

The Binary Regression Quantile Plot: Assessing the Importance of Predictors in Binary Regression Visually

EFSTATHIA BURA

Department of Statistics
The George Washington University
USA

JOSEPH L. GASTWIRTH

Division of Cancer Epidemiology and Genetics
National Cancer Institute
USA

Summary

We present a graphical measure of assessing the explanatory power of regression models with a binary response. The binary regression quantile plot and an area defined by it are used for the visual comparison and ordering of nested binary response regression models. The plot shows how well various models explain the data. Two data sets are analyzed and the area representing the fit of a model is shown to agree with the usual likelihood ratio test.

Key words: Binary response; Logistic regression; Explanatory power; Quantile plot.

1. Introduction

Prediction models estimating the probability an individual belongs to a particular category, e.g. disease-free, are widely used in applied statistics. Numerous goodness of fit statistics for the fitted model have been proposed. As some try to mimic aspects of R^2 in linear regression, they can also be regarded as measures of the predictive power of the model. Several texts and articles present the most commonly used ones and comment on their merits and shortcomings (AGRESTI, 1990, 1996; CHRISTENSEN, 1997; COLLETT, 1991; HOSMER and LEMESHOW, 1989; MCCULLAGH and NELDER, 1989; PIGEON and HEYSE, 1999).

This article introduces a new estimation method for the explanatory or predictive power of a fitted model which is based on a graph that quantifies the fitted model's predictive power. This should facilitate the communication of analytical findings to subject matter specialists. The existing measures are mainly numerical.

The proposed measure stems from the desire to visually judge and assess how well a fitted model explains the data by investigating the scatterplot of the binary response versus the regressor with the fitted response overlaid. Because predictor variables may have an arbitrarily large range, we introduce a transformed graph that plots the predicted probabilities against the percentiles of the predictor, which may be a linear combination of several independent variables. Two regions of the graph are identified with the gain in accuracy of prediction yielded by the regression curve relative to the sample proportion of successes (1's). The area of these two regions is the numerical value of our measure.

The proposed measure quantifies how much more explanatory power the fitted model has compared to the mean of the binary response variable. Based on simulations, there appears to be a monotonic relationship between the range of values of the measure and the degree to which the fitted model explains the data. In addition, the new measure is shown to be asymptotically normally distributed. Two examples are presented to illustrate its use.

2. A New Graphical Measure

Let Y denote a binary response variable taking on only two values, which we will denote by 0 and 1. Hence, Y has the Bernoulli distribution with $\Pr(Y = 1) = p$, $\Pr(Y = 0) = 1 - p$, and $E(Y) = p$. In the presence of a regressor variable X , which may affect the behaviour of the distribution of Y , $E(Y | X = x) = \Pr(Y = 1 | X = x) = p(x)$; that is, $p(x)$ is the probability that $Y = 1$ when $X = x$. It is usually assumed that

$$E(Y | X = x) = p(x) = G(\alpha + \beta x),$$

where G is a continuous cumulative distribution function, twice continuously differentiable, with likelihood function concave in β . In practice, G is often assumed to be either logistic or normal.

Given a random sample from (y, x) , the parameters are usually estimated by maximizing the likelihood function. The resulting maximum likelihood estimates $\hat{\alpha}$, $\hat{\beta}$ are used to compute the estimate of the risk $\hat{p}(x) = \hat{\Pr}(Y = 1 | X = x) = G(\hat{\alpha} + \hat{\beta}x)$.

A measure of the explanatory power of a regression with a binary response should assess how much better the model approximates $\Pr(Y = 1 | X)$ relative to p , the mean of Y . A graphical means of quantifying this difference is by using the area between the horizontal line $y = p$ and the curve $p(x)$. The larger the integral of this region with respect to the c.d.f. of X is, the greater the information about Y provided by X appears to be.

The proposed measure, namely *Total Gain* of the model $p(x)$ over p is defined by

$$TG = \begin{cases} \int_{-\infty}^{x^*} (p - p(x)) dF(x) + \int_{x^*}^{\infty} (p(x) - p) dF(x) & \text{for } \beta > 0 \\ \int_{-\infty}^{x^*} (p(x) - p) dF(x) + \int_{x^*}^{\infty} (p - p(x)) dF(x) & \text{for } \beta < 0 \end{cases}, \quad (1)$$

where F is the unknown cumulative distribution function of X , and x^* is the abscissa of the intersection of the two lines, $y = p$ and $y = p(x)$.

Notice that

$$p = E(p(x)) = \int_{-\infty}^{\infty} p dF(x) = \int_{-\infty}^{x^*} p(x) dF(x) + \int_{x^*}^{\infty} p(x) dF(x).$$

Therefore,

$$\int_{-\infty}^{x^*} (p - p(x)) dF(x) = \int_{x^*}^{\infty} (p(x) - p) dF(x). \quad (2)$$

This implies that

$$TG = \begin{cases} 2 \int_{-\infty}^{x^*} (p - p(x)) dF(x) & \text{for } \beta > 0 \\ 2 \int_{x^*}^{\infty} (p(x) - p) dF(x) & \text{for } \beta < 0 \end{cases}$$

or, generally

$$TG = 2 \left| \int_{-\infty}^{x^*} (p - p(x)) dF(x) \right|. \quad (3)$$

It is shown in Section 2.4 that $TG \leq 2p(1 - p)$, so we define the standardized total gain, TG_{std} , as follows:

$$TG_{std} = \frac{TG}{2p(1 - p)}. \quad (4)$$

2.1 The Binary Regression Quantile Plot

TG is based on the area between the logistic curve and the $p(x) = p$ line over the entire x -range of values. An alternative plot can be obtained by expressing the

relationship between y and x in terms of the percentiles of x . For each percentile t , the corresponding value of x , x_t is computed. Then, $\hat{E}(Y | X = x_t)$ is plotted against the corresponding x percentiles, where $\hat{E}(Y | X = x_t)$ is the estimate from the fitted model.

