

Predictive accuracy and explained variation[‡]

Michael Schemper^{*,†}

*Section of Clinical Biometrics, Department of Medical Computer Sciences, Vienna University, Spitalgasse 23,
A-1090 Vienna, Austria*

SUMMARY

Measures of the predictive accuracy of regression models quantify the extent to which covariates determine an individual outcome. Explained variation measures the relative gains in predictive accuracy when prediction based on covariates replaces unconditional prediction. A unified concept of predictive accuracy and explained variation based on the absolute prediction error is presented for models with continuous, binary, polytomous and survival outcomes. The measures are given both in a model-based formulation and in a formulation directly contrasting observed and expected outcomes. Various aspects of application are demonstrated by examples from three forms of regression models. It is emphasized that the likely degree of absolute or relative predictive accuracy often is low even if there are highly significant and relatively strong covariates. Copyright © 2003 John Wiley & Sons, Ltd.

KEY WORDS: Cox regression; general linear model; logistic regression; Poisson regression; prediction error

1. INTRODUCTION

One of the purposes of regression is to enable prediction of future outcomes. Occasionally the question arises if covariate knowledge of highly significant and strong covariate effects also leads to substantially improved predictability. Thus we are interested in quantifying predictive accuracy with and without the use of covariates. For this purpose, absolute measures of predictive accuracy will be suggested which are related to corresponding relative measures, some of which have been termed measures of explained variation. Both the absolute and relative measures summarize the extent to which covariates determine an outcome, or similarly, how much is known about mechanisms and conditions affecting an outcome. In this sense such measures can also be used to quantify knowledge in a given empirical research area.

In fact, ‘predictability’ and ‘understanding’ in empirical sciences are related notions. Completeness of ‘understanding’ or knowledge of a certain empirical process can generally be

* Correspondence to: Michael Schemper, Section of Clinical Biometrics, Department of Medical Computer Sciences, Vienna University, Spitalgasse 23, A-1090 Vienna, Austria.

† E-mail: michael.schemper@akh-wien.ac.at

‡ Presented at the International Society for Clinical Biostatistics Twenty-Second International Meeting, Stockholm, Sweden, 19–23 August 2001.

quantified by the departures of predicted from observed outcomes. In this paper we shall focus on regression models as an important tool to study mechanisms and conditions affecting an outcome. Furthermore, we shall use the terminology of 'prediction' though the concepts presented can similarly be used to quantify the explanatory power of models aimed solely at a concise description of the role of prognostic factors.

Predictive accuracy for individual outcomes should be distinguished from the accuracy of point estimates. These can be very precise in large samples but still may differ considerably from individual outcomes.

Let us denote an absolute measure of predictive accuracy without the use of covariates by D and with covariates by D_x . An improvement in predictive accuracy due to covariates x can be quantified by the difference $D - D_x$, the ratio D_x/D or the relative gains in predictive accuracy, $V = (D - D_x)/D$, also termed the *proportion of variation explained by covariates* or, shorter, *explained variation*.

An early and well known V -measure is the multiple R^2 of the general linear model (GLM), which can be defined by taking $D = n^{-1} \sum (y_i - \bar{y})^2$ and $D_x = n^{-1} \sum (y_i - \hat{y}_i)^2$ with y_i ($1 \leq i \leq n$) denoting the values of a continuous outcome variable and \bar{y} and \hat{y}_i unconditional and conditional expectations of y , respectively.

In the last decade, V -measures have been developed for logistic and Cox regression, often with the intent to provide a measure closely related to the R^2 of GLM, and leading to values of comparable magnitude if two models of different types could be applied to the same data set. Such measures have been reviewed and compared by Mittlböck and Schemper [1] and by Schemper and Stare [2] for logistic and Cox regression, respectively. The interested reader is also referred to references [3–13].

Two properties of the R^2 measure have been criticized. The first is that it depends on the range and distribution of covariates, and the second is that with binary responses it tends to be low even for an underlying perfect regression relationship [14]. As has been pointed out [5], both criticisms apply only to R^2 as a measure of goodness-of-fit [15], not of explained variation.

