\mu}(x_i) - \overline{y})^2
\end{aligned}$$

yield two different estimates. The 'adjusted R^2',

$$R^2_{adj,1} = 1 - \widehat{\sigma^2_{W,adj}}/\widehat{\sigma^2_{Y,adj}} = 1 - (1 - R^2)(\frac{n-1}{n-2}) \leq R^2$$

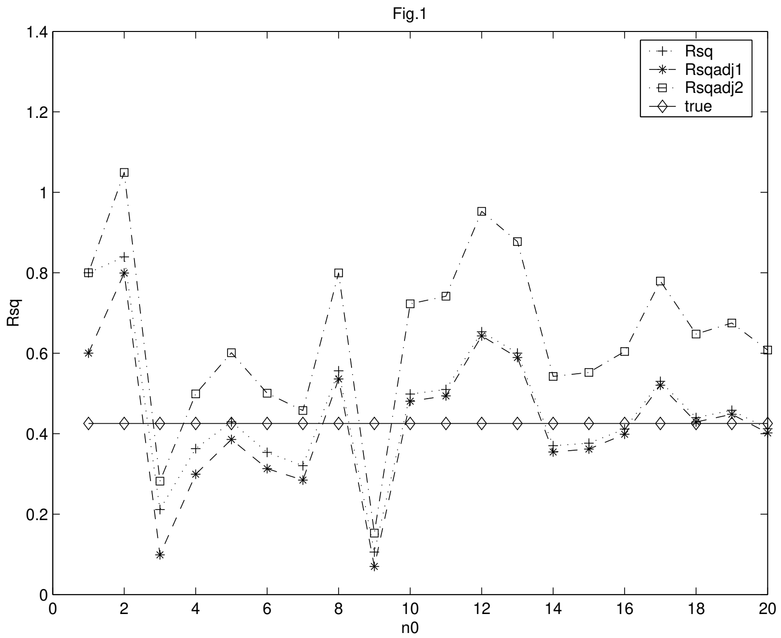

Figure 1: Estimates R^2, R^2_{adj}, R^2_{J,ζ_0} ('true') of R^2_{J,ζ_0} as functions of the number n_0 of replicates at each x_i

coincides with R^2 for large n. For $n = qn_0$

$$R^2_{adj,2} = \widehat{\sigma^2_{B,adj}} / \widehat{\sigma^2_{Y,adj}} = R^2\left(\frac{n-1}{n-n_0}\right) \geq R^2.$$

$\lim_{n_0 \to \infty} \frac{n_0 q - 1}{n_0 q - n_0} = \frac{q}{q-1}$ and $R^2_{adj,2} \geq 1$ may occur. Discussions in the literature about the appropriateness of *estimates* are presented as discussions about the appropriate *definition* of a coefficient of determination.

In fig.1 we visualize the performance of the estimators for simulated data generated according to the regression model $N(\mu_1(x), \sigma^2)$ specified in the previous section. We consider designs with equal replicates $n_i = n_0$, such that $p_i = 1/3 = n_0/n$ for $n_0 = 1, \ldots, 20$. For comparison the value R^2_{J,ζ_0} is given which was calculated using the empirical distribution of X specified by the experimental design to evaluate the integral w.r.t. X in $\sigma^2_B = var\mu_1(X)$ but otherwise using the true values of the parameters. Thus $E\widehat{J}_{KL}(p(y|\theta(X)), p(y|\theta_0)) = 0.7400$, $R^2_{J,\theta_0} = 0.4253$ are obtained.

6.1.2 Mixed model

We consider the extended regression function

$$\mu_2(x) = \alpha_1 + \beta_1(x - \overline{x}) + \beta_2(x^2 - \overline{x^2})$$

and assume that β_2 is a random coefficient: $\beta_2 \sim N(0, \tau^2)$. Marginally with $\mu_1(x)$ given in (62)

$$Y|\theta(x) \sim N(\mu_1(x), \sigma^2(x))$$

with heterogeneous variances $\sigma^2(x) = (x^2 - \overline{x^2})^2\tau^2 + \sigma^2$.

Quantities of interest Now for $\zeta_0^T = (\mu_Y/\sigma_Y^2, -\sigma_Y^{-2}/2)$

$$EJ_{KL}(p(y|\zeta(X)), p(y|\zeta_0)) = -\frac{1}{2}[1 - \sigma_Y^2 E(\frac{1}{\sigma^2(X)}) - E(\frac{(\mu(X) - \mu_Y)^2}{\sigma^2(X)})]. \quad (67)$$

Only if $\sigma^2(x) \equiv \sigma_W^2$, the right hand side of (67) reduces to σ_B^2/σ_W^2. Substituting $\sigma^2(x)$ by some average value $\overline{\sigma}_W^2$, say, one might approximate

$$EJ_{KL}(p(y|\zeta(X)), p(y|\zeta_0)) \approx -\frac{1}{2} + \frac{1}{2}(\frac{\sigma_Y^2 + \sigma_B^2}{\overline{\sigma}_W^2}) =: J_0. \quad (68)$$

The decomposition of variance then holds only approximately, $\sigma_Y^2 \approx \overline{\sigma}_W^2 + \sigma_B^2$, and

$$\begin{aligned} EJ_{KL}(p(y|\zeta(X)), p(y|\zeta_0)) &\approx \sigma_B^2/\overline{\sigma}_W^2 =: J_1 && (69) \\ &\approx (\sigma_Y^2/\overline{\sigma}_W^2) - 1 =: J_2. && (70) \end{aligned}$$

Correspondingly

$$\begin{aligned} R^2_{J,\zeta_0} &\approx \frac{J_0}{J_0 + 1} =: R^2_{J_0} && (71) \\ &\approx \frac{\sigma_B^2}{\sigma_B^2 + \overline{\sigma}_W^2} =: R^2_{J_1} && (72) \\ &\approx 1 - \frac{\overline{\sigma}_W^2}{\sigma_Y^2} =: R^2_{J_2}. && (73) \end{aligned}$$

Using the distribution of X induced by the uniform design to evaluate integrals w.r.t. X, the same numerical values $\alpha_1 = 1$, $\beta_1 = 1.2$ as in 6.1.1 and $\sigma^2 = 0.09$, $\tau^2 = 0.4$, we have $E\widehat{J}_{KL}(p(y|\zeta(X)), p(y|\zeta_0)) = 0.8929$ and $R^2_{J,\zeta_0} = 0.4717$. Hence in the mixed model there is a slightly stronger relation between X and Y than in the first model with non-random effects.