Our binary regression quantile plot is the graph corresponding to the quantile regression function defined in RAO and ZHAO (1995), who subsequently estimated it non-parametrically in RAO and ZHAO (1996). We focus on a parametric link and use ML estimates of the parameters. Alternatively, modified ML parameter estimates can be used (TIKU and VAUGHAN, 1997). The concept of the gain in predictive accuracy is also used in the analysis of expectancy curves (CSÖRGO, GASTWIRTH and ZITIKIS, to appear), and in measures of association based on proportional error reduction (GOODMAN and KRUSKAL, 1979; REYNOLDS, 1977).

The binary regression quantile plot assesses graphically the contribution of one or more covariates to the explanation of a binary variable. It differs from the *receiver-operating characteristic (ROC) curve* and the *logit rank plot* of COPAS (1999) as it does not classify subjects into high risk or low risk cases.

To obtain TG , let $t = F(x)$. Accordingly, $t^* = F(x^*)$, where $p(x^*) = p = p(F^{-1}(t^*))$. As $dt = dF(x)$,

$$\int_0^{t^*} (p - p(F^{-1}(t))) dt = \int_{-\infty}^{x^*} (p - p(x)) dF(x)$$

and,

$$\int_{t^*}^1 (p(F^{-1}(t)) - p) dt = \int_{x^*}^{\infty} (p(x) - p) dF(x)$$

yielding,

$$TG = 2 \int_0^{t^*} (p - p(F^{-1}(t))) dt. \quad (5)$$

The advantage of the new plot and (5) over (3) is that the expected responses are plotted over the unit interval as opposed to the entire x -range. The comparison of different models is facilitated by the fact that the graphs are on the unit square and the *total gain* of the fitted model over the intercept model can be visually assessed.

2.2 Simulation study

The basic concepts and properties of the *Total Gain* measure and binary regression quantile plot will be illustrated with a simulated data set. 50 values of X are generated from the $U(0, 20)$ distribution. The binary responses Y_i are simulated as

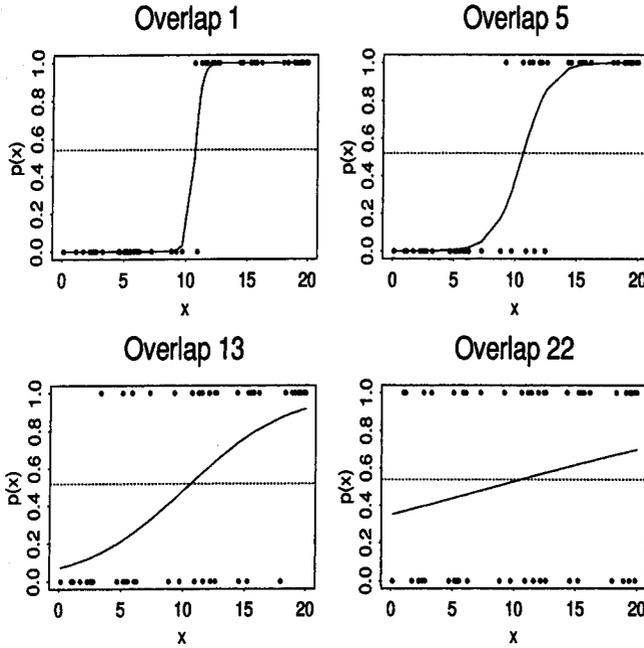


Fig. 1. Plots of the simulated data with various degrees of overlap and the corresponding fitted logistic curves. The dashed horizontal line is \bar{p}

follows: All Y values corresponding to X values smaller than 10 are set equal to zero and the remaining are set equal to one. The proportion of one's is $28/50$. In this data set there is complete separation of the Y values along the x -axis and thus

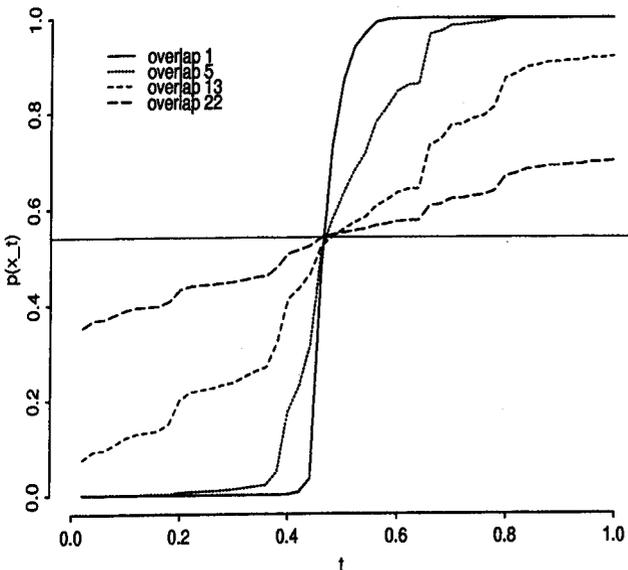


Fig. 2. Binary regression quantile plots for the simulated data with various degrees of overlap. The legend indicates the number of overlapping points. The dashed horizontal line is \bar{p}

the logistic model cannot be fitted to these data. Therefore, the following four cases will be considered: In *Overlap 1* one Y value is switched to create a minimum overlap, in *Overlap 5* five Y values are switched, in *Overlap 13* thirteen Y values are switched, and in *Overlap 22* twenty two Y values are switched. The situation is illustrated in Figure 1 where the fitted logistic curves are also plotted. Obviously, the fit of the logistic model deteriorates as the overlap increases. The lower right plot exhibits the case where the overlap is almost random across x and the fitted logistic curve is collapsing to the \bar{p} dashed horizontal line.

Figure 2 contains the corresponding *Total Gain* binary regression quantile plots. It is clear that they are ordered with the best corresponding to minimum overlap and the worst corresponding to almost random overlap. The sample based \widehat{TG} and $\widehat{TG}_{\text{std}}$ scores are as follows: 0.476 and 0.959 for *Overlap 1*, 0.423 and 0.847 for *Overlap 5*, 0.263 and 0.527 for *Overlap 13*, and 0.093 and 0.187 for *Overlap 22*. The formulae for computing the sample estimates of TG and TG_{std} can be found in Sections 2.3 and 2.4, respectively.

