MANCHESTER 1824

The University of Manchester

The Cathie Marsh Centre for Census and Survey Research

Constructing and Applying Risk Scores in the Social Sciences

CCSR Working Paper 2010-01

Ian Plews
Ian.Plewis@manchester.ac.uk

Social scientists are interested in associations between explanatory variables measured at an earlier point in time and later outcomes. In some contexts, it is useful to divide these explanatory variables into risk and protective variables although the literature is often confused about the distinction between them. This paper clarifies the distinction and shows how to assess the accuracy of risk scores generated from models that relate a binary outcome to a set of risk and protective variables. The receiver operating characteristic (ROC) curve and the logit rank plot are introduced and summary measures of accuracy derived from them. The ROC curve provides a framework for informing decisions about whether and how to intervene to prevent a poor outcome by taking account of the costs of misclassification. The ideas are applied to two examples: (i) predicting adult educational disadvantage from variables measured early in life; (ii) classifying and predicting non-respondents in longitudinal studies.

www.ccsr.ac.uk

CONSTRUCTING AND APPLYING RISK SCORES IN THE SOCIAL SCIENCES

IAN PLEWIS

SOCIAL STATISTICS,
UNIVERSITY OF MANCHESTER,
MANCHESTER M13 9PL,
UK.

Email: ian.plewis@manchester.ac.uk

Abstract

Social scientists are interested in associations between explanatory variables measured at an earlier point in time and later outcomes. In some contexts, it is useful to divide these explanatory variables into risk and protective variables although the literature is often confused about the distinction between them. This paper clarifies the distinction and shows how to assess the accuracy of risk scores generated from models that relate a binary outcome to a set of risk and protective variables. The receiver operating characteristic (ROC) curve and the logit rank plot are introduced and summary measures of accuracy derived from them. The ROC curve provides a framework for informing decisions about whether and how to intervene to prevent a poor outcome by taking account of the costs of misclassification. The ideas are applied to two examples: (i) predicting adult educational disadvantage from variables measured early in life; (ii) classifying and predicting non-respondents in longitudinal studies.

(152 words)

Keywords: risk variables, protective variables, longitudinal data, classification, prediction, ROC curve, logit rank plot, misclassification costs.

This paper translates to the social and behavioural sciences a situation that is often found in medicine. Firstly, how should we specify and estimate the relation between a single test or marker, or a set of such variables, and a 'true' state, which, in medicine, is often a disease state? Secondly, how accurate is this relation in terms of its ability correctly to classify existing cases and to predict new ones? And thirdly, how should this relation be used in decision making? There is a diversity of potential applications of this approach in the social sciences but rather few examples. One exception is the paper by Berk et al. (2009): given limited resources, on which cases should probation officers focus in order to try to prevent serious offences in the future? The ideas in this paper are illustrated by two quite different examples, both using longitudinal data: (i) should an intervention designed for young children and families in order to reduce subsequent educational disadvantage be targeted and, if so, at whom; (ii) might the problem of attrition in longitudinal studies be lessened by redirecting field work resources to those respondents more likely to drop out?

Relating one or more explanatory or predictor variables to a response or outcome variable has, for many years, been an everyday activity for quantitative researchers in the social and behavioural sciences. It is an activity that is usually manifested by the specification and estimation of statistical models, buttressed by the statistical theory developed under the broad heading of generalised linear modelling and facilitated by reliable and widely available statistical software. All this is widely accepted by statisticians and well known by researchers. What is more contested is the purpose of this kind of statistical modelling and the meaning of the estimated model coefficients, especially when applied to observational data,. Controversies are particularly marked when discussing the circumstances in which the model coefficients can be given a causal or explanatory interpretation. Many of the tools we use for classification and prediction are the same as those used in the analysis of observational data but they are used somewhat differently, with the focus more on the performance of a model as a whole than on the estimates for each variable separately. Nevertheless, as we shall see, questions about classification and prediction cannot be entirely separated from questions about explanation and cause, not least because it would be difficult to construct a model that predicts well without some understanding of the processes that lead to the outcome of interest.

An estimated model that relates a binary outcome (good/poor; condition/not condition etc.) to a set of predictors is labelled a risk score in this paper. Swets et al. (2000) use the term 'statistical prediction rule' but it is important to recognise that the practical value of such scores is to provide guidance to decision makers, not hard and fast rules. The main motivations for constructing a risk score are to classify and to predict and thus to aid decision making in conditions of uncertainty.

The thinking that lies behind the construction and use of risk scores might be applied more widely. A common approach, often found in developmental psychology for example, is to relate a set of childhood experiences and circumstances to psychopathological adolescent and adult outcomes. Researchers establish that some variables, conditional on the effects of others, are statistically significantly related to the outcome and these variables are often referred to as risk factors (although 'factor' is problematic here). This approach can be useful but, by concentrating on the estimates for particular predictors and some measure of model fit it is, at best, incomplete because we do not learn how well we can predict both the good and the poor outcomes nor how we might use the risk score to target (or to discourage) future interventions.

This paper addresses four main questions about risk scores with the two chosen examples providing some context:

(1) Which variables should be included in a risk score?
(2) How should a risk score be constructed?
(3) How should the accuracy of risk scores be assessed?
(4) How should risk scores be used in decision making?

The paper concludes by considering how the methods discussed here might be extended to more complex situations.

Candidate variables for risk scores

We generate a pool of predictor or marker variables for the outcome of interest on the basis of theory, previous research and by searching for associations in the data to hand. We are, of course, restricted to predictors that have been measured in a study that was not necessarily designed to generate a risk score. In addition, we want our risk scores to be applicable and this can lead to restrictions on the variables that we should include in our model. Take our first example: suppose we are interested in predicting whether or not a person obtains any educational qualifications by age 23 using only variables measured in early childhood. Suppose also that the actual values of these variables should be available at the case level to the service providers who are charged with targeting and delivering an intervention designed to reduce the number of young people without educational qualifications in the future. We would then only include in our risk score variables like mother's age, child's birth weight and whether or not the family lives in social housing even though including other, less 'public' variables (developmental status at age five for example), might improve its accuracy.

For outcomes like truancy from school, criminal behaviour or mental illness then the predictor variables are commonly divided into so-called risk and protective variables. Because there is some poor practice and confusion in the literature about the ways in which these variables are defined and used, we now give some attention to this issue.

Risk and protective variables

It is common practice to measure risk variables as binary splits (or factors) so that the poor outcome is more likely if the risk factor is present and less likely if it is not. Although convenient for expository purposes (Farrington, 2003), this is not in general good practice. We would usually expect to find a 'dose-response' relation between a risk variable and a binary outcome so that the more of the risk variable someone is exposed to, the more likely a poor outcome will be. This exposure can take two forms: intensity and duration. We would, for example, expect to find poorer outcomes for someone who had been brought up in extreme rather than in moderate poverty, and we would expect to find someone who had experienced persistent poverty to have a poorer outcome than someone who had lived in poverty for just a short time.

We might also expect that the relation will sometimes be non-linear: the probability of a poor outcome might be relatively insensitive to changes in the risk variable in the 'positive' part of its distribution, compared with changes in the 'negative' part. These kinds of relations are obscured when a risk variable is dichotomised.

We illustrate these points in the following way. We consider three kinds of relation between the outcome (y, measured as a probability) and the continuously measured risk variable (x), as shown in Figs. 1(a-c). In Fig. 1(a), the relation is a linear one so that the outcome become worse as the risk variable increases and better as the risk variable declines. Figs. 1(b) and 1(c) demonstrate non-linear relations. In Fig. 1(b), the outcome worsens as the risk variable increases beyond a certain point but it does not continue to improve as the risk variable declines. Stouthamer-Loeber et al. (2002) refer to the x variable here as a risk factor but not a 'promotive' factor. Other authors (for example, Woodward et al., 2002) have used the term 'compensatory' instead of 'promotive'. In Fig. 1(c), outcomes are sensitive to changes in the positive part of the distribution of the risk variable but are little affected by changes in the negative part and so x is a promotive variable but not a risk variable. In Fig. 1(a), x is both a risk variable and a promotive variable. Clearly we cannot distinguish between risk and promotion for binary variables; they are just opposite sides of the same coin and we select which term to use merely for convenience.