In the first case a covariate with a 'large' estimated regression coefficient is of little predictive value if it has almost no variability in the population from which the sample was drawn. Therefore, differences between values of R^2 among populations with the same assumed regression relationship but different spread of covariates have a natural interpretation, especially if the components of R^2 , D and D_x are taken into account.

In response to the second criticism, assume binomial data 40/100, 50/100 and 60/100 corresponding to values 0, 1 and 2, respectively, of a dose. They are fitted perfectly by logistic regression. However, viewed on an individual-by-individual basis the observations are all zeros and ones while the predictions are either 0.4, 0.5 or 0.6. Knowledge of the particular dose reduces little of the uncertainty of an individual result. While the goodness-of-fit is perfect for this model, the explained variation is low.

If R^2 is used as a measure of explained variation, however, there is a shortcoming usually overlooked: the interpretability of a relative measure such as explained variation requires knowledge of its components. Unfortunately, the components of R^2 are variances and therefore not the most intuitive measures of predictive accuracy. A more natural way to express predictive accuracy is by means of expected absolute departures of observed from predicted outcome values, that is, by the *absolute prediction error*, measured on a scale familiar to the investigator.

Alternatively, predictive accuracy could be expressed in terms of unconditional and conditional standard deviations of outcome, y , which also is on the original scale of y . Under strictly normal y , the standard deviation has an intuitive interpretation (for example, approximately two-thirds of the observed outcomes are within ± 1 standard deviation of the expected outcome value) and also is a more efficient estimate than the absolute prediction error. However, both advantages are lost already under very mild departures from normality (see reference [16] and references therein). Therefore, the absolute prediction error is preferred by this author for a unified concept of predictive accuracy, applicable to various outcome distributions.

The next section describes the estimation of the absolute prediction error and of the alternative standard deviations with continuous, binary, polytomous and survival outcomes. Section 3 illustrates application in medical studies and is followed by the concluding remarks of Section 4.

2. MEASURES OF PREDICTIVE ACCURACY

The absolute prediction error can be obtained by explicit comparison of observed and expected outcomes, leading to *direct* formulations of measures of predictive accuracy. In this case the measures are correctly estimated even if statistical assumptions for a particular regression model are violated or if the model is misspecified.

The absolute prediction error can also be derived from the parameter estimates of a particular regression model. After the data have been used to determine the model, they will no longer be needed to estimate predictive accuracy. The resulting *indirect* or *model-based* formulation of predictive accuracy permits additional insight, mathematically, but may be non-robust to misspecification and violation of assumptions of a model. Both formulations of predictive accuracy lead to identical results under modelling assumptions.

We now present *unconditional* and *conditional absolute prediction errors*, D_T and $D_{T,x}$, respectively, for various types of outcomes, T , using *direct* (\hat{D}_T and $\hat{D}_{T,x}$) and *indirect* (\tilde{D}_T and $\tilde{D}_{T,x}$) estimation, and corresponding estimates of explained variation, \hat{V}_T and \tilde{V}_T .

2.1. Predictive accuracy for models with continuous outcomes

With continuous outcomes y_i ($1 \leq i \leq n$) let \bar{y} and \hat{y}_i denote unconditional and conditional expectations of y , respectively. Then *direct* estimates of predictive accuracy can be defined by $\hat{D}_C = n^{-1} \sum |y_i - \bar{y}|$ and $\hat{D}_{C,x} = n^{-1} \sum |y_i - \hat{y}_i|$ leading to an estimate of relative gains in predictive accuracy or of explained variation by covariates x : $\hat{V}_C = (\hat{D}_C - \hat{D}_{C,x})/\hat{D}_C$.