2.3 Computation of Total Gain (TG)

Suppose a sample of size n from (Y, X) has been collected. We assume that $\beta > 0$ because one can consider either the 0's or 1's as the positive response. The c.d.f., $F(x)$, of X is continuous and will be estimated by the *empirical distribution function* F_n . Since $p(x)$ is a bounded, continuous real function (see BILLINGSLEY, 1986, Theorem 25.8, p. 344),

$$\int p(x) dF_n(x) \rightarrow \int p(x) dF(x).$$

Consequently, from (3), TG can be consistently estimated by

$$\begin{aligned} \widehat{TG} &= 2 \left\{ \int_{-\infty}^{x^*} p dF_n(x) - \int_{-\infty}^{x^*} p(x) dF_n(x) \right\} \\ &= 2 \left\{ pF_n(x^*) - \frac{1}{n} \sum_{x_i \leq x^*} p(x_i) \right\}. \end{aligned} \quad (6)$$

Let

$$\bar{p} = \frac{\# \text{ of events}}{\# \text{ of trials}} = \frac{\sum_{i=1}^n y_i}{n}. \quad (7)$$

Also, let $\hat{\alpha}$ and $\hat{\beta}$ denote the maximum likelihood estimates of the parameters α and β , and let $\hat{p}(x) = G(\hat{\alpha} + \hat{\beta}x)$. By plugging in \bar{p} , $\hat{\alpha}$, and $\hat{\beta}$ in (1), the *Total Gain* for the fitted model $\hat{p}(x)$ is given by

$$2 \left\{ \bar{p} \frac{1}{n} \sum_i I(x_i \leq x^*) - \frac{1}{n} \sum_{x_i \leq x^*} \hat{p}(x_i) \right\}.$$

The point x^* is easily computed from the equation $p = p(x)$ where the unknown parameters are replaced by estimates for the specific model of $p(x)$. For example, if G is the logistic c.d.f., then

$$p(x) = \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)} \tag{8}$$

When $\beta = 0$, $p(x) = p = \Pr(Y = 1)$, regardless of the value of X . Setting $p = p(x)$ yields $x^* = [\log(p/(1 - p)) - \alpha]/\beta$. Therefore, x^* can be estimated by

$$\hat{x}^* = \left(\log \frac{\bar{p}}{1 - \bar{p}} - \hat{\alpha} \right) / \hat{\beta} \tag{9}$$

which in turn yields the following formula for computing TG ,

$$\widehat{TG} = 2 \left\{ \bar{p} \frac{1}{n} \sum_{i=1}^n I(x_i \leq \hat{x}^*) - \frac{1}{n} \sum_{x_i \leq \hat{x}^*} \hat{p}(x_i) \right\}. \tag{10}$$

Furthermore, in the logistic regression framework, \hat{x}^* has an asymptotic normal distribution as stated in the following lemma. The proof is given in the Appendix.

Lemma 1: Suppose $(y_1, x_1), (y_2, x_2), \dots, (y_n, x_n)$ is a random sample from a binary random variable Y and a continuous random variable X . Also, suppose that the logistic regression model (8) is fitted to the data.

Let $\hat{\alpha}$ and $\hat{\beta}$ be the maximum likelihood estimates of α and β , respectively. Let \hat{x}^* be defined by (9). Then, conditionally on x_1, x_2, \dots, x_n ,

$$\sqrt{n} (\hat{x}^* - x^*) \xrightarrow[n \rightarrow \infty]{} N(0, \sigma_*^2), \tag{11}$$

where

$$\begin{aligned} \sigma_*^2 = & \frac{1}{\beta^2} \frac{1}{\left(\sum p_i q_i \right) \left(\sum x_i^2 p_i q_i \right) - \left(\sum x_i p_i q_i \right)^2} \\ & \times \left\{ \left(\frac{n \sum p_i q_i}{\sum p_i (n - \sum p_i)} - 1 \right)^2 \sum x_i^2 p_i q_i + \left(\frac{n \sum x_i q_i}{\sum p_i (n - \sum p_i)} + \frac{\alpha}{\beta} \right)^2 \sum p_i q_i \right. \\ & \left. - 2 \left(\frac{n \sum p_i q_i}{\sum p_i (n - \sum p_i)} - 1 \right) \left(\frac{n \sum x_i q_i}{\sum p_i (n - \sum p_i)} - \frac{\alpha}{\beta} \right) \sum x_i p_i q_i \right\} \end{aligned}$$

with $p_i = p(x_i)$, $q_i = 1 - p_i$.

2.4 Properties of TG

TG is a non-negative, unitless measure of the total cumulative distance between the mean, p , of Y and the response curve $p(x)$. When, Y and X are independent,

as exhibited in Figure 1 (iii), $p(x)$ is constant and therefore it coincides with p . In this case, TG is readily seen to be zero from (7). When there is clear separation of the two Y values along the x -axis; that is, when there is a straightforward relationship between X and $\Pr(Y = 1|X)$, we are led to set

$$p(x) = \begin{cases} 0 & \text{for } x \leq x^* \\ 1 & \text{for } x > x^* \end{cases}$$

if the values are distributed as in Figure 1 (i). Otherwise, $p(x)$ is defined analogously. The step function $p(x)$ fits the Y values perfectly and yields the maximum value of TG , which by (3) is:

$$TG = pF(x^*) + (1 - p)(1 - F(x^*)).$$

Recall that $pF(x^*) = (1 - p)(1 - F(x^*))$ from (5), yielding $F(x^*) = 1 - p$. Consequently the maximum value of TG is

$$TG_{\max} = 2p(1 - p). \quad (12)$$

Clearly, (26) is maximized at $p = 1/2$, so that $0 \leq TG \leq 2p(1 - p) \leq 1/2$, with the two endpoints corresponding to total independence (0) of Y and X and to perfect relationship ($\frac{1}{2}$) between Y and X , respectively. Values of TG close to 0 signify poor explanatory power, and TG values close to the upper bound signify better explanatory power. In practice, one estimates p by \bar{p} , so \widehat{TG} -values close to $2\bar{p}(1 - \bar{p})$ would indicate a near perfect explanatory power of the fitted model.