Estimation Willett and Singer ([42]: 1988) use weighted least squares estimates of $\mu(x)$ in order to quantify the 'variance explained by regression' in the original metric of Y in contrast to the decomposition of sum of squares of transformed variables proposed by Kvålseth ([21]: 1985). R^2_{J,ζ_0} refers to the metric of Y, and estimation should aim directly at $EJ_{KL}(p(y|\zeta(X)), p(y|\zeta_0))$. Otherwise invariance of (the estimate of) $EJ_{KL}(p(y|\zeta(X)), p(y|\zeta_0))$ w.r.t. the chosen transformation of Y should be checked.

For illustration we discuss an example with univariate continuous X and Y introduced by Draper and Smith ([11]: 1981, pp.112f) and resumed by Willett and Singer ([42]: 1988). In the example the variance function

$$\sigma^2(x) = 1.5329 - 0.7334x + 0.0883x^2$$

derived by Draper and Smith ([11]: 1981, p.115) is used, and the estimated regression function based on weighted least squares is

$$\widehat{\mu}_{WLS}(x) = -0.889 + 1.142x.$$

There are $n = 35$ data points.
We first estimate $EJ_{KL}(p(y|\zeta(X)), p(y|\zeta_0))$ as in (67), using $\widehat{\mu_Y} = \overline{y}$, $\widehat{\sigma_Y^2} = \frac{1}{35}\sum_\nu (y_\nu - \overline{y})^2$ and empirical means instead of expectations having evaluated $\widehat{\mu}(x)$ and $\sigma^2(x)$ for each x_ν, $\nu = 1, \ldots, 35$. Thus we obtain $E\widehat{J}_{KL}(p(y|\zeta(X)), p(y|\zeta_0)) = 35.83$, $R_{J,\zeta_0} = 0.9728$. Next we set

$$\widehat{\sigma^2_{B,WLS}} = \frac{1}{35}\sum_\nu^{35} (\widehat{\mu}_{WLS}(x_\nu) - \overline{y})^2.$$

In order to obtain approximations we try three values of $\overline{\sigma}^2_W$:

$$\widehat{\overline{\sigma}^2_{W,1}} = 1/(\frac{1}{35}\sum_\nu^{35} \frac{1}{\sigma^2(x_\nu)}) = 2.53,$$

$$\widehat{\overline{\sigma}^2_{W,2}} = \frac{1}{35}\sum_\nu^{35} \sigma^2(x_\nu) = 1.63,$$

$$\widehat{\overline{\sigma}^2_{W,3}} = \frac{1}{35}\sum_\nu^{35} (y_\nu - \widehat{\mu}_{WLS}(x_\nu))^2 = 2.02.$$

All of them do not satisfy $\widehat{\sigma_Y^2} = \widehat{\overline{\sigma}^2_W} + \widehat{\sigma^2_B}$. Insertion of these estimates in (68) results in $\widehat{J}_{0,i}$ ($R^2_{J_{0,i}}$), say, for $i = 1, 2, 3$. Substituting in (69) yields $\widehat{J}_{1,i}$ ($R^2_{J_{1,i}}$), and evaluating (70) gives $\widehat{J}_{2,i}$ ($R^2_{J_{2,i}}$) respectively. The values of the coefficients of determination are given in table 1.

Table 1

	$R^2_{J_{0,i}}$	$R^2_{J_{1,i}}$	$R^2_{J_{2,i}}$
$\widehat{\overline{\sigma}^2_{W,1}}$	0.9670	0.9655	0.9684
$\widehat{\overline{\sigma}^2_{W,2}}$	0.8709	0.8718	0.8699
$\widehat{\overline{\sigma}^2_{W,3}}$	0.8425	0.8460	0.8389

The results indicate that the choice of the variance estimator is more important than the approximation of $EJ_{KL}(p(y|\zeta(X)), p(y|\zeta_0))$. Relative to $R_{J,\zeta_0} = 0.9728$ the worst performance is obtained with $\widehat{\overline{\sigma}^2_{W,3}}$, an estimate that does not refer to the functional specification of $\sigma^2(x)$. The estimate $R^2_{J_{2,i}}$ evaluated with $\widehat{\overline{\sigma}^2_{W,3}}$ (i.e. 0.8389) was suggested by Willett and Singer ([42]: 1988). The best estimator is the one based on the mean of inverse variances, $1/\sigma^2(x)$, which according to (67) is in fact more appropriate than forming the inverse of the mean.

6.2 Example: Hald data

The Hald data set is a famous data set which has often been analyzed in order to illustrate techniques for variable selection. The data and an extensive classical analysis are given by Draper and Smith ([11]: 1981, pp.297, 591, 629). We resume this example in order to illustrate how the classical strategies of 'backward elimination' and 'forward insertion' amount to selecting a model with maximum strength of relation between X and Y.
The response variable Y is 'heat evolved in cement', and the covariates $\widetilde{X}_1, ...,$ $\widetilde{X}_4$, measure amounts of ingredients (in %) in clinkers used to produce the cement. We set $X = (\widetilde{X}_1, \ldots, \widetilde{X}_4)$. The covariates are nearly linearly dependent summing up to almost 100%. There are only $n = 13$ data points (y_ν, x_ν), where $x_\nu = (\widetilde{x}_{1\nu}, \ldots, \widetilde{x}_{4\nu})$. We assume $Y|x \sim N(\mu(x), \sigma^2)$, where $\mu(x) = \alpha + \beta^T x$, without referring to joint Normality of X and Y. We use centered covariates.

6.2.1 Quantities of interest

We again have conditional Gaussian densities with homogeneous variances. Hence the theoretical derivations in section 6.1.1 apply, and the main quantity of interest is again $R^2_{J,\zeta_0} = \sigma^2_B/\sigma^2_W$.

