ELSEVIER

Journal of Econometrics 77 (1997) 329–342

JOURNAL OF Econometrics

An R-squared measure of goodness of fit for some common nonlinear regression models

A. Colin Cameron[a], Frank A.G. Windmeijer*[,b]

[a]*Department of Economics, University of California, Davis, CA 95616-8578, USA*
[b]*Department of Economics, University College London, London WC1E 6BT, UK*

(Received September 1994; final version received January 1996)

Abstract

For regression models other than the linear model, R-squared type goodness-of-fit summary statistics have been constructed for particular models using a variety of methods. We propose an R-squared measure of goodness of fit for the class of exponential family regression models, which includes logit, probit, Poisson, geometric, gamma, and exponential. This R-squared is defined as the proportionate reduction in uncertainty, measured by Kullback–Leibler divergence, due to the inclusion of regressors. Under further conditions concerning the conditional mean function it can also be interpreted as the fraction of uncertainty explained by the fitted model.

Key words: R-squared; Exponential family regression; Kullback–Leibler divergence; Entropy; Information theory; Deviance; Maximum likelihood
JEL classification: C52; C29

1. Introduction

For the standard linear regression model the familiar coefficient of determination, R-squared (R^2), is a widely used goodness-of-fit measure whose usefulness and limitations are more or less known to the applied researcher. Application of this measure to nonlinear models generally leads to a measure that can lie

*Corresponding author.

The authors are grateful to Richard Blundell, Shiferaw Gurmu, and two anonymous referees for their helpful comments.

PII S0304-4076(96)01818-0

outside the [0, 1] interval and decrease as regressors are added. Alternative R^2-type goodness-of-fit summary statistics have been constructed for particular nonlinear models using a variety of methods. For binary choice models, such as logit and probit, there is an abundance of measures; see Maddala (1983) and Windmeijer (1995). For censored latent models, such as the binary choice and tobit models, it is possible to avoid nonlinearity by obtaining an approximation of the usual R^2 for the linear latent variable model; see McKelvey and Zavoina (1976), Laitila (1993), and Veall and Zimmermann (1992, 1995). For other nonlinear regression models R^2 measures are very rarely used.

Desirable properties of an R-squared include interpretation in terms of the information content of the data, and sufficient generality to cover a reasonably broad class of models. We propose an R-squared measure based on the Kullback–Leibler divergence for regression models in the exponential family. This measure can be applied to a range of commonly-used nonlinear regression models: the normal for continuous dependent variable $y \in (-\infty, \infty)$; exponential, gamma, and inverse-Gaussian for continuous $y \in (0, \infty)$; logit, probit, and other Bernoulli regression models for discrete $y = 0, 1$; binomial (m trials) for discrete $y = 0, 1, \ldots, m$; Poisson and geometric for discrete $y = 0, 1, 2, \ldots$

The exponential family regression model is described in Section 2. In Section 3, the R^2 measure based on the Kullback–Leibler divergence is presented. This measures the proportionate reduction in uncertainty due to the inclusion of regressors. Interpretation of the measure in terms of the fraction of uncertainty explained by the fitted model is given in Section 4. Examples are presented in Section 5. Extensions and other goodness-of-fit statistics are discussed in Section 6. Section 7 contains an application to a gamma model for accident claims data. Section 8 concludes.

2. Exponential family regression models

Following Hastie (1987), assume that the dependent variable Y has distribution in the one-parameter exponential family with density

$$f_\theta(y) = \exp[\theta y - b(\theta)]h(y), \tag{1}$$

where θ is the natural or canonical parameter, $b(\theta)$ is the normalizing function, and $h(.)$ is a known function. Different $b(\theta)$ correspond to different distributions. The mean of Y, denoted μ, can be shown to equal the derivative $b'(\theta)$, and is monotone in θ. Therefore, the density can equivalently be indexed by μ, and expressed as

$$f_\mu(y) = \exp[c(\mu)y - d(\mu)]h(y).$$

General statistical theory for regression models based on the exponential family is given in Wedderburn and Nelder (1972), Gourieroux et al. (1984), and

White (1993). The standard reference for applications is McCullagh and Nelder (1989). Regressors are introduced by specifying μ to be a function of the linear predictor $\eta = x'\beta$, where x is a vector of regressors and β is an unknown parameter vector. Models obtained by various choices of $b(\theta)$ and functions of η are called generalized linear models. More specialized results are obtained by choice of the canonical link function, for which $\eta = \theta$, i.e., θ in (1) is set equal to $x'\beta$.