The sample based estimate of TG_{std} defined by (8) is given by

$$\widehat{TG}_{\text{std}} = \frac{\widehat{TG}}{2\bar{p}(1 - \bar{p})}. \quad (13)$$

This estimated standardized version of TG can be viewed as an analog to the coefficient of determination, R^2 , in linear regression when the latter is used as a measure of predictive power or "explained risk" (KORN and SIMON, 1991), and not as a measure of goodness of fit, for the following reasons (see MITTLBÖCK and SCHEMPER, 1996): (i) it has an intuitively clear interpretation; (ii) linear transformations of the explanatory variables do not affect $\widehat{TG}_{\text{std}}$, non-linear monotonic transformations do; (iii) $0 \leq \widehat{TG}_{\text{std}} \leq 1$ from (13), with the two endpoints corresponding to complete lack of predictability and perfect predictability, respectively; and (iv) even though the values of R^2 and TG_{std} are not identical for data sets that can be analyzed by a linear model, they both follow the same trend.

In Section 2.3, \widehat{TG} was shown to be a consistent estimate of TG . The large sample distribution of \widehat{TG} , used to compute approximate confidence intervals, is given by the following:

Theorem 1: Let \widehat{TG} be defined by (10). When the density function $f(x)$ of $F(x)$ is strictly positive at x^* and $F'(x)$ is bounded in a neighborhood of x^* , then

$$\sqrt{n}(\widehat{TG} - TG) \xrightarrow[n \rightarrow \infty]{} N(0, \sigma_{TG}^2), \quad (14)$$

where

$$\sigma_{TG}^2 = 4 \left\{ F_*^2 p(1-p) + p^2 F_*(1-F_*) + F_*^2 p_c(1-p) + \frac{\int_{-\infty}^{x^*} p^2(x) dF(x)}{F_*} - \frac{\left(\int_{-\infty}^{x^*} p(x) dF(x) \right)^2}{F_*^2} + 2p(1-F_*) \int_{-\infty}^{x^*} p(x) dF(x) \right\}$$

with $F_* = \int_{-\infty}^{x^*} dF(x)$ and $p_c = E(Y | X \leq x^*) = \int_{-\infty}^{x^*} p(x) dF(x) / F_*$.

The proof is given in the Appendix.

2.5 The binary regression quantile plot for the multiple regressor case

So far we have restricted our attention to binary regressions with one independent variable X . For multiple regressors one can consider the fitted “best” predictor of $E(Y | X)$, say $\beta^T X$, as a new single regressor. Consequently, the plot can be drawn as in the univariate case, with the percentile scores of $\beta^T X$ placed on the x -axis.

In practice, for the computation of the binary regression quantile plot one fits the logistic model

$$p(x) = \frac{\exp(\alpha + \beta^T x)}{1 + \exp(\alpha + \beta^T x)} \tag{15}$$

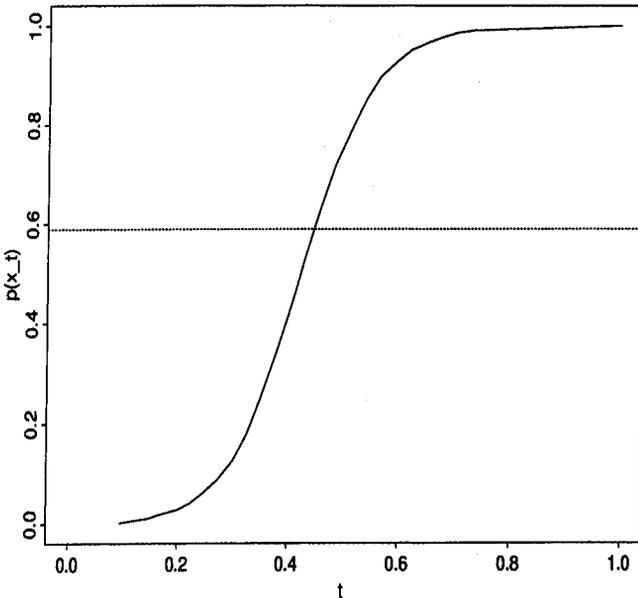


Fig. 3. Binary regression quantile plot of the logistic regression of *menarche* on *age*. The dashed horizontal line is \bar{p}

to the data and then uses $\hat{\beta}^T \mathbf{x}$ as a single predictor variable. The method applies to other link functions as well.

3. Examples

The first data set to be analysed is the *Age of menarche of 3918 Polish girls* data (MILICER and SZCZOTKA, 1966). These data were analysed by FINNEY (1971), MORGAN (1992) and SPRENT (1998). The *age* of the girls at the onset of menstruation is the only regressor variable.

Figure 3 contains the binary regression quantile plot exhibiting the *Total Gain* for the logistic regression of *menarche* on *age*. As expected, it is evident that *age* is a significant variable for modelling the probability of having reached menarche. The numerical value of TG is 0.397, and the estimate of its maximum is $2\bar{p}(1 - \bar{p}) = 0.484$. Therefore, $TG_{\text{std}} = 0.82$.

The *Kyphosis in laminectomy patients* data set, taken from HASTIE and TIBSHIRANI (1990, pp. 301–303), are retrospective measurements on 81 patients on four variables. The variable of interest is the presence (1) or absence (0) of *kyphosis*, a spinal deformity. The predictors are *age* in months at time of operation, the starting and ending range of vertebrae levels involved in the operation, indicated by *start* and *end*, and the *number* of levels involved. The regressors *start* and *number* satisfy $\text{number} = \text{end} - \text{start} + 1$. The analysis concentrates on identifying risk factors for the binary response variable *kyphosis*.

Figure 4 contains four plots. In all four graphs, the solid curve corresponds to the binary regression quantile plot from the regression of *kyphosis* on the three