For models with i.i.d. normal errors, D_C and $D_{C,x}$ can be estimated alternatively by $\tilde{D}_C = \hat{s}_y \sqrt{2/\pi}$ ($\approx 0.8\hat{s}_y$) and $\tilde{D}_{C,x} = \hat{s}_{y|x} \sqrt{2/\pi}$ ($\approx 0.8\hat{s}_{y|x}$) with \hat{s}_y and $\hat{s}_{y|x}$ denoting estimates of unconditional and conditional standard deviations of outcome y . Notice that these *indirect* estimates use $\sqrt{2/\pi}$, which is the expected absolute value of a standard normal deviate. For non-normal errors the relationship between average absolute error and standard deviation will be different and then the *indirect, model-based* estimates do not apply. The *indirect* estimate of explained variation assuming i.i.d. normal errors, $\tilde{V}_C = (\hat{s}_y - \hat{s}_{y|x})/\hat{s}_y$, always leads to smaller values than $R_{\text{adj}}^2 = (\hat{s}_y^2 - \hat{s}_{y|x}^2)/\hat{s}_y^2$. Notice also that quantifying predictive accuracy by unconditional and conditional standard deviations of outcome y instead of by average absolute prediction errors for \tilde{D}_C and $\tilde{D}_{C,x}$ differs only by a multiplicative constant (≈ 0.8) and for \tilde{V}_C not at all.

If predictive accuracy is to be estimated from small samples and/or for regression models with a relatively large number of fitted parameters, k , then no further adjustment is needed for \tilde{D}_C , $\tilde{D}_{C,x}$, \tilde{V}_C or R_{adj}^2 since they are based on appropriate estimates of residual variance. In such situations the *direct* estimates, \hat{D}_C and $\hat{D}_{C,x}$, will underestimate the true prediction error and no distribution-independent correction factor is available. However, under approximate normality of the outcomes \hat{D}_C and $\hat{D}_{C,x}$ should be replaced by $\hat{D}'_C = \hat{D}_C \sqrt{\{n/(n-1)\}}$ and $\hat{D}'_{C,x} = \hat{D}_{C,x} \sqrt{\{n/(n-k)\}}$, respectively. The correction factors reflect the adjustment in the degrees of freedom in \hat{s}_y and $\hat{s}_{y|x}$ and are based on the relationship of prediction error and standard deviation given above.

2.2. Predictive accuracy for models with binary outcomes

With binary outcomes $y_i \in \{0, 1\}$, $1 \leq i \leq n$, unconditional and conditional expectations are denoted by \bar{p} and \hat{p}_i , respectively, where \hat{p}_i can be obtained, for example, from logistic or probit regression. Then *direct* estimates of predictive accuracy can be defined by $\tilde{D}_B = n^{-1} \sum |y_i - \bar{p}| = 2n^{-1} \sum (y_i - \bar{p})^2$ and $\tilde{D}_{B,x} = n^{-1} \sum |y_i - \hat{p}_i|$. Under properly specified models $\tilde{D}_{B,x}$ can also be estimated by $2n^{-1} \sum (y_i - \hat{p}_i)^2$.

The remarkable relationship of absolute and squared error with binary data can be verified by setting $y_i = 1$ and $y_i = 0$, $n\bar{p}$ and $n(1 - \bar{p})$ times, respectively. Because of this relationship $\tilde{V}_B = (\tilde{D}_B - \tilde{D}_{B,x})/\tilde{D}_B$ is identical with the R^2 measure based on squared 'raw' residuals, recommended by various authors [1, 10–12].

Indirect estimates of predictive accuracy which do not require the original outcome values, $\tilde{D}_B = 2\bar{p}(1 - \bar{p})$ and $\tilde{D}_{B,x} = 2n^{-1} \sum \hat{p}_i(1 - \hat{p}_i)$, follow from the squared error definitions of \tilde{D}_B and $\tilde{D}_{B,x}$ by simple arithmetic under the assumption of a perfectly fitting model. Though models never are fitted 'perfectly', in practice, for properly specified models, differences between *direct* and *indirect* estimates usually are very small. Finally, $\tilde{V}_B = (\tilde{D}_B - \tilde{D}_{B,x})/\tilde{D}_B$.

