A Comment on the Coefficient of Determination for Binary Responses

D. R. COX and NANNY WERMUTH*

Linear logistic or probit regression can be closely approximated by an unweighted least squares analysis of the regression linear in the conditional probabilities provided that these probabilities for success and failure are not too extreme. It is shown how this restriction on the probabilities translates into a restriction on the range of the coefficient of determination R^2 so that, as a consequence, R^2 is not suitable to judge the effectiveness of linear regressions with binary responses even if an important relation is present.

KEY WORDS: Discrimination; Logistic model; Multiple correlation.

Much has been written in dispraise of the coefficient of determination R^2 as an overall summary of the effectiveness of a least squares equation. The dependence of R^2 on the spread of the explanatory variables encountered makes it particularly unsuited for comparisons of different studies, where a difference in spread is a feature of the design used, or of selection effects rather than of the system under investigation. However, in any study with many observations, in which the contributions of the explanatory variables are highly significant and of substantive interest, R^2 can be a useful reminder that the additional contribution of some of the variables may, in fact, only explain a small percentage of the variability in the response.

We show that this type of interpretation is misleading in linear regressions with binary responses since low values of R^2, roughly .1, are inevitable even if an important relation is present. Such linear regressions are quite widely used in various fields and, under the major proviso that the range of fitted probabilities is not extreme (e.g., between .2 and .8), are virtually indistinguishable from logistic and probit regressions; see Section 1.

Because the point-biserial coefficient (McNemar 1962) and Pearson's phi-coefficient are measures of linear relationships, they too are severely restricted in their usefulness to judge the strength of an association in just those situations where they appear to be appropriate at first sight; see Section 3.

*D. R. Cox is Warden, Nuffield College, Oxford, OX1 1NF, England. Nanny Wermuth is Professor, Psychologisches Institut, Universität Mainz, Postfach 3980, D-6500 Mainz, Germany. We are grateful to Professor F. Wolinsky, who showed in his lecture at a World Health Organization workshop on Quantitative Methods in Social Medicine (Copenhagen, April 1990), a strong relationship for a binary response together with a low value of R^2 and thus motivated the current work. We also thank the British German Academic Research Collaboration Programme for supporting our work together and Elke Korn for computing the estimates linear in probabilities via maximum-likelihood.

1. THE SIZE OF R^2 FOR BINARY RESPONSES

We consider first the case of a binary response A, with levels $i = 0, 1$, where the unconditional probability of success is denoted by $\pi_1 = \Pr(A = 1)$, and a single random explanatory variable X, with levels x, having mean μ_x and variance σ_x^2. The argument extends immediately to multiple linear regression. We suppose that it is adequate to fit by unweighted least squares the linear regression in the conditional probabilities of success,

$$\pi_{1|x} = \Pr(A = 1 \mid X = x) = E(A \mid X = x)$$
$$= \pi_1 + \beta(x - \mu_x), \qquad (1)$$

where $E_X(\pi_{1|X}) = \pi_1 = 1 - \pi_0$ and $\mathrm{var}_X(\pi_{1|X}) = \beta^2\sigma_x^2$. The result will be a close approximation to the fitting of a linear logistic regression or a linear probit model provided that, say, $.2 \leq \pi_{1|x} \leq .8$ (Cox 1966; Cox and Snell 1989). This is illustrated by Figure 1, which compares logistic, probit, and linear curves scaled to agree at 20% and 80% points.

As in the usual linear random regression, the overall variance of the response can be split up into two parts, the expected conditional variance and the variance due to the linear model since

$$E_X\{\mathrm{var}(A \mid X = x)\} = E_X(\pi_{1|X}\pi_{0|X}) = \pi_1\pi_0 - \beta^2\sigma_x^2$$

so that

$$R^2 = \beta^2\sigma_x^2/(\pi_1\pi_0) = \mathrm{var}_X(\pi_{1|X})\{\pi_1\pi_0\}^{-1}. \qquad (2)$$

The value of R^2 is determined primarily by the variance of X as that determines the variance of the conditional success rate, that is, $\mathrm{var}_X(\pi_{1|X})$, and the factor $(\pi_1\pi_0)^{-1}$ varies only slowly over the range contemplated, in fact, between 4 at $\pi_1 = \pi_0 = .5$ and 6.25 at $\pi_1 = .8$, $\pi_0 = .2$. If we confine $\pi_{1|X}$ to the interval (.2, .8), R^2 is maximized by any two-point distribution taking values at the end points. Such a distribution has $\mathrm{var}_X(\pi_{1|X}) = (.8 - .2)^2\pi_1\pi_0$ and hence leads to $R^2 = .36$.

