Tue Tjur

Coefficients of determination in logistic regression models — a new proposal: The coefficient of discrimination.

Summary.

Many analogues to the coefficient of determination R^2 in ordinary regression models have been proposed in the context of logistic regression. Our starting point is a study of three definitions related to quadratic measures of variation. We discuss the properties of these statistics, and show that the family can be extended in a natural way by a fourth statistic with an even simpler interpretation, namely the difference between the averages of fitted values for successes and failures, respectively. We propose the name "the coefficient of discrimination" for this statistic, and recommend its use as a standard measure of explanatory power. In its intuitive interpretation, this quantity has no immediate relation to the classical versions of R^2, but it turns out to be related to these by two exact relations, which imply that all these statistics are asymptotically equivalent.

0. Introduction.

Consider an ordinary linear regression model, describing the observations $y_1, \ldots, y_n$ as independent normal with common variance σ^2 and expectations

$$\mathrm{E}y_i = \mu_i = \alpha + \beta x_i + \ldots$$

where, here and in the following, "$\alpha + \beta x_i + \ldots$" is meant to indicate any linear model specification. Let $\hat{\mu}_i = \hat{\alpha} + \hat{\beta} x_i + \ldots$ denote the fitted values, and consider the following sums of squares of deviations,

$\mathrm{SSD}_{\mathrm{res}} = \sum (y_i - \hat{\mu}_i)^2$ (the residual square sum)

$\mathrm{SSD}_{\mathrm{mod}} = \sum (\hat{\mu}_i - \bar{y})^2$ (the model square sum)

$\mathrm{SSD}_{\mathrm{tot}} = \sum (y_i - \bar{y})^2$ (the total square sum)

Provided that the model has a constant term (like α above), the average $\bar{y}$ of the observations equals the average $\bar{\hat{\mu}}$ of the fitted values, which means that $\mathrm{SSD}_{\mathrm{mod}}$ can also be interpreted as the unnormalized empirical variance of the fitted values. The presence of a constant term will be assumed throughout the paper. For models without a constant term it does not make much sense to talk about explanatory power in the R^2 sense, because the whole idea is to compare the model with the trivial model that has only the constant term.

Due to this assumption and the orthogonality of the residual vector and the vector of fitted values, we have the formula

$$\mathrm{SSD}_{\mathrm{tot}} = \mathrm{SSD}_{\mathrm{mod}} + \mathrm{SSD}_{\mathrm{res}}$$

which can be interpreted as a decomposition of the total "variation" of the observations in two components, the variation explained by the model and the residual variation. Accordingly, the *coefficient of determination*

$$R^2 = \frac{\mathrm{SSD}_{\mathrm{mod}}}{\mathrm{SSD}_{\mathrm{tot}}} = 1 - \frac{\mathrm{SSD}_{\mathrm{res}}}{\mathrm{SSD}_{\mathrm{tot}}}$$

is interpreted as the fraction (or percentage, if multiplied by 100) of the total variation which is explained by the model.

Before we proceed with the generalization to logistic regression models, it is important to understand what this quantity really is. Though functionally related to the F–statistic for the overall test for "no effects at all", it should certainly not be regarded as a test statistic, since it is usually computed *after* reduction of the model by removal of insignificant terms. In statistical practice, it is just an exploratory statistic that we compute, and then we clap our hands if it is close to 1. The question is *why* we clap our hands. It is certainly not because a model with a low value of R^2 is necessarily a bad model. We can have a perfectly respectable model with a low R^2, if the design of the experiment is such that the values of the explanatory variables are kept in a (too) narrow domain.

The best interpretation of R^2 is probably as an *estimate of a parameter function.* For large values of n and under suitable asymptotic conditions, implying that the fitted values $\hat{\mu}_i$ are close to the true expectations μ_i, we have approximately

$$\frac{1}{n}\mathrm{SSD}_{\mathrm{mod}} \approx \frac{1}{n}\sum(\mu_i - \bar{\mu})^2$$

and

$$\frac{1}{n}\mathrm{SSD}_{\mathrm{res}} \approx \frac{1}{n}\sum(y_i - \mu_i)^2 \approx \sigma^2\,.$$

It follows that

$$R^2 \approx 1 - \frac{\sigma^2}{\frac{1}{n}\sum(\mu_i - \bar{\mu})^2 + \sigma^2}\,.$$

In the last fraction, the denominator can be interpreted as the variance of a randomly chosen observation y_i. Thus, $1 - R^2$ can be regarded as an estimate of the proportion between two error variances, namely the variance when we predict observations by their expectations μ_i under the model, and the variance when we predict them by their common expectation $\bar{\mu}$ as randomly chosen observations among the n we have. In this sense, R^2 measures the model's *ability to predict,* relatively to the trivial model which assumes that the observations are i.i.d.

From this it should be clear that the interpretation of R^2 is easier when the set of experimental units can be regarded as a random sample from some population, perhaps even related to a multivariate normal distribution of the covariates. For designed experiments, R^2 is less relevant — or at least its interpretation is entirely different — because it depends so strongly on the design.