With binary outcomes, in contrast to continuous ones, absolute prediction error (\tilde{D}_B and $\tilde{D}_{B,x}$) and standard deviation ($\tilde{D}'_B = \sqrt{(\tilde{D}_B/2)}$ and $\tilde{D}'_{B,x} = \sqrt{(\tilde{D}_{B,x}/2)}$) are not linearly related. They both assume a maximum value of 0.5 for $\bar{p} = 0.5$ but differ otherwise, for example, for $\bar{p} = 0.01$ \tilde{D}_B and \tilde{D}'_B assume values of 0.02 and 0.10, respectively. As standard deviations of skew distributions, such as the binomial with parameter p close to 0 or 1, are less interpretable the absolute prediction error is preferred by this author.

Under a Poisson model for rare single events the *indirect* estimates of predictive accuracy simplify to $\tilde{D}_B = 2\bar{p}$ and $\tilde{D}_{B,x} = 2n^{-1} \sum \hat{p}_i$ which are identical. Therefore, even highly significant prognostic factors cannot improve predictive accuracy for rare events and cannot explain any variation in the outcome on an individual level. However, if Poisson regression is used to analyse data sets on the level of aggregates (for example, counts of bacteria in certain space and/or time units) then *direct* estimates of predictive accuracy for continuous outcomes, similar to those of Section 2.1, could be considered and will permit estimates of explained variation larger than zero.

2.3. Predictive accuracy for models with polytomous outcomes

With polytomous outcomes $y_i \in \{0, 1, \dots, k, \dots, K\}$, $1 \leq i \leq n$, let the estimated unconditional and conditional probabilities for y_i to belong to the k th outcome category be denoted by \bar{p}_k

and \hat{p}_{ik} , respectively, where the \hat{p}_{ik} can be obtained, for example, from polytomous logistic regression. Note that $\sum_k \bar{p}_k = 1$ and $\sum_k \hat{p}_{ik} = 1$.

Then *direct* estimates of predictive accuracy can be defined by $\hat{D}_P = n^{-1} \sum |1 - \bar{p}_{k(y_i)}|$ and $\hat{D}_{P,x} = n^{-1} \sum |1 - \hat{p}_{ik(y_i)}|$ where $k(y_i)$ denotes the outcome category y_i belongs to. These measures can be interpreted as expected departures of the true outcome category from its prediction by the model. They cannot be interpreted as misclassification rates, which would require specification of classification rules based on the predicted probabilities.

For binary outcomes ($K = 1$) \hat{D}_P and $\hat{D}_{P,x}$ are identical with \hat{D}_B and $\hat{D}_{B,x}$ of Section 2.2, respectively. This is immediately evident for a binary outcome of $y_i = 1$. If $y_i = 0$ the absolute term in \hat{D}_B becomes $|0 - \bar{p}_{(y_i=1)}| = |0 - (1 - \bar{p}_{(y_i=0)})| = |1 - \bar{p}_{(y_i=0)}|$ which is the absolute term of \hat{D}_P . The identity of $\hat{D}_{B,x}$ and $\hat{D}_{P,x}$ with binary outcomes can be shown similarly.

Corresponding *indirect* estimates of predictive accuracy are $\tilde{D}_P = \sum_k \bar{p}_k(1 - \bar{p}_k)$, $\tilde{D}_{P,x} = n^{-1} \sum_k \sum_i \hat{p}_{ik}(1 - \hat{p}_{ik})$ and $\tilde{V}_P = (\tilde{D}_P - \tilde{D}_{P,x})/\tilde{D}_P$. It is easily seen that for binary outcomes ($0 \leq k \leq 1$) \tilde{D}_P and $\tilde{D}_{P,x}$ specialize to \tilde{D}_B and $\tilde{D}_{B,x}$, respectively. As with binary outcomes, *indirect* and *direct* estimates will be very similar for well fitting models. Standard deviations corresponding to the *indirect* estimates of absolute prediction errors are $\tilde{D}'_P = \sqrt{(\tilde{D}_P/2)}$ and $\tilde{D}'_{P,x} = \sqrt{(\tilde{D}_{P,x}/2)}$.