Binary choice models are an example of exponential family regression models. Then Y is Bernoulli distributed with parameter μ and density $f_\mu(y) = \mu^y(1-\mu)^{1-y}$, $y = \{0, 1\}$. This can be expressed as (1) with $\theta = \log(\mu/(1-\mu))$ and $b(\theta) = \log(1 + \exp(\theta))$. The logit regression model specifies $\mu = \exp(x'\beta)/(1 + \exp(x'\beta))$, while the probit regression model specifies $\mu = \Phi(x'\beta)$, where Φ is the standard normal cumulative distribution function. The logit model corresponds to use of the canonical link function.

The parameter vector β is estimated by the maximum likelihood (ML) estimator $\hat{\beta}$, based on the independent sample $\{(y_i, x_i),\ i = 1, \ldots, n\}$, with $f_{\mu_i}(y_i) = f_{\mu_i}(y_j)$ for $\mu_i = \mu_j$. The estimated mean for an observation with regressor x is denoted $\hat{\mu} = \mu(x'\hat{\beta})$. Throughout we assume that the model includes a constant term. The estimated mean from ML estimation of the constant only model is denoted $\hat{\mu}_0$.

3. *R*-squared based on the Kullback–Leibler divergence

A standard measure of the information content from observations in a density $f(y)$ is the expected information, or Shannon's entropy, $\mathrm{E}[\log(f(y))]$. This is the basis for the standard measure of discrepancy between two densities, the Kullback–Leibler divergence (Kullback, 1959). Recent surveys are given by Maasoumi (1993) and Ullah (1993).

Consider two densities, denoted $f_{\mu_1}(y)$ and $f_{\mu_2}(y)$ that are parameterized only by the mean. In this case the general formula for the Kullback–Leibler (KL) divergence is

$$K(\mu_1, \mu_2) \equiv 2\mathrm{E}_{\mu_1} \log[f_{\mu_1}(y)/f_{\mu_2}(y)], \tag{2}$$

where a factor two is added for convenience, and E_{μ_1} denotes expectation taken with respect to the density $f_{\mu_1}(y)$. $K(\mu_1, \mu_2)$ is the information of μ_1 with respect to μ_2 and is a measure of how close μ_1 and μ_2 are. The term divergence rather than distance is used because it does not in general satisfy the symmetry and triangular properties of a distance measure. However, $K(\mu_1, \mu_2) \geqslant 0$ with equality iff $f_{\mu_1} \equiv f_{\mu_2}$.

In addition to $f_{\mu_1}(y)$ and $f_{\mu_2}(y)$ we also consider the density $f_y(y)$, for which the mean is set equal to the realized y. Then the KL divergence $K(y, \mu)$ can be

defined in a manner analogous to (2) as

$$K(y,\mu) \equiv 2\mathrm{E}_y \log[f_y(y)/f_\mu(y)] = 2\int f_y(y)\log[f_y(y)/f_\mu(y)]\,\mathrm{d}y. \tag{3}$$

The random variable $K(y,\mu)$ is a measure of the deviation of y from the mean μ. For the exponential family, Hastie (1987) and Vos (1991) show that the expectation in (3) drops out and

$$K(y,\mu) = 2\log[f_y(y)/f_\mu(y)].$$

In the estimated model, with n individual estimated means $\hat{\mu}_i = \mu(x_i'\hat{\beta})$, the estimated KL divergence between the n-vectors $\boldsymbol{y}$ and $\hat{\boldsymbol{\mu}}$ is equal to twice the difference between the maximum log-likelihood achievable, i.e., the log-likelihood in a full model with as many parameters as observations, $l(\boldsymbol{y};\boldsymbol{y})$, and the log-likelihood achieved by the model under investigation, $l(\hat{\boldsymbol{\mu}};\boldsymbol{y})$:

$$K(\boldsymbol{y},\hat{\boldsymbol{\mu}}) = 2\sum_{i=1}^{n}[\log f_{y_i}(y_i) - \log f_{\hat{\mu}_i}(y_i)] = 2[l(\boldsymbol{y};\boldsymbol{y}) - l(\hat{\boldsymbol{\mu}};\boldsymbol{y})]. \tag{4}$$

Let $\hat{\boldsymbol{\mu}}_0$ denote the n-vector with entries $\hat{\mu}_0$, the fitted mean from ML estimation of the constant only model. We interpret $K(\boldsymbol{y},\hat{\boldsymbol{\mu}}_0)$ as the estimate of the information in the sample data on $\boldsymbol{y}$ potentially recoverable by inclusion of regressors. It is the difference between the information in the sample data on $\boldsymbol{y}$, and the estimated information using $\hat{\boldsymbol{\mu}}_0$, the best point estimate when data on regressors are not utilized, where information is measured by taking expectation with respect to the observed value $\boldsymbol{y}$. By choosing $\hat{\boldsymbol{\mu}}_0$ to be the MLE, $K(\boldsymbol{y},\hat{\boldsymbol{\mu}}_0)$ is minimized. The R-squared we propose is the proportionate reduction in this potentially recoverable information achieved by the fitted regression model:

$$R^2_{\mathrm{KL}} = 1 - K(\boldsymbol{y},\hat{\boldsymbol{\mu}})/K(\boldsymbol{y},\hat{\boldsymbol{\mu}}_0). \tag{5}$$