6.2.2 Estimation

Rao and Toutenburg ([31]: 1995, p.48) show that $R^2 = \widehat{R^2_{J,\zeta_0}}$ increases if variables are added within the linear regression model thus preserving monotonicity of the estimated quantity (in σ^2_W). For the Hald data the values of R^2 are (with $\{j, k, l\}$

indicating the sequence of variables $\widetilde{X}_i$ included in the model)

$$\begin{array}{llll}
\{1\} & 0.532 & \{3\} & 0.286 \\
\{1,2\} & 0.979 & \{3,4\} & 0.935 \\
\{1,2,4\} & 0.9823 & \{3,4,1\} & 0.981 \\
\{1,2,4,3\} & 0.9824 & \{3,4,1,2\} & 0.9824 \\
\{2\} & 0.666 & \{4\} & 0.674 \\
\{2,1\} & 0.979 & \{4,1\} & 0.972 \\
\{2,1,4\} & 0.9823 & \{4,1,2\} & 0.9823 \\
\{2,1,4,3\} & 0.9824 & \{4,1,2,3\} & 0.9824
\end{array}.$$

R^2_{adj}, now with $\widehat{\sigma^2_{W,adj}} = (n-k-1)^{-1}\sum_{\nu=1}^{n}(y_\nu - \widehat{\mu}(x_\nu))^2$, when k covariates $\widetilde{X}_i$ are included in the model, and with $\widehat{\sigma^2_{Y,adj}}$ as in section 6.1.1 is not monotone anymore. For the Hald data the sequence

$$\begin{array}{ll}
\{4\} & 0.645 \\
\{4,1\} & 0.967 \\
\{4,1,2\} & 0.9645 \\
\{4,1,2,3\} & 0.974
\end{array}$$

is obtained. Adjusted variances do penalize 'fit' (in the sum of squares) by 'model complexity' in the degrees of freedom. Yet adjusted variances exemplify, that estimates better than ml-estimates can be expected 'not to penalize sufficiently' for model complexity in model choice aiming at prediction because a coefficient of determination is an inappropriate criterion for predictive model choice.

6.2.3 Variable selection

Two principal methods are applied in variable selection: 'backward elimination' and 'forward insertion'.
Eliminating backwards one starts with the complete vector of covariates in the regression model and deletes in turn that component which least reduces the measure of dependence between X and Y until a significant drop occurs. In each step the difference of measures is assessed.
Under Normality of $Y|\theta(x)$ this reduces to the comparison of within variances, and in fact, the difference of $R^2_{I,\zeta_0}s$ or $R^2_{J,\zeta_0}s$ in this case is a standardized (by σ^2_Y) difference of within variances. The test statistic of the partial F-test used to establish significance of the difference under consideration is again a (differently) standardized difference of estimated within variances.

For the Hald data backward elimination based directly on R^2 suggests the path

$$\begin{array}{lllllllll}
\text{set} & \{1,2,3,4\} & \to & \{1,2,4\} & \to & \{1,2\} & \to & \{2\} \\
R^2 & 98.24 & & 98.23 & & 97.2 & & 66.6
\end{array}$$

with an intuitive cut-off at {1,2}. This is confirmed using partial F-tests as reported by Draper and Smith ([11]: 1981, p.306). Reversely, adding variables to the model such that the dependence between X and Y is maximally increased yields for the Hald data the steps

$$\begin{array}{lccccccc} \text{set} & \{4\} & \rightarrow & \{4,1\} & \rightarrow & \{4,1,2\} & \rightarrow & \{4,1,2,3\} \\ R^2 & 67.4 & & 97.2 & & 98.23 & & 98.24 \end{array}.$$

The cut-off point is again set at 97.2. Thus variable selection based on R^2 amounts to choosing the model with maximum dependence between X and Y taking into account the variability of the estimators of the measure of dependence.

The incremental association measured in terms of $I(\theta(X), Y)$ can be transformed to a difference of $R_I^2 s$ using (31). If joint Normality of X and Y holds this becomes a difference of multiple correlation coefficients. Standardization of this difference yields the partial correlation coefficient used to assess the contribution of the added variable to the overall association. Compare the discussion by Darbellay ([8]: 1998). As least squares estimates mimic a jointly Gaussian distribution of (X, Y) empirical multiple and partial correlation coefficients are defined analogously to the theoretical quantities even if the distributional assumption does not hold. In this spirit Draper and Smith ([11]: 1981, pp.308f) refer to empirical partial correlation coefficients for the Hald data.

6.3 Example: Doksum et al. (1994)

We analyze simulated data from a distribution introduced by Doksum et al. ([9]: 1994) in order to illustrate quantification of the local strength of association between X and Y. We also illustrate model comparison for nonparametric regression looking at the divergence as a function of the smoothing parameter.

We have univariate X and Y with

$$Y|\theta(x) \sim N(\mu(x), \sigma^2(x)),$$

where

$$\begin{aligned} \mu(x) &= (\frac{x}{10}) \exp(5 - \frac{x}{2}), \\ \sigma^2(x) &= \frac{1}{9}(1 + \frac{x}{2})^2, \end{aligned}$$

and $X \sim N(\mu_X, \sigma_X^2)$, $\mu_X = 1.2$, $\sigma_X^2 = 1/9$.

6.3.1 Quantities of interest

Global measure of dependence Considering the conditional Gaussian distributions as a two-parameter exponential family we have canonical parameters

$\zeta_1(x) = \mu(x)/\sigma^2(x)$, $\zeta_2(x) = -1/(2\sigma^2(x))$ corresponding to the canonical statistics $t_1(Y) = Y$, $t_2(Y) = Y^2$. Computation of $J(\theta(X), Y)$ according to (67) yields

$$J(\theta(X), Y) = 6.75, \quad R_J^2 = 0.87.$$

All expectations w.r.t. X were obtained by numerical integration over the range (0.0,2.4) covering 0.999 of the probability mass of X. The general definition of a coefficient of determination that we suggest thus naturally yields a coefficient for a nonparametric regression model with heterogeneous variances.