A final remark about the interpretation is that R^2 (as indicated by its name) equals the square of the empirical correlation between fitted values and observations,

$$R^2 = \left(\frac{\sum (y_i - \bar{y})(\hat{\mu}_i - \bar{y})}{\sqrt{\sum (y_i - \bar{y})^2 \sum (\hat{\mu}_i - \bar{y})^2}} \right)^2 .$$

1. Three definitions of the coefficient of determination in a logistic regression model.

Prediction in a logistic regression model can mean two different things, prediction of a single binary outcome and prediction of a relative frequency of successes in a covariate group (i.e. a group of observations with the same pattern of covariate values). The criticism of R^2 as a measure of explanatory power in logistic regression models put forward by Cox and Wermuth (1992) is obviously related to the last interpretation. Here, we are only interested in the first. For this reason, we consider only models for proper binary data, without aggregation to binomial counts. In this framework, a good prediction of an observation is only possible when the success probability is close to 1 or 0.

The justification of this choice is that it is in accordance with the usual interpretation of R^2 for linear models, which is related to the model's ability to predict a single observation. In linear models for normal observations, the equivalent of forming relative frequencies for covariate groups would be the replacement of the original observations with averages over covariate groups. This is a wellknown way of improving the accuracy and reducing the data set, but it is usually not considered a legitimite way of increasing R^2.

Let $y_i \in \{0, 1\}$ (1 for "success", 0 for "failure") denote the binary, independent responses. The model is given by

$$P(y_i = 1) = \mathrm{E} y_i = \pi_i = \frac{\exp(\alpha + \beta x_i + \dots)}{1 + \exp(\alpha + \beta x_i + \dots)} .$$

Let $\hat{\pi}_i$ denote the ML–estimates of the success probabilities, or the fitted values. An important remark here is that the average $\bar{\hat{\pi}}$ of the fitted values is equal to the average $\bar{y}$ of the binary responses, the relative frequency of successes in the data set. This identity is just one of the likelihood equations, namely the equation one gets by differentiation of the log likelihood with respect to the intercept α. This is an exclusive property of the logit models, shared by other exponential family classes like the linear normal models and the log–linear Poisson models, but not by models for binary data specified by other link functions than logit.

As in the linear case, we define

$\mathrm{SSD}_{\mathrm{res}} = \sum (y_i - \hat{\pi}_i)^2$

$\mathrm{SSD}_{\mathrm{mod}} = \sum (\hat{\pi}_i - \bar{y})^2$

$\mathrm{SSD}_{\mathrm{tot}} = \sum (y_i - \bar{y})^2$

A property of the normal linear models which is *not* shared by the logistic regression models is the orthogonality of the vector of fitted values and the vector of residuals. Accordingly, $\mathrm{SSD}_{\mathrm{tot}}$ is not the sum of $\mathrm{SSD}_{\mathrm{mod}}$ and $\mathrm{SSD}_{\mathrm{res}}$, which leaves us with two obvious candidates for the title R^2,

$$R^2_{\mathrm{mod}} = \frac{\mathrm{SSD}_{\mathrm{mod}}}{\mathrm{SSD}_{\mathrm{tot}}}$$

and

$$R^2_{\mathrm{res}} = 1 - \frac{\mathrm{SSD}_{\mathrm{res}}}{\mathrm{SSD}_{\mathrm{tot}}} \,.$$

Moreover, the interpretation of the classical R^2 as the squared empirical correlation between observations and fitted values suggests the definition

$$R^2_{\mathrm{cor}} = \left(\frac{\sum (y_i - \bar{y})(\hat{\pi}_i - \bar{y})}{\sqrt{\sum (y_i - \bar{y})^2) \sum (\hat{\pi}_i - \bar{y})^2}} \right)^2 \,.$$

These definitions are among those studied by Kvålseth (1985), and they are also mentioned in the review article by Menard (2000), as alternative expressions for essentially the same quantity. As we shall see later, there is a complicated exact relation between these three quantities. But let us first study their properties.

An important observation is that they are approximately equal, as estimates of one and the same parameter function, in the following sense. Assume that the probability estimates $\hat{\pi}_i$ are close to the true probabilites π_i, and that the sums over $i = 1, \ldots, n$ occurring in the following can, with little relative error, be replaced with the corresponding sums of expectations. The approximate results derived under these conditions can obviously be translated to proper asymptotic results, at least under the restrictive condition that the observations belong to finitely many covariate groups that grow proportionally as $n \to \infty$. In this case we know that the maximum likelihood estimates of the parameters will converge to the true parameter values, and the replacements we are going to perform of certain sums by the corresponding sums of expectations are asymptotically valid by the law of large numbers.