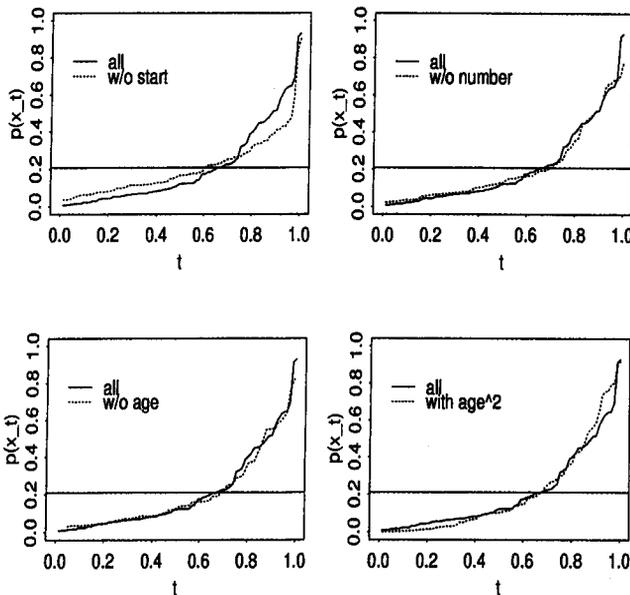


Fig. 4. Binary regression quantile plots of the multiple logistic regression of *kyphosis* on all regressors (*start*, *number*, *age*) with binary regression quantile plots for logistic regressions of *kyphosis* on all but one regressor variables, as indicated in the legends. The horizontal line is \bar{p}

regressors, *start*, *number*, and *age*. The dashed lines correspond to the regressions of *kyphosis* on all but one of the regressor variables, except for the lower right plot in which the dashed line corresponds to the percentile curve for the model with age^2 in addition to the other three regressors, *start*, *number* and *age*. For example, in the upper left plot of the figure, the dashed line represents the percentile curve of the logistic regression of *kyphosis* on *number* and *age*. The percentile scores of $\hat{\beta}^T x$ are placed on the horizontal axis, where $\hat{\beta}$ is obtained as explained in Section 3.1.5. That is, for the model with $x = (start, number, age)^T$, $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3)^T = (-0.207, 0.011, 0.411)^T$. The values are the maximum likelihood estimates of the parameters of the logistic regression model. The *S - PLUS* statistical package was used for all computations.

The plots indicate that the factor affecting the fit of the model the most is *start*, as its exclusion from the full model reduces the area between the \bar{p} line and the percentile curve noticeably. The variables *age* and *number* seem to be almost indistinguishable with respect to their importance in the logistic model. This result is in accordance with the univariate binary regression quantile plots. From Figure 4 and Table 1 it is clear that *start* is the regressor variable with the largest explanatory power. The second in importance appears to be *number*. Of course, part of its effect cannot be distinguished from *start* as they are linearly related. The variable *age*, on the other hand, is not judged to be important when it enters the logistic model linearly. Nevertheless, it seems to have a quadratic effect on the logit of *kyphosis*, as exhibited in the lower right plot in Figure 4.

The above assessments for the modelling worth of the individual regressors drawn from the binary regression quantile plots of Figure 31 are in agreement with the standard deviance analysis (COLLETT, 1991, Sec. 3.9) of the corresponding models. Table 1 contains a list of all the fitted models, the corresponding estimated *Total Gain* (\widehat{TG}) scores, as well as the standardized *Total Gain*, $\widehat{TG}_{std} = \widehat{TG}/2\bar{p}(1 - \bar{p})$, the residual deviance and the residual degrees of freedom.

4. Discussion

This paper introduces the *Total Gain* (*TG*) measure to assess the explanatory power of a binary regression model. The *Total Gain* lies between 0 and 0.5 and its standardized form is between 0 and 1; the standardized version can be considered as an analog of R^2 in linear regression. In effect, in the simple linear regression context, \widehat{TG} is a multiple of the slope estimate.

In conjunction with the binary regression quantile plot, *TG* provides a visual tool to judge the explanatory power of a nested logistic or other link regression models. Moreover, the importance of the individual regressors can be graphically assessed and compared to that of the remaining regressors.

Although our examples had moderate values of \bar{p} , the *TG* measure also applies in cases where there is a strong relationship between response and regressors, e.g.

Table 1
Analysis of the kyphosis data.

Model	\widehat{TG}	\widehat{TG}_{std}	Deviance	df
<i>start + number + age + age²</i>	0.194	0.585	54.428	76
<i>start + number + age</i>	0.171	0.514	61.380	77
<i>start + age + age²</i>	0.166	0.511	58.415	77
<i>number + age + age²</i>	0.148	0.456	63.863	79
<i>start + number</i>	0.160	0.483	64.537	78
<i>start + age</i>	0.156	0.471	65.299	78
<i>number + age</i>	0.118	0.357	71.627	78
<i>age + age²</i>	0.119	0.365	72.739	78
<i>start</i>	0.147	0.444	68.072	79
<i>number</i>	0.110	0.332	73.357	79
<i>age</i>	0.045	0.135	81.933	79

the relationship of maternal age to birth defects (HOOK, 1981), which starts at a very low probability (1 in 500) and rises to a small one (1 in 20).

Acknowledgements

The authors would like to thank the referee and the editor for their useful comments and valuable suggestions. J. L. GASTWIRTH was supported in part by National Science Foundation Grant SBR-9807731 awarded to George Washington University.

Appendix

Proof of Lemma 1: For a generalised linear model with canonical link; that is, $\eta = g(\mu) = \mathbf{x}'\boldsymbol{\beta}$ where $\mathbf{x}' = (x_1, x_2, \dots, x_k)$, $\boldsymbol{\beta} = (\alpha, \beta_1, \dots, \beta_k)'$, we have that $\sum_{i=1}^n y_i = \sum_{i=1}^n \hat{\mu}_i$ (see McCULLAGH and NELDER, 1989) or, equivalently, $\bar{p} = \frac{1}{n} \sum_{i=1}^n \hat{p}_i$.

Specifically, in a one variable logistic regression context, \hat{p}_i is given by plugging in x_i , $\hat{\alpha}$ and $\hat{\beta}$ for x , α and β in (8), respectively. By (9), \hat{x}^* can be expressed as a function of $\hat{\alpha}$ and $\hat{\beta}$, via $\hat{p}(x)$, as follows,

$$\hat{x}^* = \{\log(\sum \hat{p}(x_i)/(n - \sum \hat{p}(x_i))) - \hat{\alpha}\}/\hat{\beta}.$$

Now, the estimates $\hat{\alpha}$ and $\hat{\beta}$, have a joint asymptotic normal distribution (eg., see AGRESTI, 1990) with asymptotic covariance matrix $(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}$, where $\mathbf{W} = \text{diag}(p_i(1 - p_i))$ and the design matrix \mathbf{X} contains a column of ones.