Local strength of dependence The limiting local discrepancy, quantifying the change of the distribution due to a change in x, is given by $J_0(x) = I(x)$ as x is univariate. Regarding the Normal distributions as a one-parameter exponential family where only changes in $\mu(x)$ - though weighted by $\sigma^2(x)$ - are reflected,

$$J_0(x) = \frac{\mu'(x)^2}{\sigma^2(x)} =: J_{01}(x).$$

Considering the Normal distributions as a two-parameter exponential family

$$\begin{aligned} J_0(x) &= I(x) \\ &= (\zeta_1'(x), \zeta_2'(x)) cov \begin{pmatrix} Y \\ Y^2 \end{pmatrix} \begin{pmatrix} \zeta_1'(x) \\ \zeta_2'(x) \end{pmatrix} \\ &= \sigma^2(x)(\zeta_1'(x))^2 + 4\mu(x)\sigma^2(x)\zeta_1'(x)\zeta_2'(x) \\ &\quad +4\mu^2(x)\sigma^2(x) + 2\sigma^4(x)(\zeta_2'(x))^2 \\ &= : J_{02}(x). \end{aligned}$$

The corresponding determination curves $R^2_{J_{0i}}(x) = \sigma_X^2 J_{0i}(x)/(1 + \sigma_X^2 J_{0i}(x))$ are displayed in fig.2. Note that according to (54) $R^2_{J_{01}}(x)$ is the correlation curve. It coincides with the curve displayed in fig.4 in ([9]: Doksum et al., 1994) apart from the range of x-values. At $x = 2$ the mean function μ attains a maximum, and hence the derivative is zero.
Fig.2 shows that the rates of change $J_{01}(x)$ and $J_{02}(x)$ behave similarly in the center of the distribution of X. Where they differ slightly ($x > 1.6$) it is indicated that the local strength of relation between X and Y evaluated in the two-parameter family is weaker than in the one-parameter family

According to (57) the certainty curve is $R_I^2(x) = 1 - c.\log(2\pi e\sigma^2(x))$, where $c.$ is a constant determined by $H(Y)$. Thus this curve pointwise results from an antitone transformation of $\sigma^2(x)$. As $\sigma^{2\prime}(x) > 0$, $\sigma^2(x)$ is an increasing function of x, and the reduction of uncertainty about Y is maximum for $x = 0$. Hence the certainty curve like the correlation curve decreases in (0,2.0).

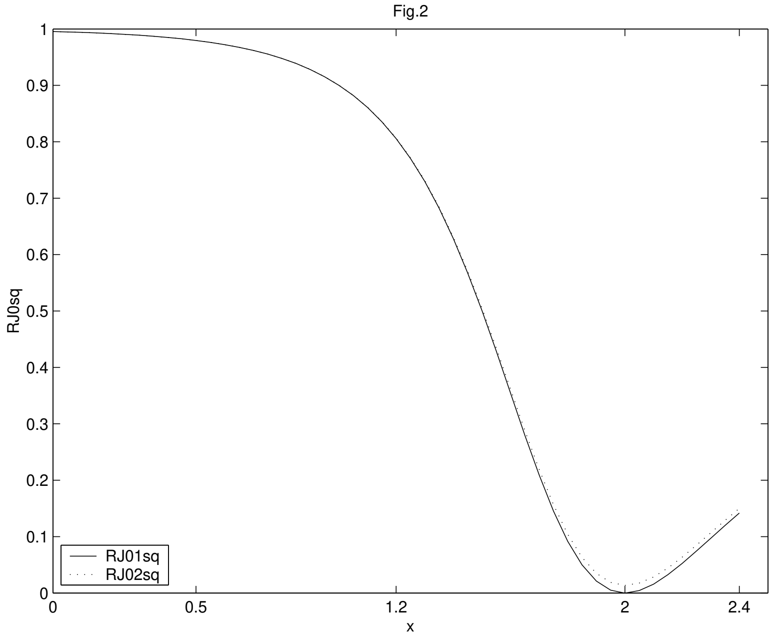

Figure 2: Determination curves corresponding to $J_0(x)$ evaluated for the Normal family as one-parameter (RJ01sq) or as two-parameter (RJ02sq) exponential family

Estimation Using 2000 simulated data points (x_ν, y_ν) we estimate σ_X^2, μ_Y, σ_Y^2 by the empirical mean and variances, and following Doksum et al.([9]: 1994) we then estimate $\mu(x)$ and $\sigma^2(x)$ by averages over neighbourhoods, where each neighbourhood of an x-value contains K points. The derivative of $\sigma^2(x)$ is calculated like $\mu'(x)$ as a ratio of differences using half the neighbourhoods. For further details see the paper by Doksum et al. ([9]: 1994). Inserting the estimated curves and evaluating expected values w.r.t. X as empirical means we obtain $\widehat{J}(\theta(X), Y)$. It is displayed as a function of K ($K = 30, 60, \ldots, 300$) in fig.3 below. The highest (estimated) strength of relation is attained for $K = 30$. Doksum et al.([9]: 1994) recommend $K = 60$ as optimal size of the neighbourhood, but we found that for higher K the estimates of the derivatives improve. It is to be expected that for over-smoothed $\widehat{\mu}$ the association between X and Y decreases, and this is confirmed in fig.3 showing $\widehat{J}(\theta(X), Y)$ as a decreasing function of K.

The estimated limiting local divergences $\widehat{J}_{01}(x)$ and $\widehat{J}_{02}(x)$ for the smoothing parameter $K = 300$, transformed into determination curves $R^2_{J_{01}}$, $R^2_{J_{02}}$ are shown in fig.4.

The estimates perform rather poorly, and better estimates are required to reliably evaluate the determination curves.

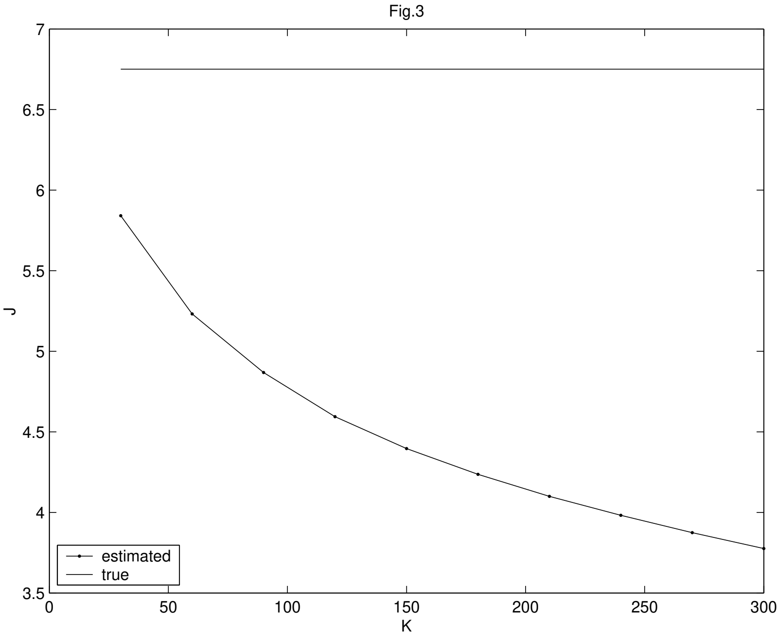

Figure 3: Estimated values of $J(\theta(X), Y)$ for smoothing parameters K with constant true value $J(\theta(X), Y) = 6.75$