For ordinal polytomous outcomes, especially with many response levels, the *direct* estimates for continuous outcomes of Section 2.1 may be preferred in practice.

2.4. Predictive accuracy for models with survival outcomes

While in logistic regression we compare single observed and predicted values for each individual, in Cox regression [17] a complete survival process is observed in time and compared with its prediction by a survival function. At a certain time t the value of this survival process, $S_i(t)$ (1 for alive at t , 0 for dead at or before t and undefined if survival time is censored before t) can be compared with its unconditional and conditional survival probabilities according to the Kaplan–Meier [18] estimator, $\hat{S}(t)$, and Cox regression, $\hat{S}(t|x_i)$, respectively. Thus for a single time t the situation is identical to regression for binary outcomes discussed above and the same measures of predictive accuracy could be used. However, we are usually interested in survival over the full follow-up range $(0, \tau)$ and hence it is natural to average over t ($0 \leq t \leq \tau$).

Let t_i , η_i and x_i denote observation time, censoring indicator (1 for censored, 0 for death) and covariate vector, respectively, for individual i ($1 \leq i \leq n$). Assume there are m distinct survival times in the sample, at times $t_{(j)}$ ($1 \leq j \leq m$), and with d_j deaths at $t_{(j)}$. Then at each distinct death time $t_{(j)}$ *indirect, model-based* estimates of predictive accuracy can be defined by $\tilde{M}(t_{(j)}) = 2\hat{S}(t_{(j)})(1 - \hat{S}(t_{(j)}))$ and $\tilde{M}(t_{(j)}|x) = 2n^{-1} \sum_i \hat{S}(t_{(j)}|x_i)(1 - \hat{S}(t_{(j)}|x_i))$. If predictive accuracy is to be expressed by standard deviations instead of by absolute prediction errors, the square root is applied to both $\tilde{M}(t_{(j)})$ and $\tilde{M}(t_{(j)}|x)$ after elimination of the multiplier ‘2’.

To obtain overall estimators of predictive accuracy with $(\tilde{D}_{S,x})$ and without covariates (\tilde{D}_S) we form weighted averages of the estimates $\tilde{M}(t_{(j)}|x)$ and $\tilde{M}(t_{(j)})$ over death times, with weights designed to compensate the attenuation in observed death due to earlier censorship: $\tilde{D}_S = w^{-1} \sum_j \hat{G}(t_{(j)})^{-1} d_j \tilde{M}(t_{(j)})$ and $\tilde{D}_{S,x} = w^{-1} \sum_j \hat{G}(t_{(j)})^{-1} d_j \tilde{M}(t_{(j)}|x)$ with $w = \sum_j \hat{G}(t_{(j)})^{-1} d_j$ and \hat{G} denoting the Kaplan–Meier estimator of the censoring or ‘potential follow-up’

distribution estimated like $S(t)$ but with the meaning of the censoring indicator η reversed. Relative gains in predictive accuracy can now be defined as $\hat{V}_S = (\hat{D}_S - \hat{D}_{S,x})/\hat{D}_S$.

Corresponding *direct* estimates of D_S , $D_{S,x}$ and V_S which are based on comparisons of observed survival processes (whenever available) and fitted survival functions have been presented by Schemper and Henderson [19]. Alternative *direct* and *indirect* estimates of predictive accuracy have been proposed by Graf *et al.* [20] and by Korn and Simon [4], respectively.

Though focus of the current section is on semi-parametric analysis of survival data, the presented measures are easily applied also to parametric survival models. With samples of only uncensored survival times, the *direct* measures for continuous data of Section 2.1 could also be considered, which results in a different notion of prediction for survival outcomes.

