Explained Residual Variation, Explained Risk, and Goodness of Fit

EDWARD L. KORN and RICHARD SIMON*

A loss function approach is used to define the concepts of explained residual variation and explained risk for general regression models. Explained risk measures the ability of the covariates in a correctly specified model to distinguish differing outcomes. Explained residual variation, which is R^2 for a linear model, estimates the explained risk with a penalty for poorly fitting models. Application of the general definitions to linear regression, logistic regression, and survival analysis is given. The importance of distinguishing the concepts of explained residual variation, explained risk, and goodness of fit is discussed.

KEY WORDS: Binary data; Coefficient of determination; Loss function; R^2; Survival analysis.

1. INTRODUCTION

The notion of a decrease in residual variation due to the utilization of a regression model actually incorporates two concepts: (1) the applicability of the model to the data and (2) the ability of the covariates in the model to distinguish differing outcomes. We shall refer to the first concept as "goodness of fit" and the second concept as "explained risk." Goodness of fit is being used here with its historical meaning of the consistency of the model with the data (Pearson 1900); it does not address the question as to whether the model is "useful" or whether better models using other covariates may fit the data. "Explained risk" is one way of quantifying how much better predictions are when using the covariates compared to when not using them; this will be made precise in the next section. The proportional decrease in residual variation, the "explained residual variation," incorporates both concepts. In part because of this, there is the potential for confusion. For the linear regression model, the usual coefficient of determination (squared multiple correlation coefficient), R^2, is the explained residual variation.

Two examples will now be given to help clarify the nomenclature. The first example involves simple linear regressions on two data sets each with eight observations at each of $x = 1, 2, \ldots 8$. Figure 1 contains plots of the data sets that have the same value of R^2, .64. Thus the explained residual variation of these regressions for both data sets is the same. However, this simple model is clearly only adequate for the first data set. An adequately fitting covariate model for the second data set would have two steps corresponding to $x \leq 4$ and $x \geq 5$. If this adequately specified model is used to fit the second data set, then $R^2 = .78$. One can think of the drop from .78 to .64 as the penalty for using a poorly fitting model.

For the second example, consider binomial data 45/100, 50/100, and 55/100 corresponding to values 1, 2, and 3, respectively, of the covariate x. A simple linear logistic regression yields predicted proportions of .45, .50, and .55, for $x = 1$, 2, and 3. The goodness of fit is perfect for this model. However, the explained risk is low; the covariate is not very useful in distinguishing outcomes. The explained residual variation of this model for this data is also low. At first glance this conclusion seems incorrect, given the closeness of the observed and predicted proportions for $x = 1$, 2, and 3. When viewed on an individual-by-individual basis, however, the observations are all zeros and ones while the predictions are either .45, .50, or .55. Viewed in this way, we see that the explained residual variation is low for this model, since the observed and predicted values are not close.

In the next section we give general definitions of explained residual variation and explained risk using a loss-function approach. This is followed in Section 3 by examples involving linear regression, binary regression, and survival analysis. We concentrate on the population parameters being estimated, rather than on the finite sampling properties of various statistics. We do consider, however, the realistic possibility that the class of models used to fit the data is misspecified. This enables us to examine the effects of poor goodness of fit on the (large-sample) properties of the explained residual variation. We end with a discussion of why we believe it is useful to distinguish the concepts of explained residual variation, explained risk, and goodness of fit.

2. GENERAL DEFINITIONS

Let $L(y, \tilde{y})$ be the loss incurred in making a prediction of $\tilde{y}$ for true observation y. Given covariate value x, the expected loss is $\int L(y, \tilde{y})\, dF(y \mid x)$. Let $\tilde{y}(x)$ be the value that minimizes this expected loss when $F(y \mid x)$ is known, and let the risk, $R(x)$, be this minimized expected loss. For example, with squared error loss, $\tilde{y}(x)$ and $R(x)$ are the conditional mean and conditional variance of Y. When the covariates are not used to predict y, we first define the null model as the mixture of the conditional models over the distribution of the covariates, that is,

$$F_0(y) \equiv \frac{1}{N} \sum_{i=1}^{N} F(y \mid x_i),$$

where the set of covariates $\{x_1, x_2, \ldots, x_N\}$ is assumed

*Edward L. Korn is Mathematical Statistician and Richard Simon is Chief, Biometric Research Branch, National Cancer Institute, Bethesda, MD 20892. The authors thank the referees for their helpful comments.