Rutter (1985) and Stattin and Magnusson (1996) define a protective variable as a variable that moderates the relation between a risk variable and the outcome. In other words, there is an important statistical interaction between the risk variable and the protective variable that is related to the outcome. For these authors, it is the interaction that is the defining characteristic of a protective variable but for others (for example, Jessor et al., 1995; Luthar et al., 2000) the interaction is regarded as necessary but not sufficient. This second group of authors give as much attention to the main effect of the protective variable, averaged over the risk variable, as they do to the interaction between the risk and protective variables. And it is not uncommon still to find the term 'protective factor' being used just as the positive end of a risk factor without any reference to an interaction. In other words, 'promotive' and 'protective' variables are elided rather than kept conceptually distinct, and this is unhelpful.

There is no universal distinction between risk and protective variables; a variable could be a risk variable for some outcomes and a protective variable for others. Although it would be possible to determine the distinction solely on empirical grounds, it will usually be more appropriate to separate risk and protective variables a priori, on theoretical grounds, and then to test for interactions in a statistical model. Just how risk and protective variables are related in combination to an outcome does, as we shall see, depend on the estimated parameters of such a statistical model.

Suppose we have an outcome y, measured as a probability of a poor outcome as before, a risk variable x_1 assumed continuous and standardised to have mean zero and S.D. = 1, with higher values representing more risk, and linear in its relation with y, and a hypothesised protective variable x_2 (again standardised) with higher values representing more protection. Then we can represent the relation between the three variables as a regression equation (ignoring, for convenience, any transformations of y (logit, probit etc.) that are likely to be used in practice):

$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_1x_2 + e$$

We expect b_1 to be positive and b_2 to be negative. For x_2 to qualify as a protective variable, $b_3 \neq 0$. If $b_3 = 0$ and $b_2 < 0$ then x_2 is just another risk variable, albeit one that is coded in the opposite direction to x_1. To illustrate how the predicted values of *y* change as x_1 and x_2 change, consider the nine combinations generated by the values –1, 0, +1 for the two explanatory variables and assume that one of the situations (a) to (c) below holds:

(a) the effects of the risk variable (b_1) and protective variable (b_2) are equal *($b_1 = -b_2$)*,
(b) the effect of the risk variable is twice as strong as the effect of the protective variable *($b_1 = -2b_2$)*,
(c) the effect of the protective variable is twice as strong as the effect of the risk variable *($b_1 = -0.5b_2$)*.
In addition, assume that for each of (a) to (c):
(d) the interaction parameter (b_3) is either positive (0.1) or negative (-0.1).

We see in Table 1 that, for combinations *(1)* and *(3)*, it is difficult to argue that x_2 really is protective because it has no effect on the probability of a poor outcome for high values of the risk variable (0.6 for *(1)* and 0.7 for *(3)*). On the other hand, for combinations *(2)* and *(6)*, x_2 is strikingly protective because a high value of the protective variable leads to the lowest probability of a poor outcome as well as moderating the effect of the risk variable so that the effect of x_1 is constant for high values of x_2 (0.4 for (*2*) and 0.3 for (*6*)). For combination *(4)* we see that the probability of a poor outcome is high when the risk variable is high even if the protective variable takes a high value, whereas for combination *(5)* it is the relative absence of protection rather than relative presence of risk that generates a poor outcome. The results in Table 1 show how inferences about risk and protection depend on the estimated values of *all* the parameters in a model and not just on the interaction parameter, and illustrate that care is needed before variables can properly be labelled ‘protective’.

Constructing risk scores

There are at least three ways of constructing risk scores for binary outcomes. The first – and the subject of most of the attention in this section – is to model the transformed probability as a function of the predictor variables. This is a straightforward application of logit or probit modelling (Freedman, 2005). The second is to employ methods of statistical learning such as classification trees and random forests as used by Berk et al. (2009). These methods are not discussed in this paper but they do appear, at least in some circumstances, to be able to improve predictions and hence decision making. The third – and least satisfactory approach - is to generate a risk score just by adding up the number of risk factors.

To return to the point made earlier: dichotomising a risk variable wastes information and can obscure its underlying relation with the outcome. As soon as we move away from considering bivariate associations between individual risk and protective variables and an outcome to considering classification and prediction from sets of these variables, our aim should be to find a function of those variables that best discriminates between the two outcome groups. Relating the probability of a poor outcome to a combination of variables by fitting a logistic or probit regression does that job and provides a set of what are essentially weights or loadings for each or risk (i.e. explanatory) variable. These weights will, in practice, vary in size to reflect the relative importance of each variable in terms of its ability to predict the outcome. Giving equal weights to a binary version of each risk variable, as is implied by a risk score (see, for one of many examples, Stattin and Magnusson (1996)), cannot provide an optimum solution. Instead, any person's risk, in terms of their predicted probability of a poor outcome, should be determined by their own combination of scores on those risk (and protective) variables shown to relate to the outcome. This then makes redundant arguments about linear or threshold effects for cumulative risk as proposed by, for example, Appleyard et al. (2005) whose approach is widely applied. They (i) dichotomise risk variables measured either on a continuous scale or as ordered categories by taking the top quartile as the cut-off; (ii) give any score above the cut point a value of one; (iii) add up these values to get an overall score; (iv) combine the small groups at the high risk extreme. Each of these four steps wastes potentially valuable information.

We now elaborate on our first example, based on data from the National Child Development Study (NCDS). See Plewis et al. (2004) for more details about the study design. We focus on the failure to obtain any educational qualifications by age 23 (27% of the sample) as our outcome measure and we choose four risk variables from early in the child's life – (1) birth weight, (2) mother's age at birth, (3) being in care before age seven and (4) living in social housing. Birth weight and mother's age are measured on continuous scales whereas 'social housing' and 'in care' are binary. We also consider one potential protective variable: maternal grandfather's social class. The social class variable is a categorical variable with six categories ordered from one (professional) to six (unskilled). Descriptive statistics are given in Table 2. Although primarily illustrative, our model is nevertheless plausible in terms of predicting the probability of a poor adult outcome (no educational qualifications) from variables measured early in the child's life. There is a substantial body of evidence (e.g. Currie and Hyson, 1999) to show that babies with low birth weight do less well in later life. We would not, however, expect the relation between birth weight and the outcome to be linear. Rather we would expect the relation between the probability of a poor outcome and birth weight to be like Fig. 1(b) – with high birth weight corresponding to low risk - because we would not expect birth weight to be a promotive variable. Turning to mother's age, again we would expect a non-linear relation, with young mothers and possibly older mothers having children with poorer outcomes. There is no choice about how to specify the relation with the binary housing and care variables. Social housing is a proxy for poor socio-economic circumstances (both of the family and the area in which they live) and so we would expect children living in social housing to do less well later on. It is well known that children who experience a period being brought up in care rather than by one or both parents fare less well in adulthood (Roy et al., 2000). The choice of grandfather's social class as a potential protective variable is motivated by the possibility that grandparents from a higher social class can offer both financial support and additional forms of capital to their children and grandchildren that could moderate the effects of the other risk variables.

We find that the best logistic regression model relating the four risk variables and one protective variable to the outcome includes (i) the reciprocal of birth weight, (ii) mother's age and mother's age squared, (iii) social housing, (iv) in care, (v) grandfather's social class and (vi) an interaction between social housing and grandfather's social class. The estimates, their standard errors and the model fit are given in Table 2.