6.4 Binomial response: example birth study

We illustrate our approach for a binomial response variable and emphasize comparison of models with different link functions. We also evaluate approximations of $EJ_{KL}(p(y|\zeta(X)), p(y|\zeta_0))$ according to results 4 and 5.

In the example the dependent variable Y indicates occurrence of an infection following birth by Caesarian section. The risk of infection is modelled depending on three binary covariates: X_1 indicates whether the Caesarian section was planned or not, X_2 indicates the presence of risk factors, and X_3 indicates prophylaxis with antibiotics. The data and analysis are given in the book by Fahrmeir and Tutz ([12]:1994, pp.29f). We use the notation introduced in section 2.3.2. Thus $\pi(x)$ denotes the risk of infection given $x = (x_1, x_2, x_3)$, and $Y|\pi(x) \sim B(1, \pi(x))$. The covariate vector X takes eight values, denoted by triples of zeros and ones, and their probabilities are given as $p(x)$ in general. Let the number of infections occurring under condition x be $k(x)$. There are $n = \sum_x n(x) = 251$ observations.

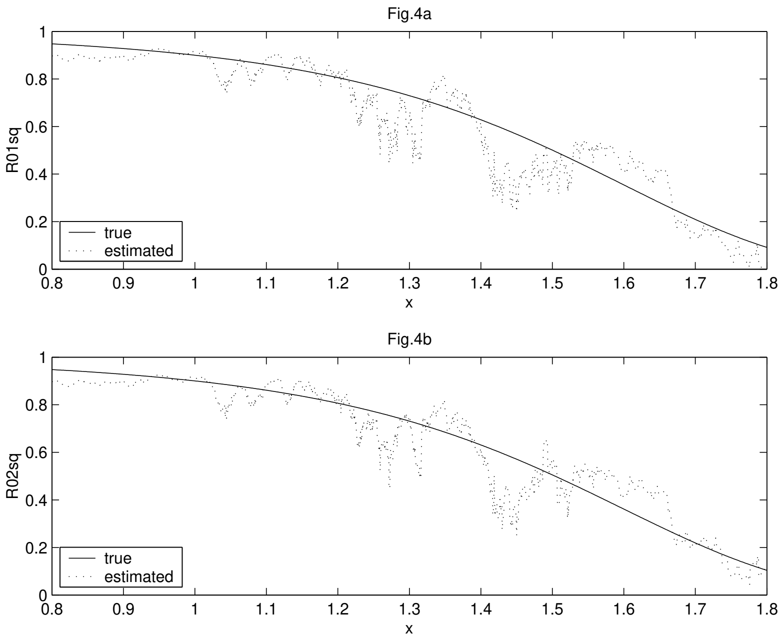

Figure 4: a) true and estimated correlation curve $\rho^2(x)$ = determination curve $R^2_{J_{01}}(x)$ with estimate based on $K = 300$ b) determination curve $R^2_{J_{02}}(x)$ with estimate based on $K = 300$

6.4.1 Quantities of interest

(i) For a Bernoulli reference distribution $B(1, \pi_0)$ we have

$$J_{KL}(p(y|\pi(x)), p(y|\pi_0)) = \sum_y ((p(y|\pi(x)) - p(y|\pi_0)) \log(\frac{p(y|\pi(x))}{p(y|\pi_0)}), \qquad (74)$$

where $p(y|\pi) = \pi^y(1-\pi)^{1-y}$, $\pi \in \{\pi(x), \pi_0\}$. Hence

$$\begin{aligned} & EJ_{KL}(p(y|\pi(X)), p(y|\pi_0)) \qquad (75) \\ = & \sum_x p(x)(\pi(x) - \pi_0)(\log \frac{\pi(x)}{1-\pi(x)} + \log \frac{1-\pi_0}{\pi_0}). \end{aligned}$$

(ii) Similarly

$$\begin{aligned} & EI_{KL}(p(y|\pi(X)), p(y|\pi_0)) \\ = & \sum_x p(x) \sum_y p(y|\pi(x)) \log \frac{p(y|\pi(x))}{p(y|\pi_0)}. \qquad (76) \end{aligned}$$

(iii) We also investigate approximations to $EI_{KL}(p(y|\pi(X)), p(y|\pi_0))$ and $EJ_{KL}(p(y|\pi(X)), p(y|\pi_0))$. First we consider the log-likelihood ratio as an approximation to $EI_{KL}(p(y|\pi(X)), p(y|\pi_0))$.

$$\begin{aligned}
& \frac{1}{n}\log\frac{\prod_x \pi(x)^{k(x)}(1-\pi(x))^{n(x)-k(x)}}{\prod_x \pi_0^{k(x)}(1-\pi_0)^{n(x)-k(x)}} \\
= & \sum_x \frac{n(x)}{n}\frac{1}{n(x)}\log\frac{\pi(x)^{k(x)}(1-\pi(x))^{n(x)-k(x)}}{\pi_0^{k(x)}(1-\pi_0)^{n(x)-k(x)}} \\
= & \sum_x \frac{n(x)}{n}[\frac{k(x)}{n(x)}\log\frac{\pi(x)}{\pi_0}+\frac{n(x)-k(x)}{n(x)}\log\frac{1-\pi(x)}{1-\pi_0}] \\
= & : I_{LR}.
\end{aligned}$$