Under these assumptions we have

$$\mathrm{SSD}_{\mathrm{mod}} \approx \sum (\pi_i - \bar{\pi})^2 ,$$

$$\mathrm{SSD}_{\mathrm{res}} \approx \sum (y_i - \pi_i)^2 \approx \sum \pi_i (1 - \pi_i)$$

and

$$\begin{aligned} \mathrm{SSD}_{\mathrm{tot}} &= \sum (y_i - \bar{y})^2 = \sum y_i^2 - \frac{1}{n} \left(\sum y_i \right)^2 \\ &= n\bar{y} - n\bar{y}^2 = n\bar{y}(1 - \bar{y}) \approx n\bar{\pi}(1 - \bar{\pi}) . \end{aligned}$$

The following computations show that these asymptotic expressions for the basic square sums do actually satisfy the relation corresponding to the exact orthogonality relation in the linear case,

$$\mathrm{SSD}_{\mathrm{mod}} + \mathrm{SSD}_{\mathrm{res}} \approx \sum (\pi_i - \bar{\pi})^2 + \sum \pi_i (1 - \pi_i)$$

$$= \sum \left(\pi_i^2 + \bar{\pi}^2 - 2\pi_i \bar{\pi} + \pi_i - \pi_i^2 \right) = \sum \left(\bar{\pi}^2 - 2\pi_i \bar{\pi} + \pi_i \right)$$

$$= n\bar{\pi}^2 - 2n\bar{\pi}^2 + n\bar{\pi} = n\bar{\pi}(1 - \bar{\pi}) \approx \mathrm{SSD}_{\mathrm{tot}} .$$

It follows that

$$R^2_{\mathrm{mod}} \approx \frac{\sum (\pi_i - \bar{\pi})^2}{n\bar{\pi}\,(1 - \bar{\pi})} = 1 - \frac{\sum \pi_i (1 - \pi_i)}{n\bar{\pi}\,(1 - \bar{\pi})} \approx R^2_{\mathrm{res}} .$$

Hence, R^2_{mod} and R^2_{res} can be regarded as estimates of the same parameter function.

For the third statistic R^2_{cor}, we have

$$R^2_{\mathrm{cor}} = \frac{\left(\sum (y_i - \bar{y})(\hat{\pi}_i - \bar{y}) \right)^2}{\sum (y_i - \bar{y})^2 \sum (\hat{\pi}_i - \bar{y})^2} = \frac{\left(\sum (y_i - \hat{\pi}_i + \hat{\pi}_i - \bar{y})(\hat{\pi}_i - \bar{y}) \right)^2}{\sum (y_i - \bar{y})^2 \sum (\hat{\pi}_i - \bar{y})^2}$$

$$= \frac{\left(\sum (y_i - \hat{\pi}_i)(\hat{\pi}_i - \bar{y}) + \sum (\hat{\pi}_i - \bar{y})^2 \right)^2}{\sum (y_i - \bar{y})^2 \sum (\hat{\pi}_i - \bar{y})^2}$$

$$\approx \frac{\left(0 + \sum (\hat{\pi}_i - \bar{y})^2 \right)^2}{\sum (y_i - \bar{y})^2 \sum (\hat{\pi}_i - \bar{y})^2} = \frac{\mathrm{SSD}^2_{\mathrm{mod}}}{\mathrm{SSD}_{\mathrm{tot}} \mathrm{SSD}_{\mathrm{mod}}} = \frac{\mathrm{SSD}_{\mathrm{mod}}}{\mathrm{SSD}_{\mathrm{tot}}} = R^2_{\mathrm{mod}} .$$

This means that R^2_{cor} and R^2_{mod} are asymptotically equivalent, and in particular that R^2_{cor} can be regarded as an estimate of the same parameter function as the two others.

Simulation studies with the π_i's spread more or less uniformly over the unit interval show that the properties of these three statistics are very similar. They have almost the same mean and the same standard deviation. As estimates, they are upwards biased. The bias is small compared to the standard deviation when the number of covariates is small, but increases with the number of parameters. However, what is most important to us here is that the pairwise differences between them are small, with a standard deviation of no more than around 20% of their individual standard deviations. This holds rather generally, for situations with $n = 100$–10,000 and 2–50 covariates. In particular, R^2_{res} and R^2_{cor} are typically very close to each other. As we shall see later, there is a good reason for this.

Thus, from a practical point of view it does not matter much whether we choose R^2_{res}, R^2_{mod} or R^2_{cor} as our measure of explanatory power. We can freely choose the one with the simplest interpretation as a function of data and fitted values. Personally I would prefer R^2_{mod} if I had to choose, because this is the maximum likelihood estimate of the parameter function it estimates.

But obviously, we must also take a look at the theoretical properties of these quantities. The least we can expect from a reasonable coefficient of determination is that it takes its values between 0 and 1, and that the extreme values 0 and 1 correspond to the properties "no explanatory power at all" (all fitted values are equal) and "perfect fit" (the fitted values coincide with the observations).