The fitted model brings out a number of issues. It predicts that, holding all other variables constant, the odds of a poor outcome are 1.3 times higher if birth weight is two standard deviations below the mean than if it is one standard deviation below but only 1.1 times higher if birth weight is two rather than one SD above the mean, thus confirming that birth weight is indeed a risk variable but not really a promotive one. For mother's age, we find that a child born to a mother two SDs below the mean age is 1.6 times more likely to have a poor outcome than a child born to a mother one SD below the mean. However, a change from one to two SDs above the mean raises the odds of a poor outcome by just 1.1, again in line with our hypothesis. Living in social housing (and hence more likely to be in poverty) and being in care raises the odds of a poor outcome by 3.3 and 2.7 respectively.

As we find an interaction between grandfather's social class and social housing, grandfather's social class satisfies the necessary condition for being a protective variable. However, the estimates given in Table 3 are such that it is difficult to regard grandfather's social class as being protective. Holding the other variables constant, we find (from the right-hand column in Table 3) that living in social housing rather than in private housing in the early years means that you are over three times as likely to have no educational qualifications in adulthood if your maternal grandfather was in the highest social class but less than twice as likely if your grandfather was in the lowest social class. This does not, of course, gainsay the advantage of having a grandfather in a high social class: the column headed 'private housing' in Table 3 shows that you are 4.5 times more likely to have no educational qualifications at age 23 if you lived in private housing and your maternal grandfather was from social class six (unskilled) rather than from social class one (professional). But having such a grandfather cannot protect you from the disadvantages of your own family's poverty (as represented by the social housing variable).

On the basis of this analysis, it is difficult to make out a case for grandfather's social class to be classified as a protective variable. We could re-classify it as a risk variable and this would improve the accuracy of the model. Instead, we re-estimate the model without grandfather's social class (on a larger sample because there are fewer missing values). The exclusion of grandfather's social class and its interaction with social housing has little effect on the other estimates except that the estimate for social housing is now smaller and the interaction between mother's age and social housing becomes important.

Our second example is rather different and draws on an analysis of the predictors of non-response at the second wave of the Millennium Cohort Study, the fourth in the series of UK birth cohort studies (see Plewis, 2007). A major reason for analyses of this kind is to construct functions that can be used to generate non-response weights and also to determine which variables might be used to impute missing data (Carpenter and Plewis, 2010). We find that a range of variables measured at wave one all combine to predict overall non-response at wave two (see Table 2 in Plewis, 2007 for the variables and the estimates from the logistic regression). Although the distinction between risk and protective variables is not as relevant here as it is in the developmental context of the first example, it is interesting to note that the model does include some interactions. For example, moving house between waves one and two and living in Northern Ireland both predict non-response but the combination of these two variables does not increase the risk of non-response. We would like to know how well this risk score classifies cases into respondents and non-respondents as this will help us to assess the utility of non-response weights derived from the risk score. We would also like to know how well it predicts drop-out in order to consider improved procedures for allocation of resources for data collection.

Assessing the accuracy of a risk score

Estimates such as those given in Table 2 provide some useful insights into why children become educationally disadvantaged but they do not tell us very much about the utility of targeting interventions at children with high risk scores. The predicted probabilities of educational disadvantage generated by the risk score that excludes grandfather's social class range from 0.12 to 0.90. This range of predicted probabilities raises the question of where the optimum cut point should be such that everyone with a predicted probability at least as great as the cut point is expected to have a poor outcome and everyone with a predicted probability less than the cut point is expected to have a good outcome. We return to this issue in the next section. First, we consider how to measure the accuracy of the risk score.

There are two related components of accuracy in the situation just described: classification (or discrimination) and prediction (Pepe, 2003). Classification refers to the conditional probabilities of having a positive test given that a person either does or does not have the undesired outcome. For most applications in the social sciences, a positive test means a score (*s*) above a chosen threshold (*t*) on a continuous scale. Prediction, on the other hand, refers to the conditional probabilities of a poor outcome given either a positive or negative score on the test.

More formally, let D and $\bar{D}$ refer to the presence and absence of the poor outcome and + ($s > t$) and – ($s \le t$) refer to positive and negative tests. Then, for classification, we are interested in P(+|D), the true positive fraction (TPF) or sensitivity of the test, and P(-|$\bar{D}$), its specificity, equal to one minus the false positive fraction (1 – FPF). For prediction, however, we are interested in P(D|+), the positive predictive value (PPV) and P($\bar{D}$|-), the negative predictive value (NPV). Classification and prediction, defined in this way, are linked by Bayes' Theorem in that:

$$P(D|+) = P(+|D)\ P(D)/P(+)$$

and

$$P(\bar{D}|-) = P(-|\bar{D})\ P(\bar{D})/P(-)$$

Hence, if the probability of a positive test (P(+) = τ) is the same as the prevalence of the poor outcome (P(D) = ρ) then inferences about classification and prediction are essentially the same. Then sensitivity equals PPV and specificity equals NPV. Generally, however, (TPF, FPF, ρ) and (PPV, NPV, τ) convey different pieces of information.

As we have seen, we estimate the relation between a set of explanatory or predictor variables and the outcome for a sample of cases for which the outcome is known, and usually with the intention of applying the score function so obtained to a sample of new cases for which the predictors are known but the outcome is unknown. The classifications and predictions that are made using this score function are nearly always imperfect. Although a positive test is generally more likely in the presence of a poor outcome than in its absence and a poor outcome is generally more likely as the test score increases, the distribution of the test score for cases with poor outcomes overlaps with the same distribution for cases with good outcomes: some cases scoring high on the test ($s > t$) turn out to have good outcomes (false positives) whereas some cases scoring low ($s \leq t$) turn out to have poor outcomes (false negatives).

In order to make more use of the information from the risk score in terms of classification and prediction, it would be helpful to summarise it. A widely used method of assessing the goodness-of-fit of models for binary or categorical outcomes is to use one of several possible pseudo-R^2 statistics. Apart from their rather arbitrary nature, which thus makes comparisons across datasets difficult, they are not useful in this context because they assess the overall fit of the model and do not distinguish between the accuracy of the model for the good and poor outcomes separately. Instead we turn to techniques that have been widely used in medicine but rarely in social science.

Receiver Operating Characteristic curves

A popular way of summarising the information about classification is to estimate the sensitivity and specificity at *each* value of the predicted probability of a poor outcome from the model and then plotting sensitivity against (1-specificity), that is TPF against FPF. This is known as a receiver operating characteristic (ROC) curve. The ROC curve is always anchored at coordinates (0,0) and (1,1) and for large samples and at least some continuously measured risk variables it is smooth with a monotonically declining but always positive slope. Krzanowski and Hand (2009) give a detailed discussion of how to estimate ROC curves. The ROC curve generated from the model for the NCDS data is given as Fig. 2. The diagonal line joining the point [0, 0] (sensitivity = 0, specificity = 1: everyone is predicted not to have a poor outcome so the cut-point on the probability scale is one) to [1, 1] (sensitivity = 1, specificity = 0: everyone is predicted to have a poor outcome so the cut-point is zero) is the ROC that would be obtained if the chosen variables that make up the risk score do not explain any of the variation in the outcome. Consequently, it is the AUC – the area enclosed by the ROC curve and the diagonal – that is of interest and this can vary from 1 down to 0.5. In Fig. 2, the AUC is 0.66 with a 95% confidence interval of 0.64 to 0.67. One way of interpreting the value of 0.66 is to say that, presented with a set of randomly chosen pairs, each pair known to contain one positive and one negative outcome, and with information on all the explanatory variables, then someone using this risk score by always classifying the case with the higher score as having a poor outcome would be correct 66% of the time compared to 50% if they chose randomly.

The AUC can be transformed to a more natural scale that varies from zero to one by calculating 2*AUC -1, often referred to as a Gini coefficient and equal to 0.31 (95% CI: 0.29 – 0.33) for Fig. 2.