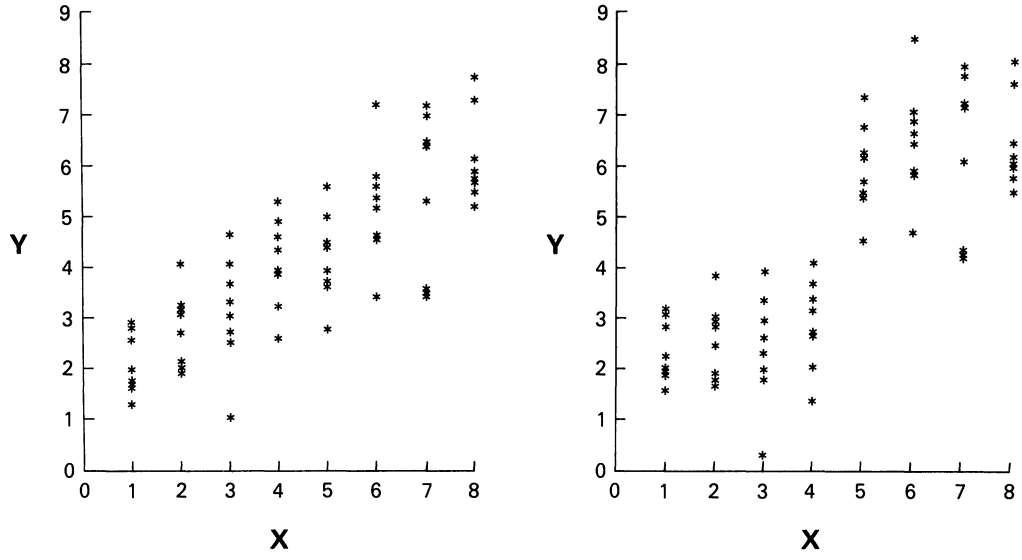

Figure 1. Two Data Sets for Which the Simple Linear Regressions Have R^2 = .64.

fixed or the cdf $F_X(x)$ is assumed given. Note that, if $F(y \mid x)$ is a regression model, then this null model is not the same as the regression model with no covariates. For example, in a normal linear regression model $F_0(y)$ is a mixture of normal distributions with varying means. We prefer our choice of null model for the following reason: Although we are interested in considering a prediction that does not utilize the covariates, we will usually not seriously consider the data to be consistent with this model. Thus, computation of a risk under the model with no covariates does not seem particularly relevant. On the other hand, the use of $F_0(y)$ computes the risk associated with not using the covariates under the realistic covariate model. Using $F_0(y)$, the null risk is defined as $R_0 = \int L(y, \tilde{y}_0)\, dF_0(y)$, where $\tilde{y}_0$ minimizes the expected loss with respect to $F_0(y)$.

The explained risk is defined as the proportional decrease in risk obtained by using the covariates:

$$\text{explained risk} \equiv \frac{R_0 - E[R(X)]}{R_0},$$

where $E[R(X)]$ is the expected value of $R(x)$ averaged over the distribution of the x's. Note that the explained risk is a population quantity; it depends only on the model and the distribution of the x's but not on the observed values of Y. Although many choices of loss function are available, squared error loss will usually be considered. For squared error loss, $R(x) = \text{var}(Y \mid x)$ and $R_0 = \text{var}_0(Y) = E[\text{var}(Y \mid X)] + \text{var}[E(Y \mid X)]$, yielding

$$\text{explained risk} = \frac{\text{var}[E(Y \mid X)]}{E[\text{var}(Y \mid X)] + \text{var}[E(Y \mid X)]}. \quad (2.1)$$