The last property requires an explanation. Formally, finite values of the parameters α, β, $\ldots$ can never result in the exact values 1 or 0 of the success probabilities. What we mean when we say "perfect fit" or "$y_i = \hat{\pi}_i$ for all i" is, of course, that the likelihood function takes its maximal value on the boundary — or more correctly, it does not take a maximal value at all, but converges to a value greater than all others when the parameter vector tends to infinity in a certain direction. More precisely, the condition is that there exists a linear combination $a + bx_i + \ldots$ of the covariates which *separates* successes and failures, in the sense that it is positive for $y_i = 1$ and negative for $y_i = 0$. In that case, taking $\alpha_N = Na$, $\beta_N = Nb$ etc., will produce a sequence of parameter values such that the likelihood converges to its supremum as $N \to \infty$ and the condition $y_i = \hat{\pi}_i$ for perfect fit is met in the limit. This is what we mean when we write $y_i = \hat{\pi}_i$. We are really talking about the degenerate situation which, in practice, is recognized by the following characteristics. The Newton–Raphson iterations tend to continue forever unless they are stopped by some "time out" limit. All or some of the parameter estimates obtained in the last iteration are unrealistically large in absolute value (typically corresponding to values of $\hat{\alpha} + \hat{\beta}x_i + \ldots$ around ± 40, which is roughly the same as $\pm\infty$ on the logit scale), and so are their reported approximate standard deviations (which, in turn, implies that the Wald tests for removal of terms from the model are useless). The model has failed in the sense that it degenerates to a deterministic model. But if we compute the fitted values, they turn out to coincide exactly (in principle only almost exactly) with the observations.

Notice that if all $\hat{\pi}_i$ are 0 or 1, then we must also have $y_i = \hat{\pi}_i$. Or, to be more precise, if a sequence $(\alpha^{(N)}, \beta^{(N)}, \dots)$ of parameter vectors can be constructed in such a way that the corresponding values of the likelihood function converge to its supremum, and all the corresponding sequences of success–probabilities $\pi_i^{(N)}$ converge to either 0 or 1, then we can not have the limit 1 for an observation which is 0 or vice versa. This is easily seen if we take a closer look at the expression for the likelihood and notice that it will always converge to zero when at least one π_i converges to $1 - y_i$. Hence it makes good sense to state the condition for "perfect fit" simply as "$y_i = \hat{\pi}_i$ for all i".

The following result shows that our proposed measures of explanatory power have — with a single exception — the properties one would expect by analogy with the properties of R^2 for linear models.

Proposition 1.

R^2_{mod} and R^2_{cor} are ≥ 0, with equality if and only if all $\hat{\pi}_i$ are equal.

R^2_{mod}, R^2_{res} and R^2_{cor} are ≤ 1, with equality if and only if $y_i = \hat{\pi}_i$ for all i.

The proof can be found in section 5.

Notice that the proposition does *not* claim that $R^2_{\text{res}} \geq 0$. It is actually possible to construct examples where R^2_{res} is negative. Take a model with a single covariate x with values $x_1 = 2$, $x_2 = 1$ and $x_3 = \cdots = x_n = 0$, and suppose we have observed $y_1 = 1$, $y_2 = 0$ and $y_3 = \cdots = y_n = 1$. For $n \geq 13$, R^2_{res} becomes negative, and for large n it seems to approach -0.25. I have no idea whether -0.25 is a lower limit for this phenomena. But the fact that R^2_{res} can come out negative — even though it seems to happen only in such degenerate cases — is more than enough to disqualify it as a candidate for the title "best choice of R^2 in logistic regression".

2. A fourth definition, the coefficient of discrimination.

Among the many exploratory methods for evaluation of a logistic regression model, my favourite is the following. Draw two "parallel histograms", i.e. two histograms, one over the other, on the same scale (in this case the unit interval) and with the same number of intervals (in this case usually 10, for very large n perhaps 20). The distributions summarized by the two histograms are, respectively, the distribution of the fitted values for the failures and the distribution of the fitted values for the successes. See figure 1 in section 3 for a concrete example. This simple figure contains a lot of information. For example, an easy way of detecting serious violations of the model assumptions (in particular misspecification of the link function or the need for transformation of covariates) is to check for inconsistencies between the heights of corresponding bars in the two histograms. If the model is correct we would, for example, expect the proportion between the counts of successes and failures with fitted values between 0.3 and 0.4 to be somewhere around 0.35/0.65, because each of the observations in play here is supposed to be a success with a probability between 0.3 and 0.4. This graphical check of the model can be regarded as an explorative version of Hosmer–Lemeshow's test (Hosmer et al. 2000), which — in spite of the criticism put forward by Hosmer et al. (1997) — is about the closest one can come in the strictly binary case to a standard analogue of the usual goodness–of–fit test for data that are aggregated to binomial counts in relatively large covariate groups. The ROC–curve (see again e.g. Hosmer et al. 2000) is also related to this figure, namely as the curve $(1 - F_0(\pi), 1 - F_1(\pi))$, $\pi \in (0, 1)$, where F_0 and F_1 are the cdf's of the empirical distributions underlying the two histograms. But in my opinion, the two histograms are much easier to interpret than the ROC–curve.

Also, the commonly reported two–by–two contingency tables of "predicted versus observed" are closely related to the two histograms. The idea behind these tables is that a good model is a model that predicts well, in the sense that almost all observations with fitted values higher than, say, 0.5 are successes, whereas almost all observations with fitted values lower than 0.5 are failures. In other words, a good model is characterized by the property that if we "predict" each observation by 1 or 0 according to whether its estimated probability of success is greater or smaller than a given threshold value, then we get a high percentage of "hits", with only a few "false positives" and "false negatives". The two histograms summarize in the simplest possible way a number of such tables, one for each cutpoint, and actually does it in a much more convenient way than a listing of the tables would do. For example, by looking at the histograms it is usually easy to get an impression of where one should place the threshold in order to minimize the total number of false predictions, if this is what we want.