One of the advantages of using ROC curves to assess prediction is that, providing the ROC curves do not cross, the estimated Gini coefficients can be compared across different models fitted to the same data. So, if we dichotomise birth weight at 2.5 Kg. (88 oz.) and mother's age at 20 and then estimate the model with four binary risk factors rather than two continuous and two binary risk variables, we find that the actual model has essentially the same form as before, but the Gini coefficient is now only 0.27 compared with 0.31 for the model with continuous (reciprocal) birth weight and linear and quadratic terms for mother's age.

Moreover, we can compare these two models with one based on a risk score obtained by adding the binary scores. In principle, this risk score varies from zero to four and must be constructed to allow for the interaction between mother's age and social housing (i.e. socially housed young mothers contribute one and not two to the risk score). In fact, we find that nobody gets a score of four in this sample and that only 3% of the sample score over one. The Gini coefficient is now only 0.26 and illustrates that the method of adding up risk factors is not only less accurate than using all the information on the risk variables but is also less accurate than using the variables separately as binary risk factors.

Turning to our second example about non-response, we find that the Gini coefficient for the model used by Plewis (2007) is 0.39 (95% CI: 0.37 to 0.41). One of the difficulties faced by researchers in this area is knowing which predictors to use when trying to classify cases into productive (or respondents) and non-productive (or non-respondents) in future waves. Consequently, it is interesting to note that the introduction of another explanatory variable into the model – whether or not the respondent gave consent at wave one to have administrative data from birth records linked to the survey data – raised the Gini coefficient only slightly (to 0.40) although the consent variable is highly predictive of future non-response. Another feature of research into non-response in longitudinal studies is that the predictors of refusal and other kinds of non-response (not located, not contacted etc.) differ. Again, we can get more insight into this by estimating separate ROC curves for these two aspects of non-response. We find that the Gini coefficient for classification into refusers and productive interviews is 0.39 (95% CI: 0.36 – 0.41) but rises to 0.52 (95% CI: 0.49 – 0.54) for classifying into other non-productives and productives.

Logit rank plots

Copas (1999) proposes the logit rank plot as a means of assessing the predictiveness of a risk score. If the risk score is derived from a logistic regression then a logit rank plot is a plot of the linear predictor from the logistic regression model against the logistic transform of the proportional rank of the risk scores. More generally, it is a plot of logit(p_i) against the logits of the proportional ranks (*r/n*) where p_i is the estimated probability of a poor outcome for case *i* and *r* is the rank position of case *i* (*i = 1..n*) on the risk score. The slope of this relation – which can vary from zero to one - is a measure of the predictive power of the risk score. The slope is scale-independent and can therefore be used to compare risk scores for the outcome of interest. A good estimate of the slope can be obtained by calculating quantiles of the variables on the y and x axes and then fitting a simple regression model.

Moskowitz and Pepe (2004) propose a slightly different approach, plotting the PPV against the rank and then estimating the slope. Because the LHS of this equation involves an inequality i.e. P(D|*s* > *t*) as opposed to Copas' P(D|*s* = *t*) where *t* is the cut point for the risk score *s*, a more complicated estimation method such as Generalized Estimating Equations is needed. Pepe et al. (2008) suggest using a predictiveness curve which is a plot of the predicted risk against the proportional rank and then linking chosen thresholds from this curve to the ROC curve.

Copas' approach is used here and works well for our second example. The estimate of the slope from the logit rank plot (Fig. 3) based on 40 quantiles when we compare refusal with being productive is 0.42 (approximate 95% CI: 0.40 – 0.44; R^2 = 0.98) but is 0.64 (approximate 95% CI: 0.61 – 0.66; R^2 = 0.98) for the comparison of other non-productive and productive, bringing out the point that other non-productives are more predictable than refusals. When we apply the method to our first example about educational disadvantage, we find that the linear regression fits less well – the estimate of the slope is 0.31 (approximate 95% CI: 0.28 – 0.34) but the R^2 is only 0.91. This arises because the distribution of the predicted probabilities from the risk score is bimodal (Fig. 4), a result of the powerful influence of the social housing variable on later educational qualifications that effectively partitions the sample into two sub-samples.

Decisions from risk scores

The ROC represents the balance between sensitivity and specificity for different cut points on the risk score. The Gini coefficient and AUC derived from the ROC provide us with an assessment of the discriminatory power of the risk score. The slope derived from the logit rank plot provides a useful summary of the predictive power of the risk score. Up to now, however, we have kept the question of the optimum cut point or threshold in the background. This optimum will depend on the prevalence of the outcome and also on the costs and benefits of the inevitable errors that occur because of the difficulties of constructing a risk score that discriminates perfectly between cases with good and poor outcomes.

Suppose that we want to intervene to prevent an undesired outcome and we want to target our resources in the most efficient way. In other words, we decide to intervene in such a way that everyone with a score (or predicted probability) above the cut point is eligible to receive the intervention and nobody with a score below the cut point receives it. One way of determining the optimum cut point is to maximise the utility or expected value of a decision in terms of its costs (generated by incorrect decisions) and benefits (generated by correct decisions).

The slope of the ROC, *a*, is just the ratio of two densities:

$a = f(s|\mathrm{D})/f(s|\bar{\mathrm{D}})$ where *s* is the risk score.

Swets et al. (2000) state that, based on earlier work in signal detection theory, the optimal cut point, a^o, is determined from the slope of the ROC curve such that:

$$a^o = k * (b_{00} + c_{01})/(b_{11} + c_{10})$$

where *k* is the odds of a good outcome (and therefore *k* is usually greater than one), *b* and *c* stand for benefits and costs and their first subscript represents the outcome (0, good; 1, poor) and the second subscript the absence (0) or implementation (1) of the intervention. We can reasonably argue that, when targeting interventions, b_{00} is zero (because the intervention is not implemented) and so the formula reduces to multiplying the odds of a good outcome, *k*, by the ratio, *R*, of the costs of intervening when a good outcome would have occurred without intervention (c_{01} - the costs of the false positives) to the net costs of intervening to prevent a poor outcome (b_{11} + c_{10}) where *b* and *c* are opposite in sign and c_{10} is the cost of the false negatives.

The value of *R* will depend on the efficacy of the intervention; it will be substantially less than one for effective interventions where the benefits exceed the costs but substantially greater than one for less effective ones. Consequently, a^o will be smaller for effective interventions than it will be for less effective ones. A relatively ineffective intervention is targeted only at those few cases with a high probability of a poor outcome - where the slope of the ROC is high, sensitivity is low but specificity is high – whereas an effective intervention is made available to many more cases with consequent higher sensitivity but lower specificity.

We can illustrate these points with our second example, predicated on the idea that we would like to prevent refusals in a longitudinal study. The proportion of refusals at wave two in the Millennium Cohort Study is 0.091 so the value of *k* is 0.909/0.091 = 10. Consider three situations: an intervention that is expected to be very successful in terms of preventing refusal so that *R* is 0.33 (benefits three times the size of costs), a moderately successful intervention with *R* = 0.8 and an intervention with a small chance of success so that *R* =1.5. The values of a^o are 3.3 (10 x 0.33), 8 and 15. Each value of a^o corresponds to a predicted probability of a poor outcome because the ROC curve is determined by the sensitivities and specificities for each predicted probability from the risk score and so *a* is positively associated with the cut-point. Hence, the implications of these values of a^o are that, for the very successful intervention, the optimum strategy is to target the intervention at the top 1% of the sample (i.e. predicted probabilities of a poor outcome > 0.4) but not to intervene at all for the intervention with at best only a moderate chance of success. This cut point is shown as point A in Fig. 5 and corresponds to the following fractions:

TPF = 0.032
FPF = 0.0055
PPV = 0.40
NPV = 0.90.

The sensitivity (i.e. TPF) is, as expected, low and the specificity (i.e. 1 – FPF) is very high. Values of 0.40 and 0.90 for PPV and NPV respectively mean that 40% of those scoring above the threshold are predicted to be refusals and 90% of those scoring below the threshold are predicted to be productive.