Since $F(y \mid x)$ will usually not be completely specified, the explained risk will need to be estimated from the (x_i, y_i) data at hand. One possible estimator is the explained residual variation,

$$\text{explained residual variation} \equiv \frac{\sum_{i=1}^N L(y_i, \hat{y}_0) - \sum_{i=1}^N L(y_i, \hat{y}(x_i))}{\sum_{i=1}^N L(y_i, \hat{y}_0)}, \quad (2.2)$$

where $\hat{y}_0$ and $\hat{y}(x_i)$ are estimators of $\tilde{y}_0$ and $\tilde{y}(x_i)$. For example, for a linear regression model, $y_i = \alpha + x_i'\beta + e_i$, the explained residual variation using squared error loss equals the usual R^2 when the maximum likelihood estimators $\hat{y}(x_i) = x_i'\hat{\beta}$ and $\hat{y}_0 = \Sigma\, y_i/N$ are used (see Sec. 3). In general, under suitable regularity conditions, the explained residual variation is a consistent estimator of the explained risk provided that the modeling assumptions used to estimate $\tilde{y}_0$ and $\tilde{y}(x_i)$ are correct. If the modeling assumptions are incorrect, then, in general, $1/N\, \Sigma_{i=1}^N L(y_i, \hat{y}(x_i))$ will converge to something larger than $E[R(X)]$. In many situations $\hat{y}_0$ will remain a consistent estimator of $\tilde{y}_0$ even under a misspecified model. In these situations, therefore, the explained residual variation will be, for large samples, an underestimate of the explained risk of the correctly specified model using the x's. In this sense, explained residual variation measures both explained risk and goodness of fit. In some applications, as will be seen later, other estimators of the explained risk exist that are not necessarily lowered by model misspecification. Care must be taken to interpret these estimators differently than explained residual variation when there is the possibility of a poorly fitting model.

3. EXAMPLES

Linear Regression

Let $y_i = \alpha_t + x_i'\beta_t + e_i$, where the e_i are iid with mean zero and variance σ_t^2, be the "true" model, that is, a model that is consistent with the data. The usual coefficient of determination, R^2, is defined by $R^2 = 1 - SSE/SST$, where SSE and SST are the error sum of squares

and the total sum of squares. Using squared error loss and maximum likelihood estimators of β_t and σ_t^2, R^2 is the explained residual variation as defined previously. For large samples, R^2 estimates the population quantity $\rho^2 = \beta_t' S_X \beta_t / (\beta_t' S_X \beta_t + \sigma_t^2)$, where S_X is the covariance of the x's (Helland 1987). Since $E(Y \mid x) = x'\beta_t$ and $\text{var}(Y \mid x) = \sigma_t^2$, the explained risk (squared error loss) for this model is, by (2.1), precisely ρ^2. Notice that even with a correctly specified model, the explained risk contains components relating to the steepness of the regression as well as the residual error (Barrett 1974).

Assume now that an incorrectly specified model is fit to the data: $y_i = \alpha + z_i'\gamma + e_i$, where the $z_i = g(x_i)$ and $g(x)$ is a vector (nonlinear) function of the vector x. Using this incorrect model, R^2 now estimates $\rho_Z^2 = \gamma_t' S_Z \gamma_t / (\beta_t' S_X \beta_t + \sigma_t^2)$ for large samples, where S_Z is the covariance of the z's, and γ_t is the least squares estimate of the regression coefficient from the regression of $x_i'\beta_t$ on z_i. It can be shown that $\rho_z^2 \leq \rho^2$ with equality only if there exists a γ such that $x_i'\beta_t = z_i'\gamma$ for all i. This is an example of the general statement that explained residual variation estimated using a misspecified model will be a (large-sample) underestimate of the explained risk of a model that is consistent with the data.