Hence in the log-likelihood ratio $p(x)$ is replaced by the relative frequency $n(x)/n$ which would also be used in conditional sampling, and the model dependent probabilities $p(y|\pi(x))$ are replaced by relative frequencies $k(x)/n(x)$, $n(x)-k(x)/n(x)$ respectively. According to (33) the corresponding approximation for $EJ_{KL}(p(y|\pi(X)), p(y|\pi_0))$ is $EJ_{KL}(p(y|\pi(X)), p(y|\pi_0)) \approx 2I_{LR}$.
Result 4 suggests the 'Wald' approximation (with $\zeta_0 = logit\ \pi_0$)

$$EJ_{KL}(p(y|\pi(X)), p(y|\pi_0)) \approx E(tr[H_M(\zeta_0)(\zeta(X)-\zeta_0)(\zeta(X)-\zeta_0)^T])$$

yielding here

$$\begin{aligned}
EJ_{KL}(p(y|\pi(X)), p(y|\pi_0)) &\approx \sum_x p(x)[var(Y|\pi_0)(logit\ \pi(x) - logit\ \pi_0)^2 \\
&\approx \sum_x \frac{n(x)}{n}\pi_0(1-\pi_0)(logit\ \pi(x) - logit\ \pi_0)^2 \quad (77)
\end{aligned}$$

Result 5 gives the 'score' approximation

$$\begin{aligned}
EJ_{KL}(p(y|\pi(X)), p(y|\pi_0)) &\approx \sum_x p(x)(\pi(x)-\pi_0)^2/var(Y|\pi_0) \\
&\approx \sum_x \frac{n(x)}{n}\frac{(\pi(x)-\pi_0)^2}{\pi_0(1-\pi_0)}, \quad (78)
\end{aligned}$$

which is the Binomial version of the ratio of 'between' to 'within' variance.

(iv) Analogously to definition 1 we define coefficients of determination of Y by X_i conditional on $X_{-i} = x_{-i}$, where X_{-i} denotes the vector of covariates without X_i,

$$R^2_{J|x_{-i}} = \frac{J(\theta(X_i; x_{-i}), Y)}{1 + J(\theta(X_i, x_{-i}), Y)} \in [0, 1]. \quad (79)$$

Here we examine the partial association between Y (infection) and X_3 (antibiotics) fixing $(x_1, x_2) = x_{-3}$. For example, for $x_{-3} = (0, 1)$ corresponds to the condition that a Caesarian section was not planned but risk factors were present.

We compare three models
- logit link without interaction between X_1 and X_2,

$$\zeta(x) = \log \frac{\pi(x)}{1-\pi(x)} = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 =: \eta_1(x)$$

- logit link with interaction between X_1 and X_2,

$$\zeta(x) = \eta_1(x) + \beta_4 x_1 x_2 =: \eta_2(x)$$

- probit link without interaction,

$$\pi(x) = \Phi(\eta_1(x)).$$

6.4.2 Estimation

We evaluate the expressions given above for the estimate $\widehat{\pi}_0 = \sum_x k(x)/n = 0.283$.

For the three models we insert ml-estimates (given by Fahrmeir and Tutz ([12]:1994) in the formulae for the discrepancies and their approximations. Under the logit link the estimated coefficients of $\eta_1(x)$ are

$$\widehat{\alpha} = -1.89, \quad \widehat{\beta}_1 = 1.07, \quad \widehat{\beta}_2 = 2.03, \quad \widehat{\beta}_3 = -3.25,$$

and for $\eta_2(x)$

$$\widehat{\alpha} = -1.39, \quad \widehat{\beta}_1 = -11.64, \quad \widehat{\beta}_2 = 1.36, \quad \widehat{\beta}_3 = -3.83, \quad \widehat{\beta}_4 = 13.49.$$

Under the probit link $\eta_1(x)$ is estimated with coefficients

$$\widehat{\alpha} = -1.09, \quad \widehat{\beta}_1 = 0.69, \quad \widehat{\beta}_2 = 1.2, \quad \widehat{\beta}_3 = -1.9.$$

We also evaluate the discrepancies for the saturated model using the cell frequencies as estimates. We dealt with zero probabilities as arguments of the logarithm using the convention $0 \log 0 = 0$. The values are displayed in fig.5.

Evaluating $EI_{KL}(p(y|\pi(X)), p(y|\pi_0))$ we obtain $R^2_{I,\pi_0} = 0.2828$, indicating that the saturated model does outperform all other models, although the model with logit link and interactions attains nearly the same coefficient of determination. The relation between X and Y in terms of the divergence turns out to be strongest in the model with the logit link and interactions. For this particular data set it even outperforms the saturated model. This is due to the case where a Caesarian section was not planned and no risk factor was present but antibiotics were given, which occurred twice without any infection observed. The probability of infection is estimated by 0.0054 in the model with logit link and interactions and hence does add a term to the divergence whereas using the relative frequencies and the convention $0 \log 0 = 0$ no term reflects this case in the divergence.

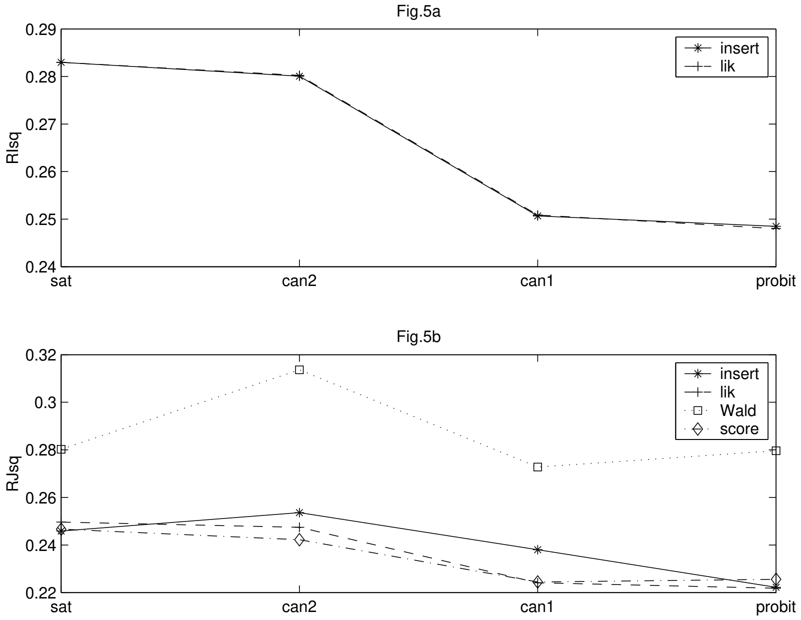

Figure 5: a) Estimates and estimates of approximations of R^2_{I,π_0} b) Estimates and estimates of approximations of R^2_{J,π_0}

We illustrate the coefficient of determination of Y by X_i conditional on x_{-i} given in (79). We evaluate the partial coefficients of determination between Y (infection) and X_3 (antibiotics) for the saturated model and for the model with logit link and interactions. The values are given in the following table.