This measure can be used for fitted means obtained by any estimation method. In the following proposition we restrict attention to ML estimation (which minimizes $K(\boldsymbol{y},\hat{\boldsymbol{\mu}})$):

Proposition 1. For ML estimates of exponential family regression models based on the density (1), R^2_{KL} defined in (5) has the following properties.

1. *R^2_{KL} is nondecreasing as regressors are added.*
2. *$0 \leqslant R^2_{\mathrm{KL}} \leqslant 1$.*
3. *R^2_{KL} is a scalar multiple of the likelihood ratio test for the joint significance of the explanatory variables.*

4. R^2_{KL} *equals the likelihood ratio index* $1 - l(\hat{\mu}; y)/l(\hat{\mu}_0; y)$ *if and only if* $l(y; y) = 0$.

5. R^2_{KL} *measures the proportionate reduction in recoverable information due to the inclusion of regressors, where information is measured by the estimated Kullback–Leibler divergence (4).*

Proof

1. The MLE minimizes $K(y, \hat{\mu})$ which will therefore not increase as regressors are added.

2. The lower bound of 0 occurs if inclusion of regressors leads to no change in the fitted mean, i.e., $\hat{\mu} = \hat{\mu}_0$, and the upper bound occurs when the model fit is perfect.

3. Follows directly from re-expressing R^2_{KL} as $2[l(\hat{\mu}; y) - l(\hat{\mu}_0; y)]/K(y, \hat{\mu}_0)$.

4. Follows directly from re-expressing R^2_{KL} as $[1 - l(\hat{\mu}; y)/l(\hat{\mu}_0; y)]\ [l(\hat{\mu}_0; y)/(l(\hat{\mu}_0; y) - l(y; y)]$.

5. See the discussion leading up to (5).

Properties 1 and 2 are standard properties often desired for R-squared measures. Property 3 generalizes a similar result for the linear regression model under normality. The relationship between likelihood ratio tests and the Kullback–Leibler divergence is fully developed in Vuong (1989). Property 4 is of interest as the likelihood ratio index, which measures the proportionate reduction in the log-likelihood due to inclusion of regressors, is sometimes proposed as a general pseudo R-squared measure. Equality occurs for the Bernoulli model, but in general the likelihood ratio index differs and, for other discrete dependent variable models, is more pessimistic regarding the contribution of regressors, as $l(y; y) \leqslant 0$. In the continuous case, large values (positive or negative) of the likelihood ratio index can arise if $l(\hat{\mu}_0; y)$ is close to zero (positive or negative). By contrast, R^2_{KL} will always be bounded by zero and one. The final property establishes an information-theoretic basis for R^2_{KL}.

An interesting aspect is that the expression for $K(y, \hat{\mu})$ in (4) equals the definition of the deviance, given in, for example, McCullagh and Nelder (1989, p. 33). Therefore R^2_{KL} can be interpreted as being based on deviance residuals, defined as the signed square root of individual contributions to the deviance. Deviance residuals have been found very useful for diagnostic checking in generalized linear models, see, e.g., Pregibon (1981), Landwehr et al. (1984), and Williams (1987); and R^2_{KL} is related to the analysis of deviance the same way as R^2 in the standard linear model is related to the analysis of variance.

4. Pythagorean decomposition for R^2_{KL}

In the linear regression model, the usual R-squared can be interpreted not only as the proportionate reduction in the total sum of squares due to inclusion of regressors, but also as the fraction of the total sum of squares explained by the regression model. This result rests on the decomposition of the total sum of squares into explained sum of squares and residual sum of squares. Such a decomposition of the sum of squares does not generally hold for exponential family regression models, which is one reason for not applying the linear regression model R-squared to other models.

For a widely used subclass of exponential family regression models that use the canonical link, R^2_{KL} has the desirable property of interpretation in terms of explained KL divergence between the fitted model and the constant-only model.

Proposition 2. For the exponential family models that use the canonical link, i.e., $0 = x'\beta$ in (1), R^2_{KL} defined in (5) can be equivalently expressed as

$$R^2_{KL} = K(\hat{\mu}, \hat{\mu}_0)/K(y, \hat{\mu}_0).$$

where $K(\hat{\mu}, \hat{\mu}_0)$ is the estimated KL divergence defined in (2) between models with fitted means $\hat{\mu}$ and $\hat{\mu}_0$, and so R^2_{KL} measures the fraction of uncertainty explained by the fitted model.