Of course, if the ROC curve were further away from the diagonal line – in other words, if the risk score classified more accurately – then the optimum strategy would change and it might, for example, be worth using an intervention with only a small chance of success on those cases most at risk. The upper curve in Fig. 5 illustrates this with point A* to the right of point A and point B* - corresponding to a less successful intervention – also appearing.

We would, ideally, apply the same reasoning to our first example. This is not, however, possible because, as intimated by Fig. 4, the densities $f(s|D)$ and $f(s|\bar{D})$ are both bimodal and therefore there is not a unique solution for a^o.

Discussion

We have seen how to separate risk variables from protective variables in a developmental context, and how these two kinds of variables can be combined in a statistical model to generate a risk score for a binary outcome. And even if the concept of protection does not apply in all situations, the model in our second example demonstrates the importance of checking for interactions between risk variables.

One of the advantages of summarising the information contained in risk scores by using Gini coefficients and slopes of logit rank plots as measures of accuracy is that it makes it easier to compare the efficacies of different risk scores. This was illustrated in our second example where the introduction of a new predictor variable into the model had little effect on its ability to classify. The analysis of the data in our first example was based on the experiences of a cohort born in Great Britain in 1958. It would be possible to use data from the 1970 cohort (BCS70) to see whether the accuracy of the risk score changes, especially as that cohort has a lower proportion of adults with no educational qualifications.

One possible conclusion to be drawn from the fact that the accuracy of the risk scores is relatively low in both our examples is that it would be better not to intervene at all in these situations, or at least not to try to target interventions at sub-groups of the population. The evidence from the ROC curves does not, however, necessarily support this view. The optimum size of the intervention group depends on the effectiveness of the intervention. For interventions likely to have only a modest effect on an outcome – and the research evidence indicates that the effect sizes or treatment effects of most social interventions relevant to the example about educational disadvantage are not large – then the analysis suggests that the intervention group should consist only of cases with a high risk of a poor outcome. In other words, interventions should be highly targeted.

On the other hand, the analysis presented here does not explicitly take into account resource constraints. Interventions are sometimes designed as a package with a set cost per case ‘treated’ and so it might not be possible to afford to target it in the way that the analysis suggests is optimum. Also, the analysis is based on the assumption that interventions are directed at just one outcome. It would be more realistic to suppose that the aim of multi-faceted interventions like the UK Children's Fund for example (Edwards et al., 2006) is to improve several outcomes. This would require a multivariate approach to the analysis and would be statistically much more challenging.

There are a number of methodological issues that warrant further investigation. We have seen, in our first example, how the risk score can generate a bimodal distribution of the predicted probabilities of a poor outcome when there is a dominant binary marker (social housing in this case). This problem might be reduced by including more predictors in the model although there are no obvious candidates in this example, given our wish to restrict ourselves to variables measured early in life. Another solution might be to substitute a categorical or continuous variable for a binary one so we might, for example, use years lived in social housing early in life rather than a measure at a particular point in time.

We might want to use as our outcome a variable with more than two categories. We did this in our second example by fitting two separate logistic regressions (productive against refusals and productive against other non-productive). This would be an unsatisfactory approach if the categories were ordered as would happen if, for example, we were interested in no, lower and higher educational qualifications as our outcome in the first example. One approach could then be to model the two continuation ratios – the odds of some qualifications and the odds of a higher rather than a lower qualification – and construct ROCs and logit rank plots from these two separate models.

We need to bear in mind that our estimated measures of accuracy will be biased upwards if the same data are used both to estimate the model parameters and to estimate accuracy. This is the problem of 'shrinkage' considered in detail by Copas (1999) and Copas and Corbett (2002). We can get round this problem in a number of ways. For small samples, we can leave out one observation at a time from the estimating process and then use the model to predict the outcome for the omitted observation. For large samples, as here, the tactic of dividing the sample into two random halves and then using the estimated model from the first half (the 'training' set) to predict the outcomes of members of the second set (the 'test' set) is a reasonable one. Applying this method to the model for educational disadvantage showed that the Gini coefficient is unchanged (0.31) although the 95% confidence interval is a little wider.

The decision rule based on the ROC curve assumes that, to maximise the utility, all available resources, in the form of an intervention, are directed at the group predicted to have a poor outcome. An alternative rule is generated by maximising a utility function that is based on the number of poor outcomes that are prevented, subject to the constraint that the available resources are fixed. Given that some poor outcomes will emerge from the group predicted to have a good outcome, it is likely, at least a priori, that some of the available resources will be allocated to the group predicted to have a good outcome. This is the problem addressed by Alberman and Goldstein (1970) and by Goldstein (1972) but not considered in this paper.

Conclusion

There are many examples in the quantitative social science literature of a binary outcome at time t being linked to a set of explanatory variables measured at times t – k (k = 1, 2..) via a logistic or probit model of some kind. The argument in this paper is not that this is an unhelpful approach but that it can be limited. The limitations arise because there tends to be too strong a focus on individual coefficients and not enough on the overall ability of the model to discriminate between good and poor outcomes, to examine the classification and prediction of good and poor outcomes separately, and to consider the implications of the model for possible interventions. By drawing on techniques such as ROC curves and logit rank plots, and by assessing the costs of misclassification, we can look at the problems of classification and prediction in a new light that could lead to policies and decisions being more firmly based on evidence.

Acknowledgements

A version of this paper formed the keynote address at the XI Congreso de Metodología de las Ciencias Sociales y de la Salud, Malaga, Sept. 2009. I would like to thank Jesús Rosel for useful comments. Some of the work for the paper was funded by UK Economic and Social Research Council grants to the International Centre for Life Course Studies in Society and Health (RES-596-28-0001) and as part of the Survey Design and Measurement Initiative (RES-275-25-0010).

References

Alberman, E. D. and Goldstein, H. (1970). The ‘At Risk’ register: A statistical evaluation. *British Journal of Preventive and Social Medicine*, 24, 129-135.

Appleyard, K., Egeland, B., van Dulmen, M. H. M. and Sroufe, L. A. (2005). When more is not better: The role of cumulative risk in child behavior outcomes. *Journal of Child Psychology and Psychiatry*, 46, 235-245.

Berk, R., Sherman, L., Barnes, G. et al. (2009). Forecasting murder within a population of probationers and parolees: A high stakes application of statistical learning. *Journal of the Royal Statistical Society, A*, 172, 191-211.

Carpenter, J. and Plewis, I. (2010). Analysing longitudinal studies with non-response: Issues and statistical methods. In M. Williams and P. Vogt (Eds), *Handbook of methodological innovations*. Newbury Park, Ca.: Sage.

Copas, J. (1999). The effectiveness of risk scores: the logit rank plot. *Applied Statistics*, 48, 165-183.

Copas, J. B. and Corbett, P. (2002). Overestimation of the receiver operating characteristic curve for logistic regression. *Biometrika*, 89, 315-331.

Currie, J. and Hyson, R. (1999). Is the impact of health problems cushioned by socioeconomic status? The case of low birth weight. *American Economic Review*, 89, 245-250.

Edwards, A., Barnes, M., Plewis, I. and Morris, K. (2006). *Working to prevent the social exclusion of children and young people: Final lessons from the national evaluation of the Children’s Fund.* London: Department for Education and Skills.

Farrington, D. P. (2003). Key results from the first forty years of the Cambridge study in delinquent development. In T. P. Thornberry and M. D. Krohn (Eds), *Taking stock of delinquency: An overview of findings from contemporary longitudinal studies*. New York: Kluwer.

Freedman, D. A. (2005). *Statistical models. Theory and practice.* Cambridge: CUP.

Goldstein, H. (1972). The allocation of resources in population screening: A decision theory model. *Biometrics*, 28, 499-518.