Now, let us for a moment forget all about the R^2–statistics of the previous section. They were based on ideas related to variance and quadratic variation, which — as also noticed by Menard (2000) — are somewhat strange concepts in a universe of binary observations. And let us, instead, take the above mentioned characterization of "a good model" as our definition of a high explanatory power. In order to boil this vague concept down to a single quantity, we must invent a statistic that measures the degree to which the upper histogram has most of its mass concentrated close to the left endpoint of the unit interval and the lower histogram has most of its mass concentrated close to the right endpoint. A simple, almost canonical, choice here is the difference between the expectations in the two distributions, i.e.

$$D = \bar{\hat{\pi}}_1 - \bar{\hat{\pi}}_0 \,,$$

where $\bar{\hat{\pi}}_1$ and $\bar{\hat{\pi}}_0$ denote the averages of fitted values for successes and failures, respectively. As a name for this quantity I propose *the coefficient of discrimination.* This name was used by Raveh (1986) for a very similar quantity (though in a somewhat different context), and it seems to describe very well what the quantity actually measures, namely the model's ability to discriminate between successes and failures.

Conceptually, D has not much in common with the R^2 measures discussed in the previous section. But as a matter of fact, we have the following result.

Proposition 2.

We have the exact relations

$$D = \frac{1}{2}\left(R^2_{\text{mod}} + R^2_{\text{res}}\right)$$

and

$$D = \sqrt{R^2_{\text{mod}} R^2_{\text{cor}}}$$

In words, D equals the arithmetic average of R^2_{mod} and R^2_{res} and the geometric average of R^2_{mod} and R^2_{cor}.

The proof is given in section 5.

An immediate consequence of this is that D is asymptotically equivalent to the R^2 measures of section 1. Moreover, combining the second relation of proposition 2 with the properties of R^2_{mod} and R^2_{cor} stated in proposition 1, we get the following nice result.

Corollary.

$D \geq 0$, with equality if and only if all $\hat{\pi}_i$ are equal.

$D \leq 1$, with equality if and only if $y_i = \hat{\pi}_i$ for all i.

Another consequence of proposition 2 is the following technicality. Implicitely, the proposition implies a relation between the three quantities R^2_{mod}, R^2_{res} and R^2_{cor}, which can be rewritten as

$$1 + \frac{1}{2}\left(\frac{R^2_{\text{res}}}{R^2_{\text{mod}}} - 1\right) = \sqrt{1 + \left(\frac{R^2_{\text{cor}}}{R^2_{\text{mod}}} - 1\right)} \,.$$

Now, consider an asymptotic situation where all the R^2 versions are close to each other. Think of the right hand side of the above identity as the value of the function $\sqrt{1+(x-1)}$ evaluated at the point $x = \frac{R^2_{\text{cor}}}{R^2_{\text{mod}}}$, which is pretty close to 1. If we replace this function by its first order Taylor expansion around 1, we get the appoximate identity

$$1 + \frac{1}{2}\left(\frac{R^2_{\text{res}}}{R^2_{\text{mod}}} - 1\right) \approx 1 + \frac{1}{2}\left(\frac{R^2_{\text{cor}}}{R^2_{\text{mod}}} - 1\right)$$

which is valid up to an error term that is known to be small of second order in $\left(\frac{R^2_{\text{cor}}}{R^2_{\text{mod}}} - 1\right)$. Thus, the absolute value of the difference between left and right side in this approximate relation is dominated by something of the form const. × $\left(\frac{R^2_{\text{cor}}}{R^2_{\text{mod}}} - 1\right)^2$. A few further rearrangements of terms, which we leave to the reader, will show that the difference $R^2_{\text{cor}} - R^2_{\text{res}}$ is of the same order of magnitude as the *square* of $R^2_{\text{cor}} - R^2_{\text{mod}}$. This explains why R^2_{res} and R^2_{cor} are (in our simulation studies, and in practice when data sets are large enough) so much closer to each other than to R^2_{mod}.

3. An example.

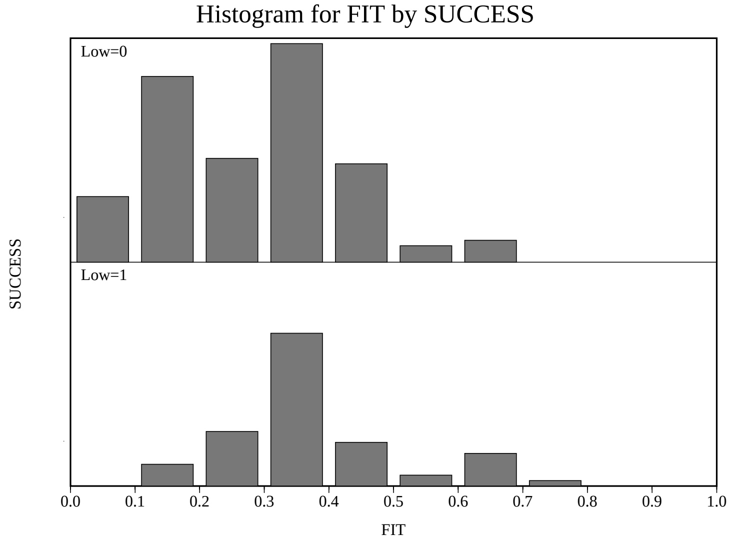

Figure 1.