Proof. Let the vector $\hat{\mu}_1 = \mu(x_1'\hat{\beta})$, and $\hat{\mu}_2 = \mu(x_2'\hat{\beta})$, with x_2 nested in x_1. For models that use the canonical link, the KL divergence exhibits the Pythagorean property (see Hastie, 1987, pp. 19–20; Simon, 1973):

$$K(y, \hat{\mu}_2) = K(\hat{\mu}_1, \hat{\mu}_2) + K(y, \hat{\mu}_1).$$

Proposition 2 follows, using the particular decomposition $K(y, \hat{\mu}_0) = K(\hat{\mu}, \hat{\mu}_0) + K(y, \hat{\mu})$. □

For models that do not use the canonical link, R^2_{KL} still satisfies all the properties in Proposition 1, in particular property 5. A decomposition is trivially obtained by use of the so-called likelihood displacement defined as (Cook, 1986; Vos, 1991)

$$LD(\hat{\mu}, \hat{\mu}_0) \equiv 2\{l(\hat{\mu}; y) - l(\hat{\mu}_0; y)\} = K(y, \hat{\mu}_0) - K(y, \hat{\mu}).$$

Then

$$R^2_{KL} = LD(\hat{\mu}, \hat{\mu}_0)/K(y, \hat{\mu}_0),$$

and can be interpreted as measuring the fraction of *empirical* uncertainty explained by the model.[1]

[1] See Hauser (1978), who analyzed the likelihood ratio index for Bernoulli and multinomial models.

5. Examples

The formulae for R^2_{KL} for a range of exponential family regression models are given in Table 1. The models are defined in, for example, McCullagh and Nelder (1989, p. 30). The column R^2_{KL} is the measure defined in (5). The final column gives the conditional mean, as a function of $\eta = x'\beta$ corresponding to the canonical link, in which case Proposition 2 also holds and R^2_{KL} can be simplified in certain cases.

For the normal distribution, with σ^2 known (or using the same estimator for the two models), R^2_{KL} given in Table 1 equals the usual coefficient of determination in the linear model. Proposition 2 applies to the linear regression model, but not to nonlinear models under normality since these do not use the canonical link.

For the linear model with nonspherical disturbances ($\text{var}(y) = \sigma^2 V$, with V known), the KL divergence can be shown to be given by

$$K(y, \hat{\mu}) = (y - \hat{\mu})' V^{-1} (y - \hat{\mu}),$$

and R^2_{KL} is

$$R^2_{KL,GLS} = \frac{(\hat{\mu} - \iota\hat{\mu}_0)' V^{-1} (\hat{\mu} - \iota\hat{\mu}_0)}{(y - \iota\hat{\mu}_0)' V^{-1} (y - \iota\hat{\mu}_0)} = 1 - \frac{(y - \hat{\mu})' V^{-1} (y - \hat{\mu})}{(y - \iota\hat{\mu}_0)' V^{-1} (y - \iota\hat{\mu}_0)}, \tag{6}$$

where ι is the vector of ones, and $\hat{\mu}_0 = (\iota' V^{-1} \iota)^{-1} \iota' V^{-1} y$. So in this case, R^2_{KL} is equal to the definition as given by Buse (1973).

For Bernoulli regression models, where y takes only the values 0 or 1, many R^2 measures have been proposed. See, for example, Maddala (1983, pp. 37–41) and Windmeijer (1995), or the output from the econometrics package SHAZAM. For these models, $l(y; y) = 0$, so that by property 4 in Proposition 1, R^2_{KL} given in Table 1 is equal to the likelihood ratio index proposed by McFadden (1974), Efron (1978) for one-way ANOVA, Pregibon (1984) who explicitly derives his measure based on deviances, and Christensen (1990). Proposition 2 applies to the logit model, but not to the probit model.

An R^2 measure is rarely reported for the Poisson model. R^2_{KL} given in Table 1 equals one of the R^2 measures proposed for this model by Cameron and Windmeijer (1996). The standard Poisson regression model specifies $\mu = \exp(x'\beta)$ which is the canonical link so that Proposition 2 applies.

Table 1 also lists R^2_{KL} for the binomial, geometric, exponential, gamma, and inverse-Gaussian regression models. For these models we have been unable to find specific R^2 measures in the literature.