In Figure 2 several distributions of $\pi_{1|X}$ are displayed, each having mean $\pi_1 = .5$ and satisfying, at least with high probability, $.2 \leq \pi_{1|x} \leq .8$. The most extreme case in one direction is given by Figure 2a, in which the explanatory variable X is binary and thus such that $\pi_{1|X}$ has two equally likely outcomes. Such a case occurs, for example, for randomization between two treatment groups with success rates $\pi_{1|(X=0)} = .2$, $\pi_{1|(X=1)} = .8$ and, as noted previously, it implies $R^2 = .36$. An intermediate case is given by Figure 2b, where the explanatory variable X has a uniform distribution such that $\pi_{1|X}$ is uniformly distributed over (.2, .8). This implies, with $\mathrm{var}_X(\pi_{1|X}) = .6^2/12 = .03$ and (2), that $R^2 = .12$. A more extreme case in the same direction is contained in Figure

© 1992 American Statistical Association

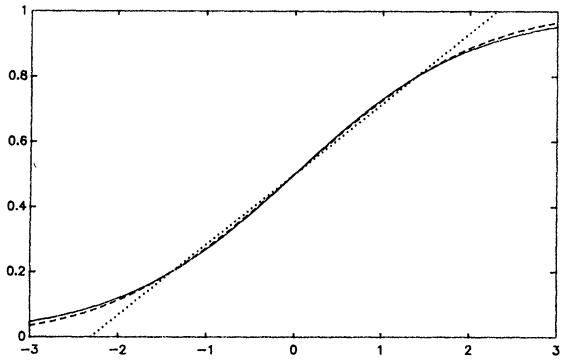

Figure 1. Logistic (solid), Probit (dashed), and Linear (dotted) Lines Scaled to Agree at 20% and 80%. The agreement is excellent if all conditional probabilities lie within the range of .2 to .8.

2c. There the distribution of X, and hence of $\pi_{1|X}$, is normal and $\text{var}(\pi_{1|X}) = .15^2$ to ensure that "most" of the success rates are between .2 and .8, that is, within two standard deviations of $.5 = E_X(\pi_{1|X})$. In this case (2) gives $R^2 = .0225/.25 = .09$.

Thus $R^2 = .36$ is the largest value of the coefficient of determination that can be achieved under circumstances where unweighted least squares fitting of the linear relation (1) in conditional success rates would be at all sensible; see Figure 1.

2. THE SIZE OF R^2 FOR GROUPED BINARY RESPONSES

Suppose, for a little extra generality, that $Y_{k,x}$ is the proportion of ones in k independent binary trials at a level x of the random explanatory variable X.

This corresponds to data from experiments in which equal numbers of individuals have the same value of the explanatory variable. There are two possible ways of applying unweighted least squares then. One application is to the individual binary responses, the other application is to group proportions of successes, that is, for $k = 2$, say, to the possible values 0, .5, or 1 of the response. The regression lines would be identical in both analyses, but the value of R^2 would be substantially larger in the second case. To see this, note that $\pi_{1|x} = E(Y_{k,X} \mid X = x)$, so that (1) is again assumed, but that

$$\begin{aligned}\text{var}(Y_{k,X}) &= E_X[\text{var}(Y_{k,X} \mid X = x)] + \text{var}_X E(Y_{k,X} \mid X = x) \\ &= \{\pi_1\pi_0 + \beta^2\sigma_x^2(k-1)\}/k\end{aligned}$$

because $\text{var}(Y_{k,X} \mid X = x) = \pi_{1|x}\pi_{0|x}/k$. Therefore, it follows, with $\beta^2\sigma_x^2 = \text{var}(\pi_{1|X})$, that

$$R^2 = k\,\text{var}_X(\pi_{1|X})\{\pi_1\pi_0 + \text{var}_X(\pi_{1|X})(k-1)\}^{-1}. \quad (3)$$

The dependence of R^2 on k, corresponding to the three situations in Figure 2, is displayed in Figure 3. For instance, with $k = 4$, the multiple correlation coefficients, respectively, are: $R^2 = .69$, $R^2 = .35$, and $R^2 = .28$.

A similar effect of increasing R^2 by least squares fitting to group means (i.e., proportions) instead of individual responses is to be expected if the group sizes are unequal. In this case the appropriate least squares analysis is that applying to the individual responses. This amounts to weighting the group proportions by the group sizes, while maximum likelihood is equivalent to a more complex iteratively determined weighting scheme.