We now briefly digress to consider a linear regression model without an intercept term to expand on our choice of null risk R_0 in the definition of explained risk. Recall that we have chosen the null model to be a mixture model over the covariates. Thus the presence or absence of an intercept term does not change the definition of explained risk. In a linear regression without an intercept, however, a reasonable alternative choice for the null model would be $E(Y \mid x) \equiv 0$. Using squared error loss, this leads to a risk of $R_0 = E(Y^2)$ and an explained residual variation $R^2 = 1 - SSE/SSU$, where SSU is the uncorrected total sum of squares (Judge, Griffiths, Hill, Lutkepohl, and Lee 1985, pp. 30–31; Kvalseth 1985). As described previously, we prefer our choice of null model since we will usually not seriously consider the data to be consistent with the model $E(Y \mid x) \equiv 0$.

Binary Regression

Let Y_i be independent Bernoulli observations and assume that $p(x) \equiv \Pr(Y_i = 1 \mid x)$. Then $E(Y \mid x) = p(x)$ and $\text{var}(Y \mid x) = p(x)[1 - p(x)]$. Using squared error loss, $\tilde{y}(x) = p(x)$, and

$$\text{explained risk} = \frac{\text{var}[p(X)]}{E[p(X)]\{1 - E[p(X)]\}}. \quad (3.1)$$

Note that the predicted values $\tilde{y}(x)$ have range [0, 1] while the observations have values {0, 1}. The predictions are not of the Y_i themselves, but instead are the probabilities that $Y_i = 1$. With this interpretation, the loss function approach to defining explained risk still applies. It is instructive to consider some special cases of binary regression: For the K-group problem, $\Pr(Y_i = 1 \mid X = k) = \pi_k$ $(k = 1, 2, \ldots K)$. If we let n_k equal the number of observations for which $X = k$, then

$$\text{explained risk} = \frac{\frac{1}{N}\sum n_k(\pi_k - \bar{\pi})^2}{\bar{\pi}(1 - \bar{\pi})},$$

where $\bar{\pi} = \Sigma\, n_k\pi_k/N$. As shown by Margolin and Light (1974), this is equivalent to the definition of τ_b given by Goodman and Kruskal (1954). It is easy to check that substituting the observed proportions $\hat{\pi}_k$ for π_k, and $\hat{\bar{\pi}}$ for $\bar{\pi}$ into the above expression yields the explained residual variation,

$$\text{explained residual variation} = \frac{\frac{1}{N}\sum n_k(\hat{\pi}_k - \hat{\bar{\pi}})^2}{\hat{\bar{\pi}}(1 - \hat{\bar{\pi}})}.$$

This is equivalent to the definition of R^2 given by Light and Margolin (1971) for this problem; see Efron (1978) for a discussion and the use of other possible loss functions. Since for a K-group problem the conditional mean given the covariates is always correctly specified, the goodness-of-fit question does not arise.

Next consider a logistic regression model for which the true model is logit $p(x) = x$, and X has a uniform distribution on $[-1, 1]$. For this model and distribution of X, $E[p(X)] = .5$ and $\text{var}[p(X)] = .019$ yielding an explained risk (squared error loss) of .076. Another commonly used loss function for binary data is entropy loss (Efron 1978): $L(y, \hat{p}) = -2[y \log(\hat{p}) + (1 - y) \log(1 - \hat{p})]$. For this loss function, $E[R(X)] = -2E\{p(X) \log p(X) + [1 - p(X)] \log[1 - p(X)]\}$ and $R_o = -2[p_0 \log p_0 + (1 - p_0) \log(1 - p_0)]$, where $p_0 = E[p(X)] = .5$, and recall that the expectation (E) is over the distribution of the x's. Using the logit model and distribution of X specified above, the explained risk (entropy loss) is calculated numerically to be .056. Suppose that this true logit model was unknown and to be estimated from data using the model logit $p(x) = \alpha + \beta x$. If the consistent maximum likelihood estimates are used to estimate (α, β), then consistent estimates $\hat{p}(x_i)$ of the predicted values will be obtained $(i = 1, 2, \ldots N)$. Substituting the sample mean and variance of these $\hat{p}_i$ into (3.1) yields a consistent estimate of the explained risk (squared error loss),