The analysis can be extended to a p-dimensional dependent variable with density in the p-parameter exponential family. Necessary results for such generalization are given in Simon (1973). Of particular interest is the multinomial

Table 1
R^2_{KL} for exponential family regression models

Distribution	KL divergence	R^2_{KL}	Canonical link[a]
Normal	$\sum(y-\mu)^2/\sigma^2$	$1-\frac{\sum(y-\hat{\mu})^2}{\sum(y-\bar{y})^2}$	$\mu=\eta$
Bernoulli	$-2\sum\{y\log\mu+(1-y)\log(1-\mu)\}$	$1-\frac{\sum\hat{\mu}\log(\hat{\mu})+(1-\hat{\mu})\log(1-\hat{\mu})}{n\{\bar{y}\log(\bar{y})+(1-\bar{y})\log(1-\bar{y})\}}$	$\mu=\frac{\exp(\eta)}{1+\exp(\eta)}$
Binomial (m)	$2\sum\left\{y\log\left(\frac{y}{\mu}\right)+(m-y)\log\left(\frac{m-y}{m-\mu}\right)\right\}$	$1-\frac{\sum\hat{\mu}\log(\hat{\mu})+(m-\hat{\mu})\log(m-\hat{\mu})}{n\{\bar{y}\log(\bar{y})+(m-\bar{y})\log(m-\bar{y})\}}$	$\mu=\frac{\exp(\eta)}{1+\exp(\eta)}$
Poisson[b]	$2\sum\{y\log(y/\mu)-(y-\mu)\}$	$1-\frac{\sum y\log(y/\hat{\mu})-(y-\hat{\mu})}{\sum y\log(y/\bar{y})}$	$\mu=\exp(\eta)$
Geometric[b]	$2\sum\left\{y\log\left(\frac{y}{\mu}\right)-(y+1)\log\left(\frac{y+1}{\mu+1}\right)\right\}$	$1-\frac{\sum y\log\left(\frac{y}{\hat{\mu}}\right)-(y+1)\log\left(\frac{y+1}{\hat{\mu}+1}\right)}{\sum y\log\left(\frac{y}{\bar{y}}\right)-(y+1)\log\left(\frac{y+1}{\bar{y}+1}\right)}$	$\mu=\frac{\exp(\eta)}{1-\exp(\eta)}$
Exponential	$-2\sum\{\log(y/\mu)+(y-\mu)/\mu\}$	$1-\frac{\sum\log(y/\hat{\mu})+(y-\hat{\mu})/\hat{\mu}}{\sum\log(y/\bar{y})}$	$\mu=\eta^{-1}$
Gamma[c]	$-2\nu\sum\{\log(y/\mu)+(y-\mu)/\mu\}$	$1-\frac{\sum\log(y/\hat{\mu})+(y-\hat{\mu})/\hat{\mu}}{\sum\log(y/\bar{y})}$	$\mu=\eta^{-1}$
Inverse Gaussian	$\sum(y-\mu)^2/(\mu^2 y)$	$1-\frac{\sum(y-\hat{\mu})^2/(\hat{\mu}^2 y)}{\sum(y-\bar{y})^2/(\bar{y}^2 y)}$	$\mu=\eta^{-2}$

[a] $\eta=x'\beta$; [b] $y\log(y)=0$ for $y=0$; [c] ν is the scale parameter.

distribution, used for example in multi-choice regression models such as multinomial and nested logit. In this case $l(\mathbf{y};\mathbf{y}) = 0$, so that R^2_{KL} equals the likelihood ratio index analyzed by Hauser (1978).

6. Discussion

Different interpretations of the coefficient of determination in the linear regression model, R^2_{OLS}, lead to different R^2 measures for nonlinear models, each with some, but not all, of the properties possessed by R^2_{OLS}. A number of the possible general approaches are given in, for example, Magee (1990) and Veall and Zimmermann (1992, 1995). The most easily interpretable measures are based on residual sums of squares, $1 - \sum_i (y_i - \hat{\mu}_i)^2 / \sum_i (y_i - \bar{y})^2$, or explained sums of squares $\sum_i (\hat{\mu}_i - \bar{y})^2 / \sum_i (y_i - \bar{y})^2$. But in nonlinear models these two measures may fall outside the unit interval, decrease as regressors are added, and differ from each other.[2]

A number of proposed measures, including R^2_{KL}, are related to LRT, the likelihood ratio test statistic for the joint significance of the slope parameters. In particular, a general measure proposed by Kent (1983), Maddala (1983), and Magee (1990) is

$$R^2_{LRT} = 1 - \exp(-LRT/n). \tag{7}$$

Kent argued that LRT/n is an estimate of the expected Kullback–Leibler information gain, the expectation being with respect to regressors x. Kent chose this particular transformation of LRT/n as it is guaranteed to lie within the unit interval, and it equals the usual multiple correlation coefficient in the regression model under normality. Maddala and Magee proposed R^2_{LRT} on grounds that in the linear model it equals R^2_{OLS}. All treat the variance σ^2 as an unknown parameter that is estimated by $n^{-1}\sum_i (y_i - \hat{\mu}_i)^2$ in the fitted model and by $n^{-1}\sum_i (y_i - \bar{y})^2$ in the constant-only model.