As a data set for illustration I have chosen the Baystate Medical Center data on low birth weights which has been used for the same purpose by Hosmer et al. (2000, page 25 and 26), and also by Hosmer et al. (1997) and Zheng and Agresti (2000). It contains information about 189 births, with the response `LOW` (indicator of low birth weight) as the binary response, to be explained by various demographic and clinical variables. For further compatibility, the model considered here is identical to the one considered by Hosmer et al. (1997), which includes most of the variables that are really significant plus the mother's age (which is insignificant, but known from other studies to be important). In standard model formula syntax, our model is

```
LOW = CONST. + AGE + LWT + RACE + SMOKE
```

where `AGE` is the mother's age, `LWT` is her weight before pregnancy, `RACE` is ethnicity coded as a factor on three levels and `SMOKE` is an indicator of (the mother being) a smoker.

The two parallel histograms of fitted values for failures and succeses are shown in figure 1. The proportions between bar heights in the two histograms seem to behave roughly as they should. For example, the two bars over the interval from 0.2 to 0.3 represent 19 "failures" and 10 "successes", which does not deviate significantly from what one would expect for 29 binary outcomes with success probabilities between 0.2 and 0.3. But as we can see, there is not much explanatory power in this model. The two histograms have a considerable overlap. The most positive we can say is that it is possible to isolate a small low–risk group, consisting of the women with $\hat{\pi} < 0.2$. This group consists of 50 women, of which only 4 gave birth to a child with critically low weight. A similar high–risk group of substantial size can not be identified (only one fitted value exceeds 0.7) Thus, even if the conclusions of this study may be scientifically interesting, they are not likely to be of much help in clinical practice. Accordingly, the summary measures of explanatory power are small,

R^2_{mod}	R^2_{res}	R^2_{cor}	D
0.101766	0.090697	0.090998	0.096231 .

4. Conclusion.

It is impossible to hide any longer that I consider the quantity D, the coefficient of discrimination, a highly recommendable R^2–substitute for logistic regression models. The corollary shows that D shares important properties with some of the more classical R^2 statistics discussed in section 1. According to proposition 2, we may even think of D as a compromise between the two quadratic measures R^2_{mod} and R^2_{cor} that were left after the disqualification of R^2_{res} (see remark following proposition 1). This relation to the classical R^2 versions is an obvious advantage if we want to assign a similar meaning to concrete values of D as to the same values of R^2 in ordinary linear regression models. In addition to this comes that D can be explained in terms and concepts that are directly related to models for binary observations on their own premisses, without any reference to strange concepts like prediction variance and quadratic variation. D is probably the measure of explanatory power that comes closest to the satisfaction of the eight ideal requirements set up by Kvålseth (1985).

A small reservation is required here. Neither D, nor any of the three R^2 versions discussed in section 1, have the property that they will automatically increase when the model is extended by an additional covariate. Usually they will, of course. A few (not too systematic) simulation studies for selected examples indicate that this happens in more than 99.9 of the cases, even when the added covariate has no effect at all. But since maximum likelihood in logistic regression is not equivalent to minimization of the residual sum of squares, nor maximization of the model sum of squares or the difference $\bar{\hat{\pi}}_1 - \bar{\hat{\pi}}_0$, it is no surprise that these statistics can "move the wrong way" when a model is extended or reduced. If this property is considered overwhelmingly important (as it seems to be e.g. for Cameron and Windmeijer 1997), a likelihood–based R^2–substitute should probably be preferred. Many such proposals have been made in the literature, but if we restrict our attention to proper *generalizations* of the usual R^2 from linear models, the one and only choice seems to be the statistic proposed by Cox and Snell (1989),

$$1 - \exp\left(-\frac{2}{n}(\hat{l} - \hat{l}_0)\right) ,$$

where $\hat{l}$ and $\hat{l}_0$ are the maxima of the log likelihood for the model in question and the model with only a constant term, respectively. The justification of this is that this is exactly the formula that expresses the usual R^2 for normal linear models in terms of the log–likelihood. Thus, it is not unreasonable at all to expect a nice behaviour of this canonical generalization of R^2. But unfortunately, it turns out that it never takes values close to 1. Its maximal value $1 - \exp(2\hat{l}_0/n)$, which is taken when $y_i = \hat{\pi}_i$ for all i, never exceeds 0.75. As noticed by Nagelkerke (1991), this can be repaired by a simple renormalization; but after this, the idea has certainly lost a bit of its intuitive appeal.

5. Proofs

Proof of proposition 1.