$$\text{estimated explained risk (squared error loss)} = \frac{\frac{1}{N}\sum (\hat{p}(x_i) - \hat{p}_0)^2}{\hat{p}_0(1 - \hat{p}_0)}, \quad (3.2)$$

where $\hat{p}_0 = \Sigma\, \hat{p}(x_i)/N = \Sigma\, Y_i/N$. Similarly, substitution of the predicted values into the formulas of $E[R(X)]$ for entropy loss yields a consistent estimate of the explained risk (entropy loss),

$$\text{estimated explained risk (entropy loss)} = 1 - \frac{\frac{1}{N}\sum \{\hat{p}(x_i) \log \hat{p}(x_i) + [1 - \hat{p}(x_i)] \log[1 - \hat{p}(x_i)]\}}{\hat{p}_0 \log(1 - \hat{p}_0) + (1 - \hat{p}_0) \log(1 - \hat{p}_0)}. \quad (3.3)$$

Haberman (1982) describes some properties of these estimators as well as generalizations to multinomial regression data.

The explained residual variations for this logistic regression model are as follows. For squared error loss,

explained residual variation (squared error loss)

$$= 1 - \frac{\sum (y_i - \hat{p}(x_i))^2}{\sum (y_i - \hat{p}_0)^2}$$

$$= \frac{\frac{2}{N}\sum y_i \hat{p}(x_i) - \frac{1}{N}\sum \hat{p}(x_i)^2 - \hat{p}_0^2}{\hat{p}_0(1 - \hat{p}_0)}. \qquad (3.4)$$

While for entropy loss,

explained residual variation (entropy loss)

$$= 1 - \frac{\sum \{y_i \log \hat{p}(x_i) + (1 - y_i) \log[1 - \hat{p}(x_i)]\}}{\sum y_i \log(1 - \hat{p}_0) + (1 - y_i) \log(1 - \hat{p}_0)}. \qquad (3.5)$$

The expression (3.5) is precisely the proportional reduction in maximized log-likelihood when using the model compared to the maximized log-likelihood of the model with intercept parameter only. Except for a small sample correction used to penalize for the number of parameters estimated, it has been suggested by Harrell (1986a) for logistic regression models (his R^2). It is also interesting to note that for the logistic regression model the explained residual variation (entropy loss) equals the estimated explained risk (entropy loss). This follows from the equations defining the maximum likelihood estimates of the regression coefficients, $\Sigma\, y_i = \Sigma\, \hat{p}(x_i)$ and $\Sigma\, y_i x_i = \Sigma\, \hat{p}(x_i)x_i$. For the particular logistic regression model considered previously, the explained residual variations are consistent estimators of the (true) explained risks, .076 and .056, for squared error and entropy loss. Notice that repeat x values are not required for either the estimated explained risks or the explained residual variations.

To examine the behavior of these estimated explained risks and explained residual variations under misspecified binary regression models, we first consider two hypothetical extreme examples and then some real data. For simplicity we restrict attention to squared error loss; similar results are obtained for entropy loss. Consider data being generated by the model logit $p(x) = x$, for x's taking on the six values -1.5, -1, $-.5$, .5, 1, and 1.5 each with probability 1/6. The $p(x)$'s corresponding to the design points are .18, .27, .38, .62, .73, and .82. The explained risk is .23. Now suppose the following incorrect model is used to fit the data: $p(x) = .1$ for $x < 0$, and $p(x) = .9$ for $x \geq 0$. The explained residual variation using this assumed model will consistently estimate .08, an underestimate of the explained risk of the true model, .23. The (estimated) explained risk of the assumed model, however, is .64. On the other hand, suppose we fit the data using the model: $p(x) = .4$ for $x < 0$, and $p(x) = .6$ for $x \geq 0$. The explained residual variation using this assumed model will consistently estimate .14, although the explained risk is .04. As a measure of the explained risk of the true covariate model, the explained residual variation is, as expected, reduced by an incorrectly specified model for these examples. Since this is not true for the estimated explained risks, care should be taken, in theory, that these quantities are interpreted differently when there is the possibility of a poorly fitting model. For example, a large estimated explained risk using (3.2) or (3.3) does not necessarily imply that the goodness of fit is high.