Jessor, R., Van Den Bos, J., Vanderryn, J. et al. (1995). Protective factors in adolescent problem behaviour: Moderator effects and developmental change. *Developmental Psychology*, 31, 923-933.

Krzanowski, W. J. and Hand, D. J. (2009). *ROC curves for continuous data*. Boca Raton, Fl.: Chapman and Hall/CRC.

Luthar, S. S., Cicchetti, D. and Becker, B. (2000). Research on resilience: Response to commentaries. *Child Development*, 71, 573-575.

Moskowitz, C. S. and Pepe, M. S. (2004). Quantifying and comparing the predictive accuracy of continuous prognostic factors for binary outcomes. *Biostatistics*, 5, 113-127.

Pepe, M. S. (2003). *The statistical evaluation of medical tests for classification and prediction*. Oxford: OUP.

Pepe, M. S., Feng, Z., Huang, Y. et al. (2008). Integrating the predictiveness of a marker with its performance as a classifier. *American Journal of Epidemiology*, 167, 362-368.

Plewis, I. (2007). Non-response in a birth cohort study: The case of the Millennium Cohort Study. *International Journal of Social Research Methodology*, 10, 325-334.

Plewis, I., Calderwood, L., Hawkes, D. and Nathan, G. (2004). *National Child Development Study and 1970 British Cohort Study technical report: Changes in the NCDS and BCS70 populations and samples over time (1st ed.).* London: Institute of Education, University of London.

Roy, P., Rutter, M. and Pickles, A. (2000). Institutional care: Risk from family background or pattern of rearing? *Journal of Child Psychology and Psychiatry*, 41, 139-149.

Rutter, M. (1985). Resilience in the face of adversity: protective factors and resistance to psychiatric disorder. *British Journal of Psychiatry*. 147, 598-611.

Stattin, H. and Magnusson, D. (1996). Antisocial development: A holistic approach. *Development and Psychopathology*, 8, 617-645.

Stouthamer-Loeber, M., Farrington, D. P., Loeber, R. et al. (2002). Risk and promotive effects in the explanation of persistent serious delinquency in boys. *Journal of Consulting and Clinical Psychology*, 70, 111-123.

Swets, J. A., Dawes, R. M. and Monahan, J. (2000). Psychological science can improve diagnostic decisions. *Psychological Science in the Public Interest*, 1, 1-26.

Woodward, L. J., Fergusson, D. M. and Horwood, L. J. (2002). Deviant partner involvement and offending risk in early adulthood. *Journal of Child Psychology and Psychiatry*, 43, 177-190.

Table 1. Predicted probabilities of a poor outcome for varying risk and protective variable scores

Risk (x_1)	Protective (x_2)	Interaction (x_1*x_2)	*(1):* b_1 = 0.1 b_2 = -0.1 b_3 = 0.1	*(2):* b_1 = 0.1 b_2 = -0.1 b_3 = -0.1	*(3):* b_1 = 0.2 b_2 = -0.1 b_3 = 0.1	*(4):* b_1 = 0.2 b_2 = -0.1 b_3 = -0.1	*(5):* b_1 = 0.1 b_2 = -0.2 b_3 = 0.1	*(6):* b_1 = 0.1 b_2 = -0.2 b_3 = -0.1
-1	-1	1	0.6	0.4	0.5	0.3	0.7	0.5
-1	0	0	0.4	0.4	0.3	0.3	0.4	0.4
-1	1	-1	0.2	0.4	0.1	0.3	0.1	0.3
0	-1	0	0.6	0.6	0.6	0.6	0.7	0.7
0	0	0	0.5	0.5	0.5	0.5	0.5	0.5
0	1	0	0.4	0.4	0.4	0.4	0.3	0.3
1	-1	-1	0.6	0.8	0.7	0.9	0.7	0.9
1	0	0	0.6	0.6	0.7	0.7	0.6	0.6
1	1	1	0.6	0.4	0.7	0.5	0.5	0.3

Note

b_0 = 0.5 throughout.

Table 2. Risk and protective variables: descriptive statistics and estimates from a logistic regression model applied to NCDS data

Explanatory variable	**Mean (S.D.)**	**Estimate (S.E.)**
Birth weight (reciprocal)	118 (18) [(a)]	125 (16)
Mother's age	28 (5.6)	-0.013 (0.005)
Mother's age squared	n.a.	0.002 (0.001)
Social housing	0.40 (0.49)	1.2 (0.15)
In care	0.018 (0.13)	1.0 (0.17)
Grandfather's social class	0.24 (0.43) [(b)]	0.30 (0.030)
Social housing * grandfather's social class	n.a.	-0.14 (0.044)

Notes

[(a)] Mean (SD) for untransformed birth weight

[(b)] Proportion 'middle class'

Sample size = 10279 except for grandfather's social class (n = 8518).

Model fit: χ^2 = 638 (7df); n = 8518.

Table 3. Predicted change in odds and relative odds of a poor outcome by grandfather's social class

Grandfather's social class	**Private housing**	**Relative odds; public:private**
1 (reference class)	1	3.5
2	1.4	3.0
3	1.8	2.6
4	2.5	2.3
5	3.3	2.0
6	4.5	1.7