Table 2

Estimated coefficients of determination $R^2_{J,\pi_0|x_{-3}}$

	saturated	logit, interaction
$x_{-3} = (x_1, x_2)$		
$(0, 1)$	0.1758	0.2470
$(1, 1)$	0.3445	0.3206

That is, the relation between infection and antibiotics turns out to be stronger if a Caesarian section was not planned, than in the case where a section was planned beforehand.

7 Discussion

7.1 Quantities of interest and estimates

We elaborated the idea to define coefficients of determination as measures of dependence, particularly based on the directed divergence stressing the explanatory power of X as reduction of uncertainty about Y or, based on the divergence, stressing the discriminatory power of X in telling apart conditional distributions of Y. In order to apply coefficients of determination in model comparison essentially two versions have been focussed on:
- for the comparison of models defined within an encompassing joint distribution of X and Y: $J(\theta(X), Y)$ and $I(\theta(X), Y)$
- for the comparison of models with (conditional) sampling distributions belonging to the same (often exponential) family of distributions:
$EJ_{KL}(p(y|\theta(X)), p(y|\theta_0))$ and $EI_{KL}(p(y|\theta(X)), p(y|\theta_0))$.
It is hard to decide whether to use the directed divergence or the divergence, and we do not recommend any of them exclusively. The most appealing feature of the divergence is the representation (result 1) as a functional of the log-odds ratio function Ψ^0 characterizing the association between X and Y (in the sense that their joint distribution is determined by the marginal distributions and Ψ). This feature yields a simple representation of the divergence in exponential families (result 3) where Ψ is bi-affine. The main advantage of the directed divergence is its decomposition in terms of entropies (28) and the monotonicity properties (discussed in section 4).

Model comparison using a reference density is essentially based on $-E_X E(\log p(y|\theta(X))|\theta_0)$. For the comparison of models without any common reference model the conditional entropy, that is the model specific expected value of the log density, $H(Y|\theta(X)) = -E_X E(\log p(y|\theta(X))|\theta(X))$ is used instead.

We did not consider covariates defined on an ordinal scale although measures of association have been applied in this case as well (e.g. [33]: Särndal, 1974). The adaptation of the coefficients of determination we propose to this set-up requires further study.

The clear distinction between theoretical and estimated quantities may stimulate further research on properties of estimates. The interpretation and use of coefficients of determination though is determined by the concept underlying their definition rather than properties of their estimates. We emphasize that in our view estimates of theoretically defined coefficients of determination empirically describe a feature of a regression *model*, namely dependence between a response Y and covariates X, and are not devised as measures of goodness of fit for a data set at hand. As ml-estimates of coefficients of determination often happen to be measures of goodness of fit this interpretation has been wide spread and implemented in sampling definitions of coefficients of determination (for example Cameron and Windmeijer ([5]:1997)).

7.2 Model choice

Model comparison w.r.t. the explanatory power of a model or the richness of a family of distributions as described in the paper is different from model choice if the target is *prediction* of future observations. In model choice it is often the *predictive* potential of a model that is of interest, and hence predictive densities are compared. The approximation of a predictive density by a likelihood needs to be corrected, and such a correction often yields a criterion for model choice of the form 'fit+complexity', for example AIC or its Bayesian extension DIC ([38]: Spiegelhalter et al., 2002). The target criterion for model choice $-E\log p(\widetilde{Y}|\widehat{\theta}(x))$, where $\widetilde{Y}$ denotes a future observation and $\widehat{\theta}$ a parameter estimate, provides the link between these approaches and coefficients of determination, particularly $H(Y|\theta(X))$.

Appendix

Proof of result 3:

$$\begin{aligned}
& J_{KL}(p(y|\zeta(x)), p(y|\overline{\zeta})) \\
&= \int (p(y|\zeta(x)) - p(y|\overline{\zeta}))\log(p(y|\zeta(x))/p(y|\overline{\zeta}))\,dy \\
&= E[(\zeta(x)-\overline{\zeta})^T t(Y) - M(\zeta(x)) + M(\overline{\zeta})|\zeta(x)] \\
&\quad -E[(\zeta(x)-\overline{\zeta})^T t(Y) - M(\zeta(x)) + M(\overline{\zeta})|\overline{\zeta}] \\
&= (\zeta(x)-\overline{\zeta})^T[E(t(Y)|\zeta(x)) - E(\,t(Y)|\overline{\zeta})].
\end{aligned}$$

Taking expectations w.r.t. x yields

$$\begin{aligned}
J(\theta(X),Y) &= E_X[(\zeta(X)-\overline{\zeta})^T\, E(t(Y)|\zeta(X))] \\
&= E_X[(\zeta(X)-\overline{\zeta})^T(E(t(Y)|\zeta(X)) - E\,t(Y))] \\
&= tr\{cov_X(\zeta(X),\, E(t(Y)|\zeta(X)))\}
\end{aligned}$$

which is (12). $\square$

Proof of result 5:
Set $\tau(x) = f(\zeta(x))$ or $\zeta(x) = f^{-1}(\tau(x))$. (14) can be elaborated as

$$\begin{aligned}
J(\theta(X),Y) &= E[(\zeta(X)-\overline{\zeta})^T(\tau(X)-\bar{\tau})] \\
&= E[\zeta(X)^T(\tau(X)-\bar{\tau})] \\
&= E[(\tau(X)-\bar{\tau})^T f^{-1}(\tau(X))] \\
&= E[u(\tau(X))] \qquad (say).
\end{aligned}$$

A second order Taylor expansion of $u(\tau(x))$ in $\bar{\tau}$ then yields the result. $\square$

References

[1] Bell, C.B. (1962). Mutual Information and Maximal Correlation as Measures of Dependence. Ann.Math.Statist. **33**, 587-595.