In practice, unless fitted models are grossly inconsistent with the data, there may be only small differences in the estimated explained risks and explained residual variations. We now give some examples. Efron (1978) considered binary toxoplasmosis data on 697 subjects sampled from 34 cities in El Salvador. For these data, Efron showed that a K-group model using cities as the grouping variable fits the data significantly better than a linear logistic model using the annual rainfall of each city as the covariate. Using the linear logistic model and squared error loss, the estimated explained risk and the explained residual variation are both less than .001. Another example is given by Haberman (1982), who considered the attitudes of 1,305 male subjects toward women staying at home. Haberman fit the data with a linear logistic model with the covariate being the number of years of education of the respondent. The estimated explained risk and explained residual variation using squared error loss for this model are .121 and .122. To misspecify the model, we instead use as a covariate the square root of the number of years of education. The estimated explained risk and explained residual variation for this misspecified model are .114 and .118. For a final example, consider the data of Lee (1974) which consist of complete remission status (yes or no) and six covariates on 27 cancer patients. Harrell (1986a) performed a stepwise logistic regression of these data with an entry criterion of p value $\leq .3$, which led to the inclusion of three covariates along with the intercept. Since the second covariate enters the regression with a p value of .26, we restrict attention here to the model that includes only the intercept and the single most statistically significant covariate ("LI"). The estimated explained risk and explained residual variation using squared error loss for this fitted model are .298 and .270. To misspecify the model, we use the square of LI as the covariate. The estimated explained risk and explained residual variation using squared error loss for this misspecified model are .24 and .23. We see that for these examples, the estimated explained risks and explained residual variations using squared error loss are almost identical even with misspecified models. Since these models were all logistic regressions using entropy loss, the estimated explained risk would actually equal the explained residual variation for each of these examples.

Survival Analysis

The distinguishing feature of survival data is the presence of censored observations. Because of these observations, the immediate application of the loss function approach to explained residual variation described in Section 2 cannot be applied. However, the explained risk can still frequently be estimated. Consider first a K-group problem with Kaplan–Meier survival curve estimates of $1 - F(y \mid k)$ for $k = 1, 2, \ldots, K$. To estimate the explained risk with squared error loss, the mean and variance of $F(y \mid k)$ are required. These can be directly estimated from the Kaplan–Meier curves provided that the curves drop to zero. With long-term survivors or with considerable censoring, this may not be true. One possible solution to this problem is to use a loss function that might be more relevant for survival data (Korn and Simon 1990). For example, the loss can be taken to be zero for any prediction of survival over ten years for an observed survival over ten years. Additionally, we may wish to incur less loss with a prediction of six years for an observed survival of five years than for a prediction of two years for an observed survival of one year. These considerations might suggest a loss function of the form $L(y, \hat{y}) = \log[\min(y, 10) + c] - \log[\min(\hat{y}, 10) + c]$, where c is some constant. Use of such a loss function will minimize or eliminate the effect of any extrapolation of the Kaplan–Meier curves down to zero. Once a loss function is decided on, the individual Kaplan–Meier curves can be used to estimate $E[R(X)]$ as follows: For group k, let $\hat{F}(t \mid k)$ be one minus the Kaplan–Meier survival curve estimate, and let $\hat{y}_k$ be the value of y which minimizes