From (9) we see that \hat{x}^* is a smooth function of $\hat{\alpha}$, $\hat{\beta}$. By a straightforward application of the delta method we obtain that $\sqrt{n}(\hat{x}^* - x^*) \Rightarrow AN(0, \sigma_*^2)$ where σ_*^2 is given in the statement of Lemma 1.

Proof of Theorem 1: Recall from (10) that \widehat{TG} is twice the difference of the two statistics, $\bar{p}F_n(\hat{x}^*)$ and $\frac{1}{n} \sum_{x_i \leq \hat{x}^*} \hat{p}(x_i)$. By (7) and the central limit theorem we have

$$\bar{p} \doteq p + \sqrt{\frac{p(1-p)}{n}} \epsilon_1 \tag{16}$$

where ϵ_1 is a standard normal random variable. Also, $F_n(\hat{x}^*)$ is the empirical c.d.f. of X at \hat{x}^* . Therefore,

$$F_n(\hat{x}^*) = \frac{1}{n} \sum_{i=1}^n I(x_i \leq \hat{x}^*) = F_n(x^*) + \frac{1}{n} \sum_{i=1}^n D_i I(x_i \in S)$$

where $S = [\min(\hat{x}^*, x^*), \max(\hat{x}^*, x^*)]$, and $D_i = 1$ when $\hat{x}^* < x^*$, $D_i = -1$ when $\hat{x}^* > x^*$. BAHADUR (1966) showed that in an interval $I_n = (x^* - a_n, x^* + a_n)$ about any point x^* , where $a_n = (\log n)/n^{1/2}$, the empirical c.d.f. satisfies

$$F_n(x) = F_n(x^*) + F(x) - F(x^*) + o_p(n^{-1/2}) \tag{17}$$

uniformly in I_n . As $|\hat{x}^* - x^*|$ is $o_p(n^{-1/2})$ with probability arbitrarily close to 1, $\hat{x}^* \in I_n$ for large n . Letting \hat{x}^* be x in (17) and using the first order Taylor expansion for $F(\hat{x}^*) - F(x^*)$ yields

$$F_n(\hat{x}^*) = F_n(x^*) + f(x^*) \Delta + o_p(n^{-1/2}) \tag{18}$$

where $\Delta = \hat{x}^* - x^*$. Similar uses of this representation appear in GASTWIRTH (1971) and GHOSH (1971).

From (16) and (18) it follows that

$$\begin{aligned} \bar{p}F_n(\hat{x}^*) &= pF_n(x^*) + pf(x^*) \Delta + f(x^*) \Delta \sqrt{\frac{p(1-p)}{n}} \epsilon_1 \\ &\quad + F_n(x^*) \sqrt{\frac{p(1-p)}{n}} \epsilon_1 + o_p(n^{-1/2}) \end{aligned} \tag{19}$$

Again from the central limit theorem we have that $F_n(x^*) = F_* + \sqrt{\frac{F_*(1-F_*)}{n}} \epsilon_2$, where $F_* = F(x^*)$ and ϵ_2 is a standard normal random variable.

Substitution in (19) yields

$$\begin{aligned} \bar{p}F_n(\hat{x}^*) &= pF_* + p \sqrt{\frac{F_*(1-F_*)}{n}} \epsilon_2 + pf(x^*) \Delta \\ &\quad + f(x^*) \Delta \sqrt{\frac{p(1-p)}{n}} \epsilon_1 + F_* \sqrt{\frac{p(1-p)}{n}} \epsilon_1 \\ &\quad + \sqrt{\frac{F_*(1-F_*)}{n}} \sqrt{\frac{p(1-p)}{n}} \epsilon_1 \epsilon_2 + o_p(n^{-1/2}) \end{aligned}$$

By Lemma 1, $\sqrt{n}(\hat{x}^* - x^*)$ has an asymptotic normal distribution with mean zero and variance σ_*^2 . Thus,

$$\bar{p}F_n(\hat{x}^*) = pF_* + pf(x^*) \Delta + F_* \sqrt{\frac{p(1-p)}{n}} \epsilon_1 + p \sqrt{\frac{F_*(1-F_*)}{n}} \epsilon_2 + o_p(n^{-1/2}) \quad (20)$$

with

$$\epsilon_1 = \frac{\sum_{i=1}^n (y_i - p)}{\sqrt{np(1-p)}} \sim AN(0, 1), \quad \epsilon_2 = \frac{\sum_{i=1}^n (I_i^* - F_*)}{\sqrt{nF_*(1-F_*)}} \sim AN(0, 1) \quad (21)$$

where $I_i^* = I(x_i \leq x^*)$. The second term of \widehat{TG} is

$$\frac{1}{n} \sum_{x_i \leq \hat{x}^*} \hat{p}(x_i) = \frac{1}{n} \sum_{x_i \leq x^*} \hat{p}(x_i) + \frac{1}{n} \sum_{x_i \in S} \hat{p}(x_i).$$

Recall that $\hat{p}(x_i) = \exp(\hat{\alpha} + \hat{\beta}x_i)/(1 + \exp(\hat{\alpha} + \hat{\beta}x_i))$, and that $(\hat{\alpha}, \hat{\beta})$ have a \sqrt{n} -asymptotic distribution (AGRESTI, 1990). Hence, $\hat{p}(x_i) = p(x_i) + a_{1i}(\hat{\alpha} - \alpha) + a_{2i}(\hat{\beta} - \beta) + o_p(n^{1/2})$, with $a_{1i} = \left. \frac{\partial p(x)}{\partial \alpha} \right|_{x=x_i}$ and $a_{2i} = \left. \frac{\partial p(x)}{\partial \beta} \right|_{x=x_i}$. Furthermore,

$$\sum_{x_i \in S} \hat{p}(x_i) = \hat{p}(x^*) nf(x^*) \Delta + o_p(n^{1/2}).$$

Combining terms yields

$$\begin{aligned} \frac{1}{n} \sum_{x_i \leq \hat{x}^*} \hat{p}(x_i) &= \frac{1}{n} \sum_{x_i \leq x^*} p(x_i) + (\hat{\alpha} - \alpha) \frac{1}{n} \sum_{x_i \leq x^*} a_{1i} + (\hat{\beta} - \beta) \frac{1}{n} \sum_{x_i \leq x^*} a_{2i} \\ &\quad + p(x^*) f(x^*) \Delta + (\hat{\alpha} - \alpha) a_{1*} f(x^*) \Delta \\ &\quad + (\hat{\beta} - \beta) a_{2*} f(x^*) \Delta + o_p(n^{-1/2}) \end{aligned}$$

where a_{1*} , a_{2*} are a_1 and a_2 evaluated at x^* . Neglecting terms of lower order than $n^{-1/2}$ and using $p(x^*) = p$ yields

$$\widehat{TG} \cong 2 \left\{ pF_* + F_* \sqrt{\frac{p(1-p)}{n}} \epsilon_1 + p \sqrt{\frac{F_*(1-F_*)}{n}} \epsilon_2 - \frac{1}{n} \sum_{x_i \leq x^*} p(x_i) + o_p(n^{-1/2}) \right\}$$