3. EXAMPLES

Three examples from multiple linear, logistic and Cox regressions will permit numerical comparisons of predictive accuracy and explained variation based on absolute prediction error, standard deviation and variance. They can also give an impression of the usefulness of quantifying predictive accuracy and explained variation in practice. All data sets used are available at <http://www-unix.oit.umass.edu/~statdata/>. Statistical calculations of the suggested measures utilized programs in SAS and S-plus which are available at http://www.akh-wien.ac.at/imc/biometrie/index_en.htm.

3.1. The extent to which birth weight is determined by mother's characteristics

Data set As part of a larger study at the Bayside Medical Center in Springfield, Massachusetts, the weight (in grams) of 189 newborns and various characteristics of their mothers have been obtained. The data set has been analysed by Hosmer and Lemeshow [21]. The following prognostic factors to explain birth weight have been considered: age of mother (in years, AGE); weight of mother at last menstrual period (in pounds, LWT); race (White, Black, other, RACE1, RACE2); smoking during pregnancy (no, yes, SMOKE); history of premature labour (no, yes, PTL); history of hypertension (no, yes, HT) and of uterine irritability (no, yes, UI).

Analysis Multiple linear regression of birth weight on the factors considered gives significant parameter estimates $\hat{\beta}$ for UI ($\hat{\beta} = -492$, $p = 0.0004$), RACE1 ($\hat{\beta} = -474$, $p = 0.0017$), SMOKE ($\hat{\beta} = -327$, $p = 0.0024$), RACE2 ($\hat{\beta} = -340$, $p = 0.0032$), HT ($\hat{\beta} = -577$, $p = 0.0044$) and LWT ($\hat{\beta} = 4.1$, $p = 0.017$). Note that birth weight is approximately normally distributed with mean and standard deviation of 2945 and 729, respectively.

Predictive accuracy From Table I we learn that the partly strong and highly significant prognostic factors do not translate into accurate prediction and that these factors can only account for 12 per cent of the variation in birth weight. We recognize that the uncorrected direct estimate \hat{V}_C is slightly larger than the appropriately corrected one (based on \hat{D}'_C and $\hat{D}'_{C,x}$) which is identical to the estimate of explained standard deviation of birth weight. The estimate of explained variance by R^2_{adj} (= 22 per cent) is substantially higher and the corresponding estimates of predictive accuracy are not easily interpretable.

Table I. Predictive accuracy and explained variation for birth weight study.

Variation measure	Predictive accuracy without/with prognostic factors	Explained variation
Absolute prediction error		
<i>Direct</i> estimation, unadjusted	591/509	0.14
<i>Direct</i> estimation, adjusted	593/520	0.12
<i>Indirect</i> estimation	583/516	0.12
Standard deviation	729/644	0.12
Variance	531474/415972	0.22

Table II. Predictive accuracy and explained variation for prostate cancer study.

Variation measure	Predictive accuracy without/with prognostic factors	Explained variation
Absolute prediction error		
<i>Direct</i> estimation	0.48/0.34	0.29
<i>Indirect</i> estimation	0.48/0.34	0.29
Standard deviation	0.49/0.41	0.15
Variance	0.24/0.17	0.29

3.2. The extent to which certain prognostic factors determine whether prostate cancer has penetrated the prostatic capsule

Data set A prostate cancer data set from The Ohio State University Comprehensive Cancer Center has been presented in Hosmer and Lemeshow [21]. Statistical analysis of the 376 completely documented patients should determine whether potential prognostic factors measured at a baseline exam could be used to predict whether the tumour has penetrated the prostatic capsule. The following prognostic factors have been considered: age (in years, AGE); race (White, Black, RACE); results of digital rectal exam (no (= 1), unilobular (= 2) and bilobular (= 3) nodule, DPROS); detection of capsular involvement (no, yes, DCAPS); prostatic specific antigen value (in mg/ml, range 0.3–139.7, PSA); tumour volume from ultrasound (in cm³, range 0–97.6, VOL), and total Gleason score (range 0–9, GLEASON).