$$\sum_i L(t_i, y)[\hat{F}(t_i \mid k) - \hat{F}(t_i^- \mid k)], \tag{3.6}$$

where $t_1 < t_2 < \cdots < t_i < \cdots$ are the distinct death times in group k and $\hat{F}(t_i^- \mid k)$ is the proportion dead just before time t_i. For general loss functions, $\hat{y}_k$ will need to be found numerically. Then $E[R(X)] = 1/N \sum n_k R_k$, where R_k equals (3.6) with $\hat{y}_k$ substituted for y and n_k is the number of observations in group k. To estimate the null risk, R_0, the estimated survival distribution of the mixture distribution $\hat{F}_0(y) = 1/N \sum n_k \hat{F}(y \mid k)$ can be used in a similar manner, where now the t_i range over the death times in all the groups. As with the other K-group problems discussed, the question of model adequacy does not arise.

If a parametric regression is modeled for the survival data, then the explained risk will be a function of the unknown parameters. Substitution of consistent estimates of these parameters will yield a consistent estimate of the explained risk. It is difficult to say in general what the impact will be of estimating these parameters using a misspecified model. For example, an incorrectly specified linear model with unknown scale parameter will cause a decrease in the estimated explained risk using squared error loss (Section 3). On the other hand, it is easy to specify an exponential regression model with known scale parameter for which a misspecified model will increase the estimated explained risk. The specification of general conditions on models for one or the other type of behavior is an area for further research. If, however, a parametric or semiparametric model is used only to define a small number of risk groups based on the covariates, then the explained risk of this grouping can be estimated as a K-group problem.

4. DISCUSSION

In this article we have given examples to demonstrate that explained residual variation measures both the explained risk and the goodness of fit of a covariate model. Since there will usually be interest in both these concepts, why bother to separate them? We believe the separation may be helpful in clarifying some points of confusion. For example, a commonly voiced concern with the usual coefficient of determination, R^2, is that when there are repeated x values, R^2 will always be less than 1.0 (Chang and Afifi 1987; Draper and Smith 1981, p. 42; Healy 1984). Additionally, R^2 is very dependent on the distribution of the x's (Bock and Herrendorfer 1976; Helland 1987; Ranney and Thigpen 1981; Weisberg 1985, pp. 73–76). If R^2 is viewed as a measure of goodness of fit, then these *are* disturbing properties. Viewing R^2 as a measure of explained risk, however, it is quite natural that it should be less than 1.0 when there are repeat x values; in this situation the covariate model has not explained all the risk. Additionally, choosing data points with more dispersed x values leads to more dispersed y values and more risk to explain. Therefore, for a given covariate model, we would expect the explained risk to be higher with more dispersed x values.

Another area of possible confusion is in the use of these measures with binary data (e.g., logistic regression). For these applications, there are measures of goodness of fit that compare grouped observed and expected values; see Hosmer and Lemeshow (1989, pp. 136–149) for a review. One can also calculate measures of explained residual variation, for example, as the proportional reduction in log-likelihood (Harrell 1986a). Finally, there also exist other measures of explained risk (Haberman 1982). Which of these measures to use in a specific application may not be obvious, however, classifying the measures into these three categories can be helpful. For example, one would not want to use a goodness-of-fit measure as the criterion for choosing variables in a stepwise logistic regression.

The loss function approach given in Section 2 is just one approach to explained risk and explained residual variation. For example, Kent (1983) and Kent and O'Quigley (1988) suggested measures related to the Kullback–Leibler information gain. Harrell (1986b) used the proportional likelihood explained in a semiparametric survival analysis. Unlike the same approach for logistic regression, this cannot be considered an explained residual variation of the form (2.2). A completely different approach is to measure rank correlations between

the observed and predicted values (Harrell, Lee, Califf, Pryor, and Rosati 1984; Harrell 1986a). We suggest that with whatever approach is taken, it is useful to distinguish between measures of explained residual variation, explained risk, and goodness of fit.

[*Received December 1989. Revised May 1990.*]

REFERENCES

Barrett, J. P. (1974), "The Coefficient of Determination—Some Limitations," *The American Statistician*, 28, 19–20.