It is obvious that $R^2_{\text{mod}} \geq 0$, with equality if and only if all $\hat{\pi}_i$ are equal, and that $R^2_{\text{res}} \leq 1$ with equality if and only if $y_i = \hat{\pi}_i$ for all i. $R^2_{\text{mod}} \leq 1$ is equivalent to

$$\sum(\hat{\pi}_i - \bar{y})^2 \leq \sum(y_i - \bar{y})^2$$

or

$$\sum \hat{\pi}_i^2 - n\bar{y}^2 \leq \sum y_i^2 - n\bar{y}^2$$

or

$$\sum \hat{\pi}_i^2 \leq \sum y_i^2$$

or (since $\sum y_i^2 = \sum y_i = \sum \hat{\pi}_i$)

$$\sum \hat{\pi}_i^2 \leq \sum \hat{\pi}_i$$

or

$$\sum \hat{\pi}_i(1 - \hat{\pi}_i) \geq 0$$

which is trivially correct with equality if and only if all $\hat{\pi}_i$ are either 1 or 0. The inequalities $0 \leq R^2_{\text{cor}} \leq 1$ are immediate consequences of the fact that R^2_{cor} is a squared correlation coefficient. From another wellknown property of the correlation coefficient, we conclude that we have $R^2_{\text{cor}} = 1$ if and only if there is a linear relation of the form $\hat{\pi}_i = a + by_i$, and this is easily seen to imply the precise condition for perfect fit, or $\hat{\pi}_i = y_i$ for all i.

It remains to prove that $R^2_{\text{cor}} = 0$ if and only if all the $\hat{\pi}_i$ are equal. This is a trivial consequence of the following result, which we give the form of a lemma for later use.

Lemma 1. *We have the inequality*

$$\sum (y_i - \bar{y})(\hat{\pi}_i - \bar{y}) \geq 0$$

with equality if and only if $\hat{\pi}_i = \bar{y}$ *for all* i.

Proof. We have

$$\sum (y_i - \bar{y})(\hat{\pi}_i - \bar{y}) = \sum y_i(\hat{\pi}_i - \bar{y}) = \sum y_i \hat{\pi}_i - \sum y_i \bar{y}\,.$$

In order to make an indirect proof, assume that this quantity is strictly negative, or that

$$\sum y_i \hat{\pi}_i < \sum y_i \bar{y}\,.$$

Under this assumption we can show that the value of the log likelihood with the fitted values $\hat{\pi}_i$ inserted becomes *smaller* than the value it takes when all $\hat{\pi}_i$ are replaced with their average $\bar{\hat{\pi}}$. Since this is in contradiction with the fact that the values $\hat{\pi}_i$ maximize the log–likelihood, the assumption must be wrong. The detailed argument goes as follows.

By a straightforward computation, our assumption implies that we also have the "complementary" inequality

$$\sum (1 - y_i)(1 - \hat{\pi}_i) < \sum (1 - y_i)(1 - \bar{y})\,,$$

and these two inequalities can conveniently be written on the short form

$$\bar{\hat{\pi}}_1 < \bar{y} < \bar{\hat{\pi}}_0\,,$$

where $\bar{\hat{\pi}}_1$ and $\bar{\hat{\pi}}_0$ denote the averages of the fitted values among successes and failures, respectively.

Now, by Jensen's inequality and the concavity of the log curve,

$$\sum y_i \log \hat{\pi}_i \leq \Big(\sum y_i\Big) \log \bar{\hat{\pi}}_1$$

and

$$\sum (1-y_i) \log(1-\hat{\pi}_i) \leq \Big(\sum (1-y_i)\Big) \log(1-\bar{\hat{\pi}}_0)$$

Since log is strictly increasing, we have further

$$\Big(\sum y_i\Big) \log \bar{\hat{\pi}}_1 < \Big(\sum y_i\Big) \log \bar{\hat{\pi}}$$

and

$$\Big(\sum (1-y_i)\Big) \log(1-\bar{\hat{\pi}}_0) < \Big(\sum (1-y_i)\Big) \log(1-\bar{\hat{\pi}})\,.$$

Combining the two chains of inequalities and adding them, we get

$$\sum y_i \log \hat{\pi}_i + \sum (1-y_i) \log(1-\hat{\pi}_i) < \sum y_i \log \bar{\hat{\pi}} + \sum (1-y_i) \log(1-\bar{\hat{\pi}})$$

which expresses that the log likelihood can be strictly increased by replacement of the maximum likelihood estimates with parameter values for which only the intercept parameter is nonzero. Thus we have reached a contradiction, and we must conclude that the inequality of the proposition is always satisfied.