[2] Bernardo, J.M. and Rueda, R. (2002). Bayesian Hypothesis Testing: a Reference Approach. Int.Statist.Rev. **70**, 351-372.

[3] Blyth, S. (1994). Local Divergence and Association. Biometrika **81**, 579-584.

[4] Bjerve, S. and Doksum, K. (1993). Correlation Curves: Measures of Association as Functions of Covariate Values. Ann.Statist. **21**, 890-902.

[5] Cameron, A.C. and Windmeijer, A.G. (1997). An R-squared Measure of Goodness of Fit for Some Common Nonlinear Regression Models. J. of Econometrics **77**, 329-342.

[6] Cover, T.M. and Thomas, J.A. (1991). *Information Theory.* Wiley: New York.

[7] Cramer, E.M. and Nicewander, W.A. (1979). Some Symmetric, Invariant Measures of Multivariate Association, Psychometrika 44, 43-54.

[8] Darbellay, G.A. (1998). Predictability: an Information-Theoretic Perspective. In: A. Procházka et al. (eds.). *Signal Analysis and Prediction.* Birkhäuser: Boston 249-262.

[9] Doksum, K., Blyth, S., Bradlow, E., Meng, X.-L. and Zhao, H. (1994). Correlation Curves as Local Measures of Variance Explained by Regression. J.Amer.Statist.Ass. **89**, 571-582.

[10] Doksum, K. and Samarov, A. (1995). Nonparametric Estimation of Global Functionals and a Measure of the Explanatory Power of Covariates in Regression. Ann.Statist. **23**, 1443-1473.

[11] Draper, N.R. and Smith, H. (1981). *Applied Regression Analysis.* Wiley: New York.

[12] Fahrmeir, L. and Tutz, G. (1994). *Multivariate Statistical Modelling Based on Generalized Linear Models.* Springer: New York.

[13] Glahn, H. (1969). Some Relationships Derived from Canonical Correlation Theory. Econometrica **37**, 252-256.

[14] Gleason, T.C. (1976). On Redundancy in Canonical Analysis. Psycho.Bull. **83**, 1004-1006.

[15] Härdle, W. (1990). *Applied Nonparametric Regression.* Cambridge Univ. Press: Cambridge.

[16] Healy, M.J.R. (1984). The Use of R^2 as a Measure of Goodness of Fit. J.R.Statist.Soc. **A147**, 608-609.

[17] Hooper, J. (1959). Simultaneous Equations and Canonical Correlation Theory. Econometrica **27**, 245-256.

[18] Jessop, A. (1995). *Informed Assessments.* Ellis Horwood: New York.

[19] Joe, H. (1989). Relative Entropy Measures of Multivariate Dependence. J.Amer.Statist.Ass. **84**, 157-164.

[20] Kent, J.T. (1983). Information Gain and a General Measure of Correlation. Biometrika **70**, 163-173.

[21] Kvålseth, T.O. (1985). Cautionary Note about R^2. The Amer.Statist. **39**, 279-285.

[22] Kullback, S. (1968). *Information Theory and Statistics.* Dover Publ.: Mineola, New York (2nd ed.).

[23] Lavergne, P. and Vuong. Q.H. (1998). An Integral Estimator of Residual Variance and a Measure of Explanatory Power of Covariates in Nonparametric Regression. Nonpar. Statist. **9**, 363-380.

[24] van der Linde, A. (2004). On the Association between a Random Parameter and an Observable. Test **13**, 85-111.

[25] McKay, R.J. (1977). Variable Selection in Multivariate Regression: an Application of Simultaneous Test Procedures. J.R.Statist.Soc. **B39**, 371-380.

[26] Magee, L. (1990). R^2 Measures Based on Wald and Likelihood Ratio Joint Significance Tests. J.Amer.Statist.Ass. **44**, 250-253.

[27] Mardia, K.V., Kent, J.T. and Bibby, J.M. (1995). *Multivariate Analysis.* Academic Press: New York (10th ed.).

[28] Nagelkerke, N.J.D. (1991). A Note on a General Definition of the Coefficient of Determination. Biometrika **78**, 691-692.

[29] Osius, G. (2000). The Association between Two Random Elements: a Complete Characterization in terms of Odds Ratios. Metrika, to appear.

[30] Rao, C.R. (1973). *Linear Statistical Inference and Its Applications.* Wiley: New York.

[31] Rao, C.R. and Toutenburg, H. (1995). *Linear Models*. Spinger: New York.

[32] Renyi, A. (1959). On measures of dependence. Acta.Math.Acad.Sci. Hungar.**10**, 441-451.

[33] Särndal, C.E. (1974). A Comparative Study of Association Measures. Psychometrika **39**, 165-187.

[34] Silvey, S.D. (1964). On a Measure of Association. Ann.Math.Statist. **35**, 1157-1166.

[35] Simon, G. (1973). Additivity of Information in Exponential Family Probability Laws. J.Amer.Statist.Ass. **68**, 478-482.

[36] Soofi, E.S., Retzer, J.J. and Yasai-Ardekani, M. (2000). A Framework for Measuring the Importance of Variables with Applications to Management Research and Decision Models. Decision Sciences **31**, 595-625.

[37] Soofi, E.S.and Retzer, J.J. (2002). Information Indices: Unifications and Applications. J. of Econometrics **107**, 17-40.

[38] Spiegelhalter, D.J , Best, N.G., Carlin, B.P. and van der Linde, A. (2002). Bayesian Measures of Model Complexity and Fit (with discussion). J.R.Statist.Soc **B64**, 583-639.

[39] Stewart, D. and Love, W. (1968). A General Canonical Index. Psycho. Bull. **70**, 160-163.

[40] Theil, H. (1987). How Many Bits of Information Does an Independent Variable Yield in a Multiple Regression ? Statist.&Prob.Lett. **6**, 107-108.

[41] Theil, H. and Chung, C. (1988). Information-Theoretic Measures of Fit for Univariate and Multivariate Regression. The Amer. Statistician **42**, 249-252.

[42] Willett, J.B. and Singer, J.D. (1988). Another Cautionary Note about R^2: its Use in Weighted Least-Squares Regression Analysis. J.Amer.Statist.Ass. **42**, 236-238.