Magee (1990) also proposed a measure based on a Wald test rather than likelihood ratio test:

$$R^2_W = W/(n + W),$$

where W is the Wald test statistic for joint significance of the slope parameters. As Magee (1990) notes, R^2_W does not necessarily increase when regressors are added, and another drawback of the measure is the lack of invariance of W to the parameterization of the model. In the linear model R^2_W equals R^2_{OLS}, where the variance σ^2 is again treated as an unknown parameter.

[2] In the special case that the nonlinear model is based on a linear latent variable model, Veall and Zimmermann (1992, 1995) advocate estimating the latter measure for the underlying latent variable. This approach cannot be applied to most of the models considered here.

This different treatment of the scale parameter needs to be emphasized. The discussion of R^2_{KL} was restricted to exponential family models where the scale parameter is known. This includes Bernoulli, Poisson, geometric, and exponential, which have no scale parameter, and binomial for which the scale parameter (the number of trials m) is usually known. For models with unknown scale parameter, R^2_{KL} is easily extended if the KL divergence is multiplicative in the scale parameter. Then the scale parameter cancels out from numerator and denominator in R^2_{KL}, leaving the same formulae for R^2_{KL} as Table 1 with no need to estimate the scale parameter. This is the case for the normal (σ^2 unknown) and gamma (ν unknown) distributions. By contrast, the motivation of R^2_{LRT} assumes estimation of any scale parameter, since if σ^2 is known in the linear regression model under normality, (7) yields $R^2_{LRT} = 1 - \exp(\sum_i\{(y_i - \hat{\mu}_i)^2 - (y_i - \bar{y})^2\}/\{n\sigma^2\})$ rather than R^2_{OLS}.

For exponential family models with known (or no) scale parameter,

$$R^2_{LRT} = 1 - \exp(-R^2_{KL}\, K(\boldsymbol{y}, \hat{\boldsymbol{\mu}}_0)/n),$$

and therefore R^2_{LRT} takes maximum value of $1 - \exp(-K(\boldsymbol{y}, \hat{\boldsymbol{\mu}}_0)/n)$ when $R^2_{KL} = 1$. So a measure with upper bound of one is

$$R^2_{LRTu} = \frac{1 - \exp(-LRT/n)}{1 - \exp(-K(\boldsymbol{y}, \hat{\boldsymbol{\mu}}_0)/n)}.$$

This equals the R^2 measure for the multinomial logit model given by Cragg and Uhler (1970) and discussed in Maddala (1983, pp. 39–40). Note that $R^2_{LRTu} > R^2_{KL}$ for $0 < R^2_{KL} < 1$. There are clearly many ways to generate an R^2 measure based on the likelihood ratio test that lies between 0 and 1 and increases as regressors are added. R^2_{KL} has the additional advantage of interpretation in terms of proportionate reduction in recoverable information.

An interesting question is generalization of R^2_{KL} to any model specification estimated by maximum likelihood. By (4)

$$R^2_{KL} = 1 - \frac{l_{max} - l_{fit}}{l_{max} - l_0} = \frac{l_{fit} - l_0}{l_{max} - l_0}, \tag{8}$$

where l_{fit}, l_0, and l_{max} denote, respectively, the log-likelihood in the fitted model, the log-likelihood in the intercept-only model, and the maximum log-likelihood achievable. Thus R^2_{KL} equals the fraction of the maximum potential likelihood gain (starting with a constant-only model) achieved by the fitted model. This definition works well in cases such as exponential family models with known scale parameter where l_{max} is well-defined.[3] But in other cases, such as the

[3] See also Merkle and Zimmermann (1992), who proposed use of R^2_{KL} as defined in (8) for the Poisson model.

normal with σ^2 unknown, $l_{\max}$ is not defined.[4] Even where $l_{\max}$ is defined, it should be noted that it does not necessarily equal the log-likelihood evaluated at $\hat{\mu} = y$.[5]