The low value of R^2 is not to be explained via the inappropriateness of criteria based on sums of squares and least squares fitting. Various generalizations of R^2 for nonnormal models, essentially based on likelihood, have been suggested—for example, the notion of the explanatory power of a hypothesis (Good 1960). The most directly interpretable approach is to compare the fit of a model under analysis with that of a baseline model, for example, in the present context, a model with constant probability of success for all individuals. If $\hat{L}_f$ and $\hat{L}_b$ are the corresponding maximized likelihoods based on n independent observations $P = (\hat{L}_f/\hat{L}_b)^{1/n}$ is the geometric mean "improvement" per observation produced by fitting the more elaborate model. For the normal-theory linear model, $R^2 = 1 - P^{-2}$, so that this equation could be regarded as a likelihood-based generalization. For the two-point example above, treating the baseline model to be one that assigns constant probability 1/2 to the binary response, $P = .8^{.8}\ .2^{.2}/.5 = 1.2126$ so that the newly defined R^2 is .032, that is, again small, essentially because the "improvement" in likelihood per observation is not great.

3. IMPLIED RESTRICTIONS ON THE DISTANCE IN THE MEANS OF THE EXPLANATORY VARIABLE

The restriction of not too extreme conditional probabilities for the linear model (1) to be appropriate not only implies restrictions on the size of the coefficient of determination but also on the distance in means of the ex-

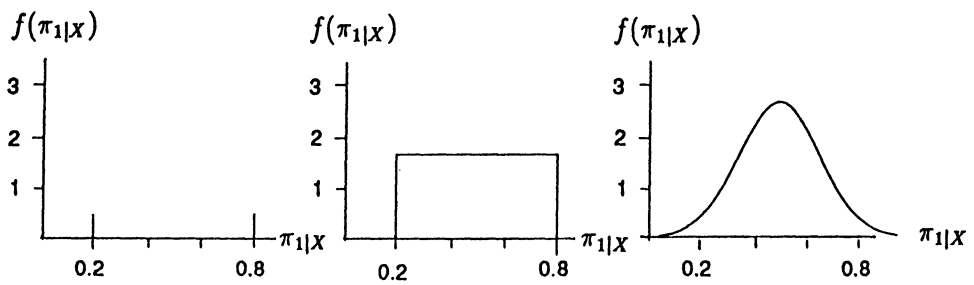

Figure 2. Possible Distributions of the Conditional Probability of Success $\pi_{1|X}$ if the Explanatory Variable is (a) Binary, (b) Equally Distributed, (c) Normally Distributed with $.2 \le \pi_{1|X} \le .8$ and $\pi_1 = .5$.

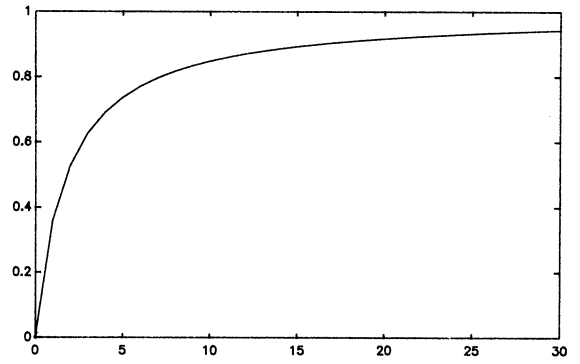

Figure 3. Increase of R^2 as Depending on Group Size k Corresponding to the Three Examples of Figure 2. The explanatory variable is (a) binary (solid), (b) equally distributed (dashed), and (c) normally distributed (dotted line).

planatory variable X, given success $A = 1$ and failure $A = 0$.

To see this we reverse the argument of Section 2, denoting the conditional means of X by $\mu_{x|i}$. The overall mean of X and the covariance between the binary response A and X can then be expressed as

$$\mu_x = E_A E(X \mid A = i) = \pi_0 \mu_{x|0} + \pi_1 \mu_{x|1}$$

and

$$\begin{aligned} \operatorname{cov}(A, X) &= E_A\{iE(X \mid A = i)\} - E(A)E(X) \\ &= \pi_0 \pi_1 (\mu_{x|1} - \mu_{x|0}). \end{aligned}$$