Analysis Multiple logistic regression of tumour penetration on the factors considered gives significant estimates of odds ratios γ for GLEASON ($\hat{\gamma} = 2.6$, $p < 0.00001$), DPROS ($\hat{\gamma} = 2.1$, $p = 0.0008$) and PSA ($\hat{\gamma} = 1.03$, $p = 0.008$).

Predictive accuracy Table II shows that there is no difference in the results of direct and indirect estimates of the absolute prediction error and that explained variance and explained absolute prediction error are identical as well. The estimates of predictive accuracy by the variance are not very intuitive. Results by the standard deviation substantially differ from those by the absolute prediction error.

Despite impressive odds ratios and p -values, predictive accuracy as measured by an absolute prediction error of 0.34 is disappointing as is the proportion of variation attributable to the factors considered (= 0.29).

Table III. Predictive accuracy and explained variation for primary biliary cirrhosis study.

Variation measure	Predictive accuracy without/with prognostic factors	Explained variation
Absolute prediction error		
<i>Direct</i> estimation	0.38/0.23	0.40
<i>Indirect</i> estimation	0.38/0.23	0.39
Standard deviation	0.42/0.30	0.29
Variance	0.19/0.12	0.39

3.3. The proportion of variation in survival of primary biliary cirrhosis patients attributable to known risk factors

Data set A randomized trial in primary biliary cirrhosis of the liver was conducted at the Mayo Clinic from 1974 to 1984. For a total of 312 patients survival times and status (60 per cent censored) as well as several prognostic factors have been obtained and analysed by Fleming and Harrington [22]. In our analysis we used the following prognostic factors: age (in years, AGE); presence of oedema (no, yes, EDEMA); albumin (in mg/dl, range 2–4.6, ALBUMIN); log(serum bilirubin) (serum bilirubin in mg/dl, range – 1.2–3.3, LOG-BIL); log(prothrombin time) (prothrombin time in seconds, range 2.2–2.8, LOGPRO). The time range of medical interest adequately covered by the sample is 0–12 years.

Analysis Multiple Cox regression gives significant estimates of hazard ratios γ for LOG-BIL ($\hat{\gamma} = 2.4$, $p < 0.00001$) and ALBUMIN ($\hat{\gamma} = 0.37$, $p < 0.0001$), AGE ($\hat{\gamma} = 1.03$, $p = 0.0001$), LOGPRO ($\hat{\gamma} = 25.6$, $p = 0.001$).

Predictive accuracy Unconditional and conditional values of predictive accuracy on the scale of survival probability are $\hat{D}_S = 0.38$ and $\hat{D}_{S,x} = 0.23$, respectively. Thus knowledge of the strong prognostic factors reduces the absolute errors of predictions of survival probability in the first 12 years after registration to the study by 0.15. Relative gains in predictive accuracy according to \hat{V}_S amount to 40 per cent which for studies of survival is rather high. From Table III we learn that direct and indirect estimates of predictive accuracy and explained variation by the absolute prediction error are almost identical and that explained variation by squared and absolute prediction errors is identical. Results by standard deviations differ substantially.

4. CONCLUDING REMARKS

Often in practice, covariates can explain only a small fraction of the variation among outcome values, regardless of the type of regression model. Therefore it is important for medical investigators to know that even strong and highly significant covariates of a study may not automatically translate into sufficiently accurate prediction or close determination of individual outcome values.

For this purpose the absolute prediction error appears to be more suitable than the squared error or variance because of its interpretability on a natural scale. Standard deviations share

this property but are unsuitable measures of variability for highly skew distributions such as the binomial with parameter values close to zero or one. In the examples for binary and survival outcomes conditional predictive accuracy according to standard deviation appears to be worse than according to absolute prediction error. This is due to the dominating influence of the large squared differences from the extreme tail of the distribution on standard deviation. The absolute prediction error is well interpretable with all types of outcome distributions.