7. Application

To illustrate the behaviour of R^2_{KL} and other R^2 measures we perform an analysis of the cost of claims for damage to an owner's car for privately owned vehicles with comprehensive cover. The data used is the same as in Baxter et al. (1980) (see also McCullagh and Nelder, 1989, p. 298). The data set consists of cell average cost of claims for each of 123 cells, where the cells are determined by eight categories of policy-holders age, four categories of vehicle age, and four categories of car group (cells with no claims are excluded). A gamma distribution is assumed, with log-likelihood

$$\sum_i \{\nu_i(-y_i/\mu_i - \log \mu_i + \nu_i \log y_i + \log \nu_i) - \log \Gamma(\nu_i)\};$$

conditional mean $\mu_i = (x_i'\beta)^{-1}$, corresponding to the canonical link function for gamma; and scale parameter $\nu_i = \nu \cdot w_i$, where ν is a scalar and the weight w_i equals the number of claims within each cell i. In constructing R^2_{KL} the parameter ν is assumed known and factors out. By contrast, in computing R^2_{LRT} and R^2_W ν is treated as unknown and needs to be separately estimated, and LRT is computed as $2\{l(\hat{\mu}, \hat{\nu}; y) - l(\hat{\mu}_0, \hat{\nu}_0; y)\}$. For comparative purposes we additionally calculate R^2_W, R^2_{LRT}, and R^2_{LRTu} for ν known and equal to 1, which corresponds to the exponential. The estimation results for the mean μ are independent of the value of ν.

The results as presented in Table 2 are given for three different models: PA has seven dummies for categories of policy-holders age; PA + CG additionally includes dummies for three categories of car group; PA + CG + VA additionally includes dummies for three categories of vehicle age. The values of the three measures are very similar for the gamma model with ν estimated, but they differ quite substantially when ν is set equal to 1 (exponential) in which case R^2_W (for the first two models) and R^2_{LRT} (for all models) are much higher than R^2_{KL}. For the measure based on the Wald statistic these higher values occur due to the

[4]Assume that each y_i is drawn from $N(\mu_i, \sigma^2)$, where $\mu_i = y_i$ and $\sigma^2 \to 0$. Then the density of each y_i, and hence the log-likelihood for the sample, becomes infinite. For the negative binomial model, where a similar problem arises, Cameron and Windmeijer (1996) propose setting the scale parameter to its estimate in the fitted model.

[5]For example, consider the log-normal, $\log y_i \sim N(\theta_i, 1)$ in which case $\mu_i = \exp(\theta_i + 0.5)$. The log-density of y_i is maximized w.r.t. θ_i at $\theta_i = \log y_i$, and hence is maximized w.r.t. μ_i at $\mu_i = y_i \exp(0.5) \neq y_i$.

Table 2
Results R^2's for car insurance data

	ν estimated			$\nu = 1$		
VAR	R^2_{KL}	R^2_W	R^2_{LRT}	R^2_W	R^2_{LRT}	R^2_{LRTu}
PA	0.127	0.138	0.134	0.410	0.487	0.490
PA CG	0.478	0.514	0.496	0.737	0.920	0.925
PA CG VA	0.808	0.800	0.820	0.799	0.986	0.991

PA: policy holder's age; CG: car group; VA: vehicle age.
$\hat{\nu}_0 = 0.203$; $\hat{\nu} = 0.231$ for PA; $\hat{\nu} = 0.378$ for PA CG; $\hat{\nu} = 1.004$ for PA CG VA.

fact that the variances in the exponential model are smaller than the estimated variances in the gamma model ($\hat{\nu} < 1$) for the first two models. For the measure based on the likelihood ratio statistic the reason is the smaller value of $l(\hat{\mu}_0, 1, y)$ as compared to $l(\hat{\mu}_0, \hat{\nu}_0; y)$.[6] The differences between R^2_{LRT} and R^2_{LRTu} are small due to the fact that the term $1 - \exp(-K(y, \hat{\mu}_0)/n)$ is close to 1.

The R^2 measures clearly convey the message that the full model provides a very good fit for this data. While the R^2's may appear high to those familiar with cross-section data, the full model does actually fit the data well, as even standard weighted nonlinear least squares, i.e., minimize $\sum_i w_i(y_i - 1)/(x_i'\beta))^2$ gives an $R^2_{KL,GLS}$, as defined in (6), equal to 0.79 in the model with all categories included.

8. Conclusions

For exponential family regression models, the Kullback–Leibler divergence can be used to construct an R^2 measure of goodness of fit, denoted R^2_{KL}, that measures the proportionate reduction in uncertainty due to the inclusion of regressors, lies between 0 and 1 and is nondecreasing as regressors are added. R^2_{KL} corresponds to the usual coefficient of determination in the linear regression under normality. In Bernoulli models, such as probit and logit, R^2_{KL} coincides with the likelihood ratio index, supporting use of this index rather than the many other competing R^2 measures. R^2_{KL} can also be used for other regression models in the exponential family, such as Poisson, geometric, binomial, exponential, and gamma, for which R^2 measures do not generally appear to be available. For models with canonical link function, R^2_{KL} can additionally be interpreted as the fraction of uncertainty explained by the fitted model.