Furthermore, the coefficient of determination (2) can be written as

$$R^2 = \frac{\operatorname{cov}(A, X)^2}{\operatorname{var}(A)\sigma_x^2} = \frac{(\mu_{x|1} - \mu_{x|0})^2 \pi_0 \pi_1}{\sigma_x^2} \tag{4}$$

because $\beta = \operatorname{cov}(A, X)/\sigma_x^2$ and $\operatorname{var}(A) = \pi_0\pi_1$. For a binary explanatory variable this reduces, with $\theta_{ij} = \Pr(A = i, X = j)$ and $\delta_j = \Pr(X = j)$, to

$$R^2 = \frac{(\theta_{00}\theta_{11} - \theta_{01}\theta_{10})^2}{\pi_0 \pi_1 \delta_0 \delta_1} \tag{5}$$

because in that case $\mu_{x|i} = \theta_{i1}/\pi_i$ so that $\operatorname{cov}(A, X) = \theta_{00}\theta_{11} - \theta_{01}\theta_{10}$. Equation (4), together with the result of Section 2, implies that the distance in means $|\mu_{x|1} - \mu_{x|0}|$ will range from $1.2\sigma_x$ at $\pi_1 = \pi_0 = .5$, to $1.5\sigma_x$ at $\pi_1 = .8$, $\pi_0 = .2$.

Furthermore, the square roots of the right sides in Equations (4) and (5) are just population equivalents of McNemar's point biserial coefficient and of Pearson's phi-coefficient, respectively. Thus, these coefficients represent special ways of expressing the usual standardized measure for linear association, the simple product-moment correlation coefficient. However, they do not vary between -1 and 1. Instead, in situations in which unweighted least squares fitting of the linear association between a binary response A and an explanatory variable X is sensible, the largest value of the phi-coefficient is .6 even if $\pi_1 = \delta_1 = .5$ and a typical value of McNemar's coefficient is still considerably smaller.

A formally closely related but conceptually different problem is discriminant analysis with two normal distributions having the same variance and an associated conditional logistic regression. Remarkably (Fisher 1938) the discriminant function can be estimated by a formal linear regression with population groups as binary response on the continuous variable even though the conditional relation is linear logistic. In this situation large values of R^2 can be achieved via populations whose means are so far apart that a linear regression would be a poor approximation indeed to the conditional binary logistic relation.

4. EXAMPLE

In a study on patients from a pain clinic (Schmitt 1990) an ordinal variable called "stage of chronic pain" of a patient has been constructed and has been related to success of stationary pain treatment. For 58 male patients the results displayed in Table 1 were obtained. There is a clear linear relation in the conditional probabilities: The higher the stage of chronic pain, the lower is the probability of a successful treatment, but $R^2 = .124$ for the

Table 1. Scores for Stage of Chronic Pain (x), Counts, and Estimated Probabilities of Success of Treatment When Leaving the Clinic (A)

Stage of chronic pain x	Total count n_x	Number of successes $n_{1\|x}$	Probabilities of treatment success estimated by: Observed relative frequencies $n_{1\|x}/n_x$	Linear logistic regression[a]	Linear probit analysis[b]	Linear regression in binary responses via: Least squares[c]	Maximum likelihood[d]
6	8	7	.88	.75	.76	.75	.78
7	9	5	.56	.64	.64	.63	.64
8	15	6	.40	.50	.50	.50	.51
9	14	6	.43	.36	.36	.37	.37
10	10	3	.30	.24	.24	.24	.23
11	2	0	.00	.15	.15	.11	.01

NOTE: Sample size $n = 58$ patients.
[a] $\ln(\hat{\pi}_{1|x}/\hat{\pi}_{0|x}) = -4.52 + .57x$.
[b] $\Phi^{-1}(\hat{\pi}_{1|x}) = 2.82 - .35x$.
[c] $\hat{\pi}_{1|x} = 1.52 - .13x$.
[d] $\hat{\pi}_{1|x} = 1.61 - .14x$.

regression with individual responses. With group proportions $R^2 = .813$ for weights equal to group sizes.

[*Received May 1990. Revised November 1990.*]

REFERENCES

Cox, D. R. (1966), "Some Procedures Connected With the Logistic Qualitative Response Curve," in *Research Papers in Statistics: Essays in Honour of J. Neyman's 70th Birthday,* ed. F. N. David, London: John Wiley, pp. 55–71.

Cox, D. R., and Snell, E. J. (1989), *Analysis of Binary Data* (2nd ed.), London: Chapman and Hall.

Fisher, R. A. (1938), "The Statistical Utilization of Multiple Measurements," *The Annals of Eugenics,* 8, 376–386.

Good, I. J. (1950), "Weight of Evidence, Corrobaration, Explanatory Power and the Utility of Experiments," *Journal of the Royal Statistical Society,* Ser. B, 22, 319–331.

McNemar, Q. (1962), *Psychological Statistics* (3rd ed.), New York: John Wiley.

Schmitt, N. (1990), "Stadieneinteilung chronischer Schmerzen," unpublished medical dissertation, University of Mainz.