Fig. 1(a) x is both a risk and a promotive variable

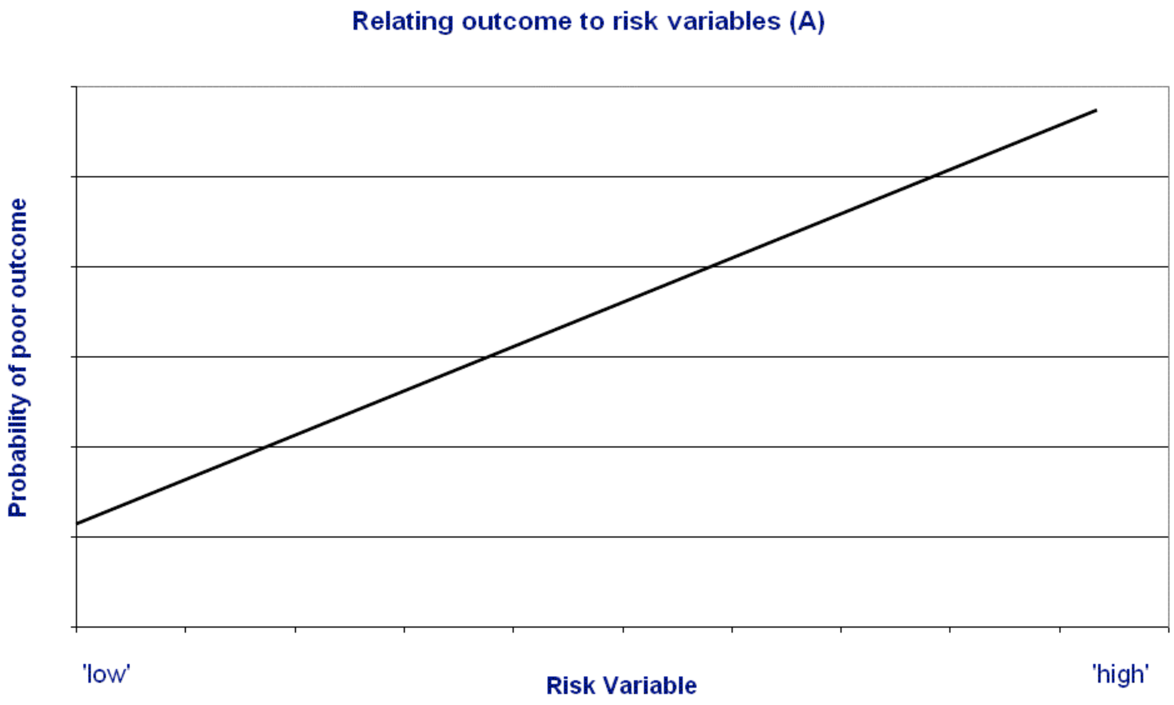


Fig. 1(b) x is a risk but not a promotive variable

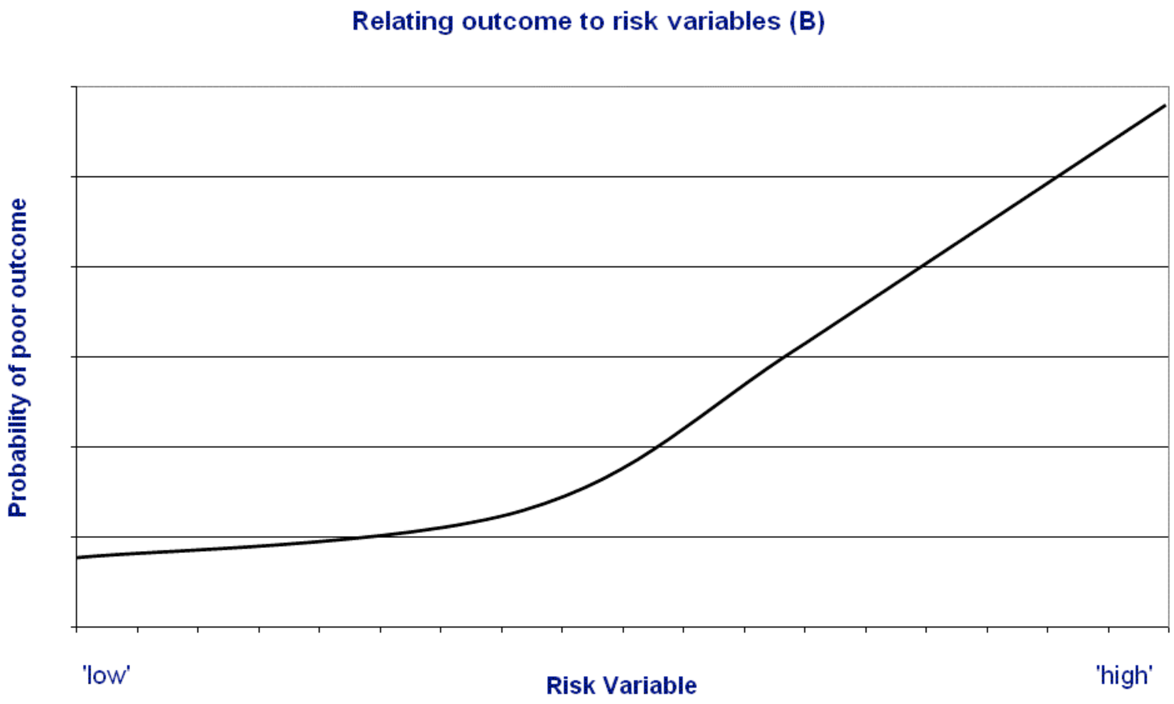


Fig. 1(c) x is a promotive but not a risk variable

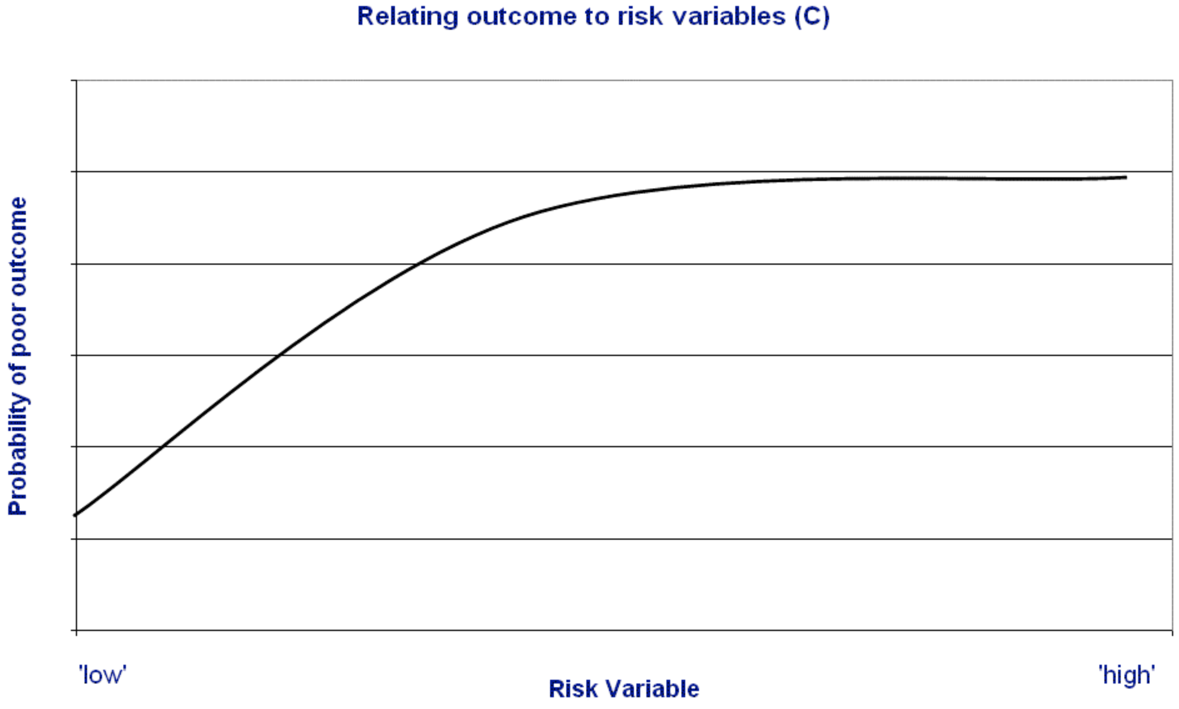
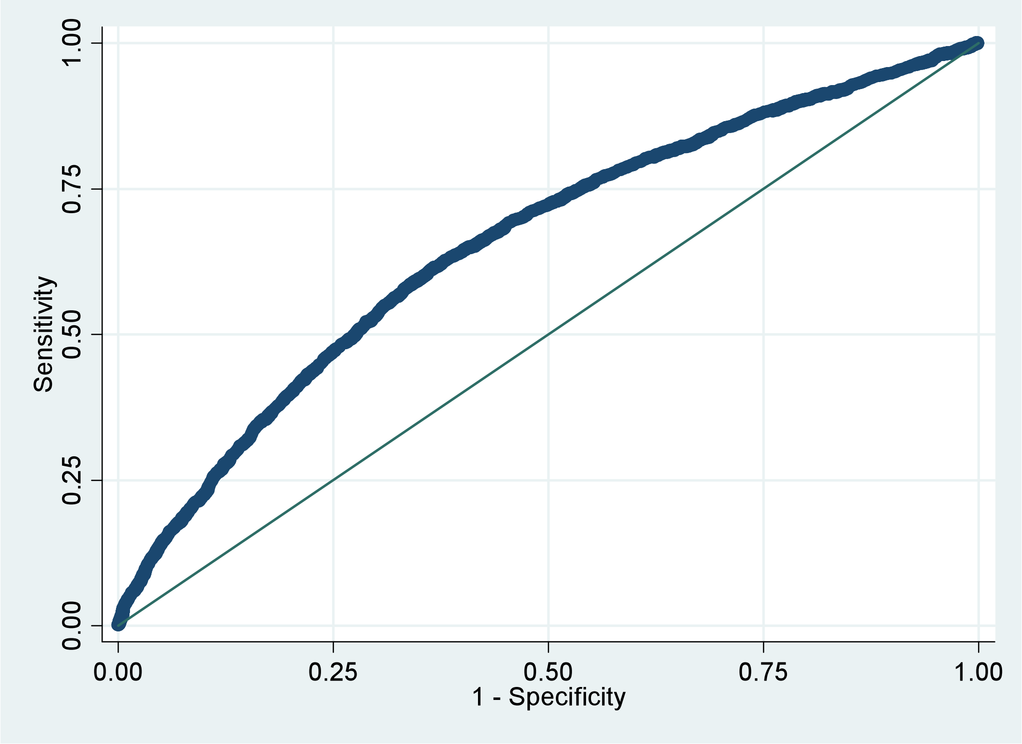