Most often $\hat{D}_{T,x}$ will be compared with \hat{D}_T , but occasionally the question arises whether prediction can be improved if some ‘experimental’ covariates, v , are added to ‘established’ covariates, u . Then construction of the quantities $\hat{D}_{T,uv}/\hat{D}_{T,u}$, $\hat{D}_{T,u} - \hat{D}_{T,uv}$ or $(\hat{D}_{T,u} - \hat{D}_{T,uv})/\hat{D}_{T,u}$ can address the resulting issue of the relative importance of covariates [23, 24].

In this paper we have focused on regression models. Measures of predictive accuracy and explained variation can similarly be applied to evaluate the performance of neural networks, of expert judgement or of clinical classification schemes.

ACKNOWLEDGEMENTS

The author wishes to thank Terry Smith, Houston, and an anonymous referee for helpful comments on an earlier draft of the paper.

REFERENCES

- Mittlböck M, Schemper M. Explained variation for logistic regression. *Statistics in Medicine* 1996; **15**(19): 1987–1997.
- Schemper M, Stare J. Explained variation in survival analysis. *Statistics in Medicine* 1996; **15**(19):1999–2012.
- Kent JT, O’Quigley J. Measures of dependence for censored survival data. *Biometrika* 1988; **75**(3):525–534.
- Korn LK, Simon R. Measures of explained variation for survival data. *Statistics in Medicine* 1990; **9**(5): 487–503.
- Korn LK, Simon R. Explained residual variation, explained risk and goodness of fit. *American Statistician* 1991; **45**(3):201–206.
- Schemper M. The explained variation in proportional hazards regression. *Biometrika* 1990; **77**(1):216–218 (correction: *Biometrika* 1994; **81**(3):631).
- Schemper M. Further results on the explained variation in proportional hazards regression. *Biometrika* 1992; **79**(1):202–204.
- Graf E, Schumacher M. An investigation on measures of explained variation in survival analysis. *Statistician* 1995; **44**:497–507.
- Henderson R. Problems and prediction in survival-data analysis. *Statistics in Medicine* 1995; **14**(2):161–184.
- Ash A, Shwartz M. R^2 : A useful measure of model performance when predicting a dichotomous outcome. *Statistics in Medicine* 1999; **18**(4):375–384.
- Buyse M. Letter to the Editor. R^2 : A useful measure of model performance when predicting a dichotomous outcome. *Statistics in Medicine* 2000; **19**(2):271–274.
- Zheng B, Agresti A. Summarizing the predictive power of a generalized linear model. *Statistics in Medicine* 2000; **19**(13):1771–1781.
- Menard S. Coefficients of determination for multiple logistic regression analysis. *American Statistician* 2000; **54**(1):17–24.
- Cox DR, Wermuth N. A comment on the coefficient of determination for binary responses. *American Statistician* 1992; **46**(1):1–4.
- Kvalseth TO. Cautionary note about R^2 . *American Statistician* 1985; **39**(4):279–285.
- Hampel F. Mean deviation. In *Encyclopedia of Biostatistics, volume 3*, Armitage P, Colton T (eds). Wiley: New York, 1998; 2488–2489.
- Cox DR. Regression models and life tables. *Journal of the Royal Statistical Society, Series B* 1972; **34**: 187–220.
- Kaplan EL, Meier PM. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association* 1958; **53**:457–481.
- Schemper M, Henderson R. Predictive accuracy and explained variation in Cox regression. *Biometrics* 2000; **56**(1):249–255.

20. Graf E, Schmoor C, Sauerbrei W, Schumacher M. Assessment and comparison of prognostic classification schemes for survival data. *Statistics in Medicine* 1999; **18**(17/18):2529–2545.
21. Hosmer DW, Lemeshow S. *Applied Logistic Regression*, 2nd edn. Wiley: New York, 2000.
22. Fleming RT, Harrington DP. *Counting Processes and Survival Analysis*. Wiley: New York, 1991.
23. Healy MJR. Measuring importance. *Statistics in Medicine* 1990; **9**(6):633–637.
24. Schemper M. The relative importance of prognostic factors in studies of survival. *Statistics in Medicine* 1993; **12**(24):2377–2382.