To compute the asymptotic distribution of \widehat{TG} we need to calculate the covariance of ϵ_1 and ϵ_2 . Using (21), it follows that $\text{Cov}(\epsilon_1, \epsilon_2) = E(\epsilon_1 \epsilon_2)$ with

$$E(\epsilon_1 \epsilon_2) = \frac{E(\sum I_i^* y_i) - F_* np - npF_* + npF_*}{\sqrt{np(1-p)} \sqrt{nF_*(1-F_*)}} = \frac{E(\sum I_i^* y_i) - npF_*}{\sqrt{np(1-p)} \sqrt{nF_*(1-F_*)}}$$

Let $p_c = E(I_i^* y_i)$. Thus, $p_c = \int_{-\infty}^{x^*} p(x) dF(x) / \int_{-\infty}^{x^*} dF(x)$. But $\sum_{i=1}^n I_i^* y_i = \sum_{i=1}^{N^*} y_i$ where N^* equals the number of x_i 's that are smaller than or equal to x^* . Therefore, $E(\sum I_i^* y_i) = nF_* p_c$ by FELLER (1966, p. 167), which in turn implies that

$$E(\epsilon_1 \epsilon_2) = \frac{nF_*(p_c - p)}{\sqrt{np(1-p)} \sqrt{nF_*(1-F_*)}} .$$

Furthermore, $\frac{1}{n} \sum_{x_i \leq x^*} p(x_i) = \frac{1}{n} \sum_{i=1}^n I_i^* p(x_i)$. Again by FELLER (1966),

$$E\left(\frac{1}{n} \sum_{x_i \leq x^*} p(x_i)\right) = F_* \frac{\int_{-\infty}^{x^*} p(x) dF(x)}{\int_{-\infty}^{x^*} dF(x)} = \int_{-\infty}^{x^*} p(x) dF(x)$$

with variance

$$\text{Var}\left(\frac{1}{n} \sum_{x_i \leq x^*} p(x_i)\right) = \frac{1}{n^2} \left\{ nF_* \text{Var}(I_i^* p(x_i)) + \frac{\left(\int_{-\infty}^{x^*} p(x) dF(x)\right)^2}{F_*^2} nF_*(1-F_*) \right\} .$$

$$\text{As } \text{Var}(I_i^* p(x_i)) = \int_{-\infty}^{x^*} p^2(x) dF(x) / F_* - \left(\int_{-\infty}^{x^*} p(x) dF(x) / F_*\right)^2 ,$$

$$\text{Var}\left(\frac{1}{n} \sum_{x_i \leq x^*} p(x_i)\right) = \frac{1}{n} \left\{ \int_{-\infty}^{x^*} p^2(x) dF(x) - \left(\int_{-\infty}^{x^*} p(x) dF(x)\right)^2 \right\} .$$

Moreover,

$$\begin{aligned} \text{Cov}\left(\frac{1}{n} \sum_{i=1}^n I_i^* p(x_i), \epsilon_2\right) &= \frac{1}{n\sqrt{nF_*(1-F_*)}} \left\{ E\left(\sum_{i=1}^n I_i^* p(x_i) (I_i^* - F_*)\right) \right. \\ &\quad \left. + E\left(\sum_{i \neq j}^n I_i^* p(x_i) (I_j^* - F_*)\right) \right\} . \end{aligned}$$

Since $x_i \perp\!\!\!\perp x_j$ for $i \neq j$, the covariance equals

$$(1 - F_*) \int_{-\infty}^{x^*} p(x) dF(x) / \sqrt{nF_*(1 - F_*)}.$$

Now, \widehat{TG} can be written as $\widehat{TG} - 2 \left\{ pF_* - \int_{-\infty}^{x^*} p(x) dF(x) \right\} \doteq \epsilon$, where

$\epsilon \stackrel{d}{=} N(0, \sigma_\epsilon^2)$, with

$$\begin{aligned} \sigma_\epsilon^2 = 4 \left\{ F_*^2 \frac{p(1-p)}{n} + p^2 \frac{F_*(1-F_*)}{n} + \frac{1}{n} F_*^2 p(p_c - p) \right. \\ \left. + \frac{1}{n} \left(\frac{\int_{-\infty}^{x^*} p^2(x) dF(x)}{F_*} - \frac{\left(\int_{-\infty}^{x^*} p(x) dF(x) \right)^2}{F_*^2} \right) \right. \\ \left. + 2p \sqrt{\frac{F_*(1-F_*)}{n}} \frac{1}{\sqrt{nF_*(1-F_*)}} (1-F_*) \int_{-\infty}^{x^*} p(x) dF(x) \right\}. \end{aligned}$$

Note that,

$$2pF_* - \int_{-\infty}^{x^*} p(x) dF(x) = 2 \int_{-\infty}^{x^*} (p - p(x)) dF(x) = TG$$

by (3). Also, let $\sigma_{\widehat{TG}}^2 = n\sigma_\epsilon^2$. After simplifying the formula for σ_ϵ^2 , we obtain Theorem 1.