Fig. 2 ROC curve



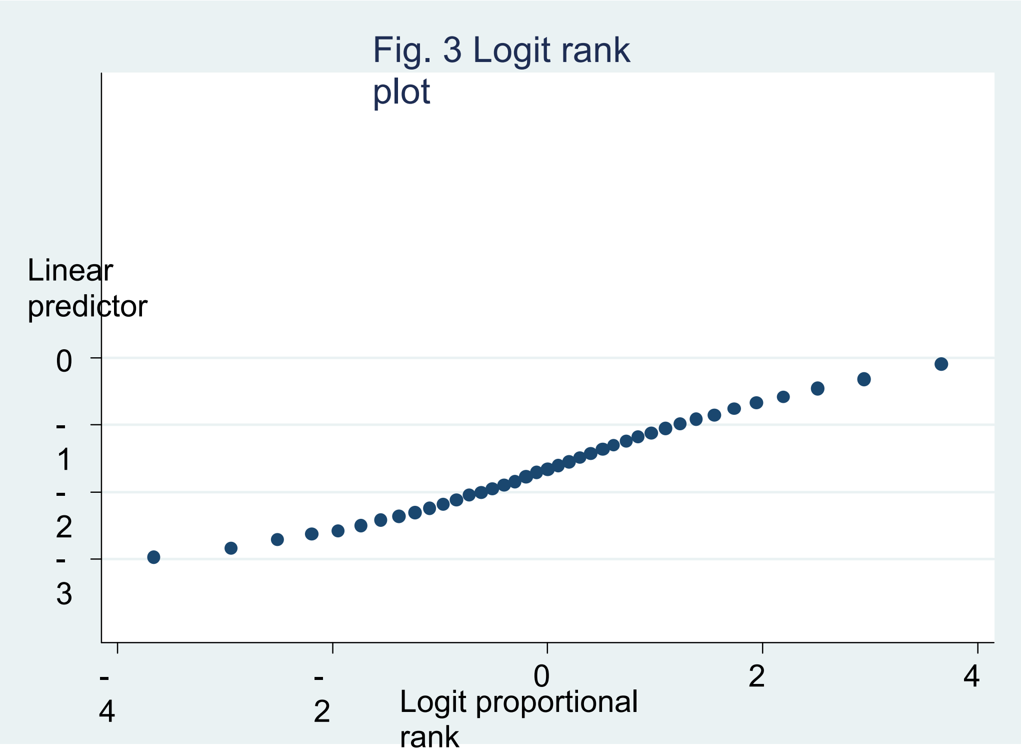


Fig. 3 Logit rank plot

Fig. 4 Predicted probabilities of no educational qualifications

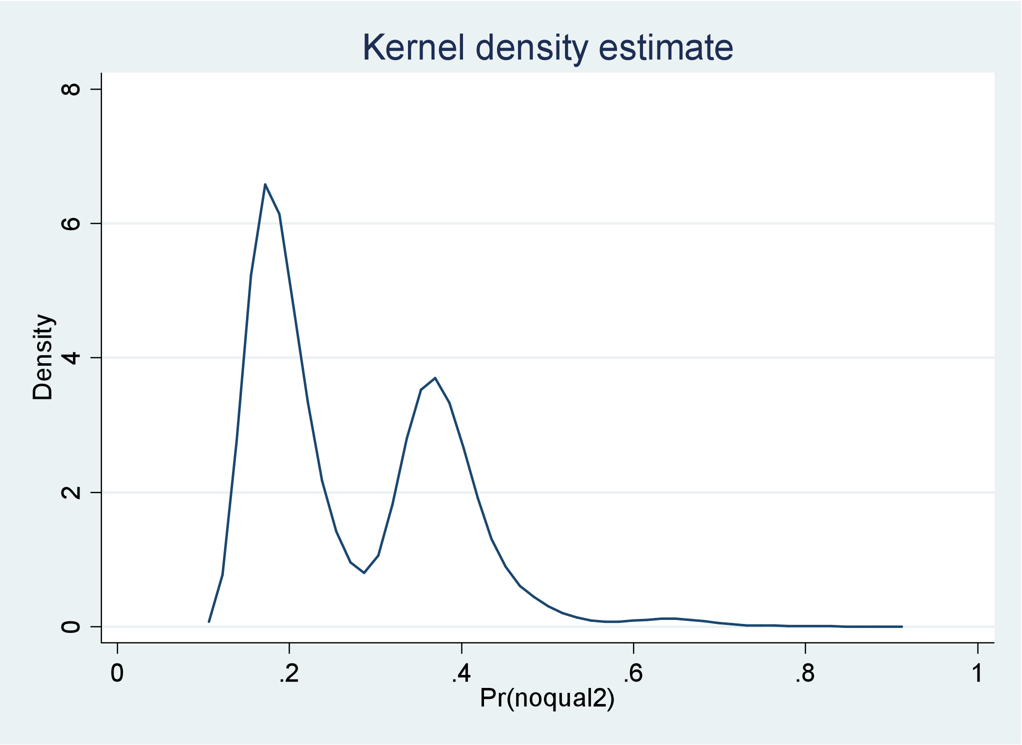
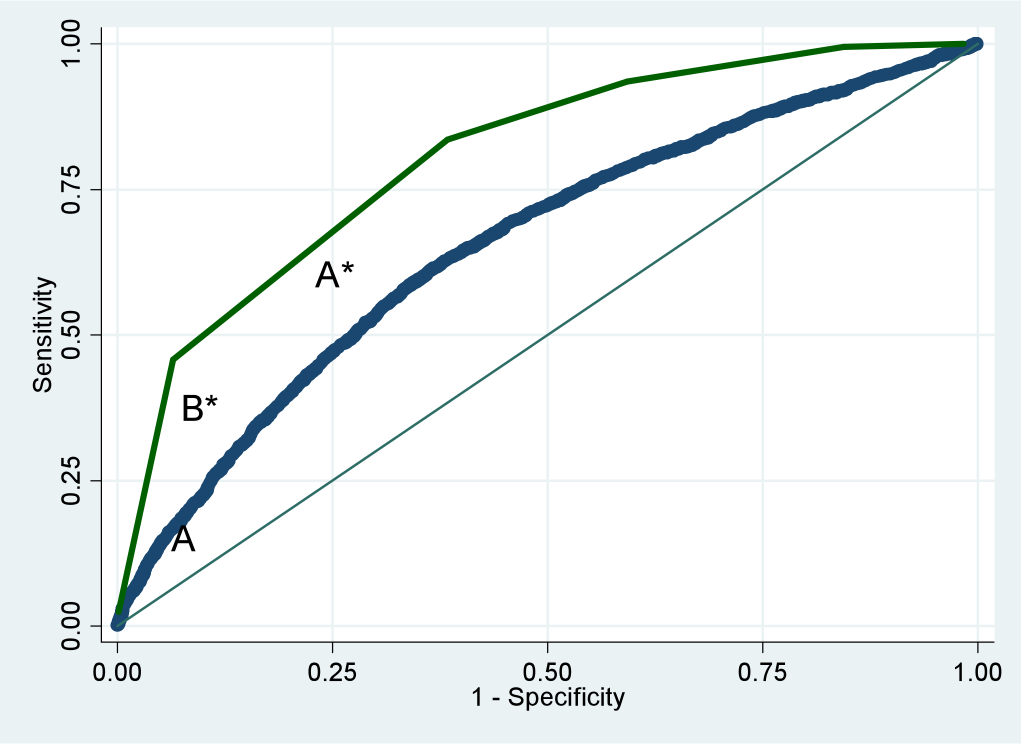


Fig. 5 Two ROC Curves


Sensitivity
A*
B*
A
1 - Specificity