[6] Equivalently, R^2_{LRT} takes on very high values when LRT is computed as $2\{l(\hat{\mu}, \hat{\nu}; y) - l(\hat{\mu}_0, \hat{\nu}, y)\}$.

References

Baxter, L.A., S.M. Coutts, and G.A.F. Ross, 1980, Applications of linear models in motor insurance, Proceedings of the 21st International Congress of Actuaries, Zurich, 11–29.

Buse, A., 1973, Goodness of fit in generalized least squares estimation, The American Statistician 27, 106–108.

Cameron, A.C. and F.A.G. Windmeijer, 1996, R-squared measures for count data regression models with applications to health care utilization, Journal of Business and Economic Statistics 14, 209–220.

Christensen, R., 1990, Log-linear models (Springer-Verlag, New York, NY).

Cragg, J. and R. Uhler, 1970, The demand for automobiles, Canadian Journal of Economics 3, 386–406.

Cook, R.D., 1986, Assessment of local influence, Journal of the Royal Statistical Society B 48, 133–169.

Efron, B., 1978, Regression and ANOVA with zero-one data: Measures of residual variation, Journal of the American Statistical Association 73, 113–121.

Gourieroux, C., A. Montfort, and A. Trognon, 1984, Pseudo maximum likelihood methods: Theory, Econometrica 52, 681–700.

Hastie, T., 1987, A closer look at the deviance, The American Statistician 41, 16–20.

Hauser, J.A., 1978, Testing the accuracy, usefulness, and significance of probabilistic choice models: An information-theoretic approach, Operations Research 26, 406–421.

Kent, J.T., 1983, Information gain and a general measure of correlation, Biometrika 70, 163–173.

Kullback, S., 1959, Information theory and statistics (Wiley, New York, NY).

Laitila, T., 1993, A pseudo-R^2 measure for limited and qualitative dependent variable models, Journal of Econometrics 56, 341–356.

Landwehr, J.M., D. Pregibon, and A.C. Shoemaker, 1984, Graphical methods for assessing logistic regression models, Journal of the American Statistical Association 79, 61–83.

Maasoumi, E., 1993, A compendium to information theory in economics and econometrics, Econometric Reviews, 137–181.

Maddala, G.S., 1983, Limited dependent and qualitative variables in econometrics (Cambridge University Press, Cambridge).

Magee, L., 1990, R^2 measures based on Wald and likelihood ratio joint significance tests, The American Statistician 44, 250–253.

McCullagh, P. and J.A. Nelder, 1989, Generalized linear models, 2nd ed. (Chapman and Hall, London).

McFadden, D., 1974, Conditional logit analysis of qualitative choice behaviour, in: P. Zarembka, ed., Frontiers in econometrics (Academic Press, New York, NY) 105–142.

McKelvey, R.D. and W. Zavoina, 1976, A statistical model for the analysis of ordinal level dependent variables, Journal of Mathematical Sociology 4, 103–120.

Merkle, L. and K.F. Zimmermann, 1992, The demographics of labor turnover: A comparison of ordinal probit and censored count data models, Recherches Economiques de Louvain 58, 283–307.

Nelder, J.A. and R.W.M. Wedderburn, 1972, Generalized linear models, Journal of the Royal Statistical Society A 135, 370–384.

Pregibon, D., 1981, Logistic regression diagnostics, Annals of Statistics 9. 705–724.

Pregibon, D., 1984, Data analytic methods for matched case-control studies, Biometrics 40, 639–651.

Simon, G., 1973, Additivity of information in exponential family probability laws, Journal of the American Statistical Association 68, 478–482.

Ullah, A., 1993, Entropy, divergence and distance measures with econometric applications (Department of Economics, University of California, Riverside, CA).

Veall, M.R. and K.F. Zimmermann, 1992, Pseudo-R^2's in the ordinal probit model, Journal of Mathematical Sociology 4, 103–120.

Veall, M.R. and K.F. Zimmermann, 1995, Pseudo-R^2 measures for some common limited dependent variable models, Journal of Economic Surveys, forthcoming.

Vos, P.W., 1991, A geometric approach to detecting influential cases, Annals of Statistics 19, 1570–1581.

Vuong, Q.H., 1989, Likelihood ratio tests for model selection and non-nested hypothesis, Econometrica 57, 307–333.

White, H., 1993, Estimation, inference and specification analysis (Cambridge University Press, Cambridge).

Williams, D.A., 1987, Generalized linear model diagnostics using the deviance and single-case deletions, Applied Statistics 36, 181–191.

Windmeijer, F.A.G., 1995, Goodness-of-fit measures in binary choice models, Econometric Reviews 14, 101–116.