References

- AGRESTI, A., 1990: *Categorical Data Analysis*. Wiley, New York.
 AGRESTI, A., 1996: *An Introduction to Categorical Data Analysis*. Wiley, New York.
 BAHADUR, R. R., 1966: A Note on Quantiles in Large Samples. *The Annals of Mathematical Statistics* **37**, 577–580.
 BILLINGSLEY, P., 1986: *Probability and Measure*. Wiley, New York.
 CHRISTENSEN, R., 1997: *Log-linear Models and Logistic Regression*. Springer, New York.
 COLLETT, D., 1991: *Modelling Binary Data*. Chapman & Hall, London.
 COPAS, J., 1999: The effectiveness of risk scores: the logit rank plot. *Applied Statistics* **48**, 165–183.
 CSÖRGO, M., GASTWIRTH, J. L. and ZITIKIS, R. (to appear). Statistical Foundations for the Analysis of Expectancy Curves. *Technical Report, Carleton University*.

- FELLER, W., 1966: *An Introduction to Probability Theory and its Applications*. Vol. 2. Wiley, New York.
- FINNEY, D. J., 1971: *Probit Analysis*. 3rd edn. Cambridge University Press, Cambridge.
- GASTWIRTH, J. L., 1971: On the Sign Test for Symmetry. *Journal of the American Statistical Association* 66, 821–823.
- GHOSH, J. K., 1971: A New Proof of the Bahadur Representation of Quantiles and an Application. *The Annals of Mathematical Statistics* 42, 1957–1961.
- GOODMAN, L. A. and KRUSKAL, W. H., 1979: *Measures of Association for Cross Classifications*. Springer-Verlag, New York.
- HASTIE, T. J., and TIBSHIRANI, R. J., 1990: *Generalized Additive Models*. Chapman & Hall, London.
- HOOK, E. B., 1981: Rates of Chromosomal Abnormalities at Different Maternal Ages. *Obstetrics and Gynecology* 58, 282–285.
- KORN, E. L. and SIMON, R., 1991: Explained Residual Variation, Explained Risk, and Goodness of Fit. *The American Statistician* 45, 201–206.
- HOSMER, P. W. and LEMESHOW, S., 1989: *Applied Logistic Regression*. Chapman & Hall, New York.
- MCCULLAGH, P. and NELDER, J. A., 1989: *Generalized Linear Models*. Chapman & Hall, New York.
- MILICER, H. and SZCZOTKA, F., 1966: Age at menarche in Warsaw girls in 1965. *Human Biology* 38, 199–203.
- MITTLBÖCK, M. and SCHEMPER, M., 1996: Explained Variation for Logistic Regression. *Statistics in Medicine* 15, 1987–1997.
- MORGAN, B. J. T., 1992: *Analysis of Quantal Response Data*. Chapman & Hall, London.
- PIGEON, J. G. and HEYSE, J. F., 1999: An improved goodness of fit statistic for probability prediction models. *Biometrical Journal* 41, 71–82.
- RAO, C. R. and ZHAO, L. C., 1995: Convergence theorems for empirical cumulative quantile regression function. *Mathematical Methods of Statistics* 4, 81–91.
- RAO, C. R. and ZHAO, L. C., 1996: Law of the iterated logarithm for empirical cumulative quantile regression functions. *Statistica Sinica* 6, 693–702.
- REYNOLDS, H. T., 1977: *The Analysis of Cross-Classifications*. The Free Press, New York.
- SPRENT, P., 1998: *Data Driven Statistical Methods*. Chapman & Hall, London.
- TIKU, M. L. and VAUGHAN, D. C., 1997: Logistic and Nonlogistic Density Functions in Binary Regression with Nonstochastic Covariates. *Statistica Sinica* 6, 693–702.

EFSTATHIA BURA
Department of Statistics
The George Washington University
2201 G Street NW
Washington, DC 20052
USA

Received, October 1999
Revised, January 2000
Accepted, January 2000

JOSEPH L. GASTWIRTH
Biostatistics Branch
Division of Cancer Epidemiology and Genetics
National Cancer Institute
Rockville, MD 20892
USA

Book Review

JEAN-PAUL CHILÈS and PIERRE DELFINGER: *Geostatistics. Modeling Spatial Uncertainty*. J. Wiley & Sons, New York, ISBN 0-471-08315-1, 1999, 695 pp., £ 80.95.

Geostatistics is a central part of spatial statistics with many applications also in fields outside of the geosciences, for example in biology and ecology. It studies phenomena that are correlated in space and time, by means of the theory of regionalised variables. This field is in very active development, which is indicated not only by many interesting applications but also by the fact that various books of different levels have recently been published in the area.

In this context the book by Chilès and Delfiner plays a prominent role. The reviewer predicts that it will soon become the handbook of the field. The community of spatial statisticians will be grateful to the authors for producing such a great and carefully written work.

Here is a list of the chapter headings, which shows the modern content and exposition:

1. Preliminaries (18 pages);
2. Structural analysis (which means variography; 121 pages);
3. Kriging (81 pages);
4. Intrinsic model of order k (61 pages);
5. Multivariate methods (83 pages);
6. Nonlinear methods (74 pages);
7. Conditional simulations (144 pages);
8. Scale effects and inverse problems (43 pages).

The authors come from the famous school of Georges Matheron at the Center for Geostatistics in Fontainebleau. Consequently, the book has the strengths of that school: very good statistical modeling and thorough mathematical analysis. Perhaps the statistical methods are not so thoroughly studied; here the older book by Cressie (1993) may still be superior. For example, little is said on model tests. A highlight of the book is the detailed description of simulation methods on 144 pages. Very valuable is also the material on modeling of porous media and determination of permeability in the last two chapters. Much material in the book is so modern that even specialists will make surprising and enlightening discoveries. Also many side remarks give important information and testify to the authors' enormous experience. As the reviewer was told, the book contains parts of Matheron's unpublished work.

It corresponds to the character of a handbook that many ideas are presented in a spirit of peaceful coexistence. The Matern class of variograms, which is considered by some researchers as so important and fashionable, is only briefly described on 13 lines, called K -Bessel model, while the spherical variogram, about which an author such as M. Stein utters that he does not understand why it is so popular, is called the 'geostatistician's best friend'. Also the level of the book is quite variable: there are pieces of deep and abstract mathematics (including passages which applied workers might consider as mathematical games) close together with simple examples and elementary facts. With the intention to introduce ideas from adjacent branches of spatial statistics, there are passages which have the character of a review paper. By no means this book is an introduction for beginners, since many ideas are explained only briefly; often the explanations have the character of memoranda for the insider.

This book is an indispensable tool and reference for any one doing applied work in the field of geostatistics.