Bock, J., and Herrendorfer, G. (1976), "The Use of the Coefficient of Determination in Linear Regression I," *Biometrische Zeitschrift*, 18, 251–257.

Chang, P. C., and Afifi, A. A. (1987), "Goodness-of-Fit Statistics for General Linear Regression Equations in the Presence of Replicated Responses," *The American Statistician*, 41, 195–199.

Draper, N. R., and Smith, H. (1981), *Applied Regression Analysis* (2nd ed.), New York: John Wiley.

Efron, B. (1978), "Regression and ANOVA With Zero-One Data: Measures of Residual Variation," *Journal of the American Statistical Association*, 73, 113–121.

Goodman, L. A., and Kruskal, W. H. (1954), "Measures of Association for Cross Classifications," *Journal of the American Statistical Association*, 49, 732–764.

Haberman, S. J. (1982), "Analysis of Dispersion of Multinomial Models," *Journal of the American Statistical Association*, 77, 568–580.

Harrell, F. E., Jr., Lee, K. L., Califf, R. M., Pryor, D. B., and Rosati, R. A. (1984), "Regression Modelling Strategies for Improved Prognostic Prediction," *Statistics in Medicine*, 3, 143–152.

Harrell, F. E., Jr. (1986a) "The LOGIST Procedure," in *SUGI Supplemental Library User's Guide* (Ver. 5 ed.), Cary, NC: SAS Institute, Inc, pp. 269–293.

——— (1986b) "The PHGLM Procedure," in *SUGI Supplemental Library User's Guide* (ver. 5 ed.), Cary, NC: SAS Institute Inc, pp. 437–466.

Healy, M. J. R. (1984), "The Use of R^2 as a Measure of Goodness of Fit," *Journal of the Royal Statistical Society*, Ser. A, 147, 608–609.

Helland, I. S. (1987), "On the Interpretation and Use of R^2 in Regression Analysis," *Biometrics*, 43, 61–69.

Hosmer, D. W., Jr., and Lemeshow, S. (1989), *Applied Logistic Regression*, New York: John Wiley.

Judge, G. G., Griffiths, W. E., Hill, R. C., Lutkepohl, H., and Lee, T-C (1985), *The Theory and Practice of Econometrics* (2nd ed.), New York: John Wiley.

Kent, J. T. (1983), "Information Gain and a General Measure of Correlation," *Biometrika*, 70, 163–173.

Kent, J. T., and O'Quigley, J. (1988), "Measures of Dependence for Censored Survival Data," *Biometrika*, 75, 525–534.

Korn, E. L., and Simon, R. (1990), "Measures of Explained Variation for Survival Data," *Statistics in Medicine*, 9, 487–503.

Kvalseth, T. O. (1985), "Cautionary Note About R^2," *The American Statistician*, 39, 279–285.

Lee, E. T. (1974), "A Computer Program for Linear Logistic Regression Analysis," *Computer Programs in Biomedicine*, 4, 80–92.

Light, R. J., and Margolin, B. H. (1971), "An Analysis of Variance for Categorical Data," *Journal of the American Statistical Association*, 66, 534–544.

Margolin, B. H., and Light, R. J. (1974), "An Analysis of Variance for Categorical Data, II: Small Sample Comparisons With Chi Square and Other Competitors," *Journal of the American Statistical Association*, 69, 755–764.

Pearson, K. (1900), "On the Criterion That a Given System of Deviations From the Probable in the Case of a Correlated System of Variables Is Such That It Can Be Reasonably Supposed to Have Arisen From Random Sampling," *Philosophical Magazine*, Ser. 5, 50, 157–173.

Ranney, G. B., and Thigpen, C. C. (1981), "The Sample Coefficient of Determination in Simple Linear Regression," *The American Statistician*, 35, 152–153.

Weisberg, S. (1985), *Applied Linear Regression*, (2nd ed.) New York: John Wiley.