It is obvious that if all $\hat{\pi}_i$ are equal (and thereby equal to $\bar{\hat{\pi}} = \bar{y}$), then the inequality of the lemma becomes an equality. Conversely, suppose that this inequality degenerates to an equality. In order to prove that $\hat{\pi}_i = \bar{y}$ for all i, we can reuse the arguments of the first part of the proof in a weakened form, where the assumption $\bar{\hat{\pi}}_1 < \bar{y} < \bar{\hat{\pi}}_0$ is replaced with $\bar{\hat{\pi}}_1 \leq \bar{y} \leq \bar{\hat{\pi}}_0$. Just like in the first part of the proof, we can show that

$$\sum y_i \log \hat{\pi}_i \leq \Big(\sum y_i\Big) \log \bar{\hat{\pi}}_1 \leq \Big(\sum y_i\Big) \log \bar{\hat{\pi}}\,.$$

The first inequality here follows from Jensen's inequality and the concavity of the log curve, the second from the fact that the log curve is increasing. The only difference is that none of the inequalities are sharp here. In the same way, we can derive the "complementary" inequalities, where y_i is replaced with $1 - y_i$ and $\hat{\pi}_i$ with $1 - \hat{\pi}_i$. Adding these two chains we end up with the same old inequality between the two values of the log–likelihood, the only difference being that this time it is not sharp. Now, the important observation is that if as much as a single one of the four inequalites in the two chains is sharp, then the relation between the two values of the log–likelihood becomes sharp too. And since this is not possible, we *must* have equality all the way through the two chains. If the inequality following from Jensen's inequality is an equality, we must (since log is strictly concave) have that all the $\hat{\pi}_i$ for successes are equal. Similarly, the similar equality in the complementary chain means that all the $\hat{\pi}_i$ for the failures are equal. And if the inequalities based on monotonicity of the log–curve are equalities, we must have $\bar{\hat{\pi}}_1 = \bar{\hat{\pi}}$ and $\bar{\hat{\pi}}_0 = \bar{\hat{\pi}}$, since log is strictly increasing. Obviously, the only possibility left is that all $\hat{\pi}_i$ are equal.

Proof of proposition 2.
Define

$$\mathrm{SPD} = \sum (y_i - \hat{\pi}_i)(\hat{\pi}_i - \bar{y})\,.$$

Then,

$$\mathrm{SSD}_{\mathrm{tot}} = \mathrm{SSD}_{\mathrm{res}} + \mathrm{SSD}_{\mathrm{mod}} + 2\mathrm{SPD}$$

and from lemma 1 it follows that

$$\mathrm{SPD} + \mathrm{SSD}_{\mathrm{mod}} = \sum (y_i - \bar{y})(\hat{\pi}_i - \bar{y}) \geq 0\,.$$

Now,

$$\begin{aligned}
\frac{1}{2}\left(R^2_{\mathrm{mod}} + R^2_{\mathrm{res}}\right) &= \frac{1}{2}\left(\frac{\mathrm{SSD}_{\mathrm{mod}}}{\mathrm{SSD}_{\mathrm{tot}}} + 1 - \frac{\mathrm{SSD}_{\mathrm{res}}}{\mathrm{SSD}_{\mathrm{tot}}}\right) \\
&= \frac{\mathrm{SSD}_{\mathrm{mod}} + \mathrm{SSD}_{\mathrm{tot}} - \mathrm{SSD}_{\mathrm{res}}}{2\mathrm{SSD}_{\mathrm{tot}}} \\
&= \frac{\mathrm{SSD}_{\mathrm{mod}} + \mathrm{SSD}_{\mathrm{res}} + \mathrm{SSD}_{\mathrm{mod}} + 2\mathrm{SPD} - \mathrm{SSD}_{\mathrm{res}}}{2\mathrm{SSD}_{\mathrm{tot}}} \\
&= \frac{2\mathrm{SSD}_{\mathrm{mod}} + 2\mathrm{SPD}}{2\mathrm{SSD}_{\mathrm{tot}}} = \frac{\mathrm{SSD}_{\mathrm{mod}} + \mathrm{SPD}}{\mathrm{SSD}_{\mathrm{tot}}} = \frac{\sum (y_i - \bar{y})(\hat{\pi}_i - \bar{y})}{\mathrm{SSD}_{\mathrm{tot}}} \\
&= \frac{\sum (y_i - \bar{y})(\hat{\pi}_i - \bar{y})}{\sqrt{\mathrm{SSD}_{\mathrm{tot}}\mathrm{SSD}_{\mathrm{mod}}}} \sqrt{\frac{\mathrm{SSD}_{\mathrm{mod}}}{\mathrm{SSD}_{\mathrm{tot}}}} = \sqrt{R^2_{\mathrm{cor}} R^2_{\mathrm{mod}}}\,.
\end{aligned}$$

Thus, the two right hand sides in the proposition are equal. It remains to prove one of the two identities. The last one can be proved as follows.

$$\begin{aligned}
D &= \frac{\sum y_i \hat{\pi}_i}{\sum y_i} - \frac{\sum (1 - y_i)\hat{\pi}_i}{\sum (1 - y_i)} \\
&= \frac{\left(\sum y_i \hat{\pi}_i\right)\left(\sum (1 - y_i)\right) - \left(\sum (1 - y_i)\hat{\pi}_i\right)\left(\sum y_i\right)}{\left(\sum y_i\right)\left(\sum (1 - y_i)\right)} \\
&= \frac{n \sum y_i \hat{\pi}_i - \left(\sum y_i\right)\left(\sum y_i \hat{\pi}_i\right) - \left(\sum y_i\right)\left(\sum \hat{\pi}_i\right) + \left(\sum y_i\right)\left(\sum y_i \hat{\pi}_i\right)}{\left(\sum y_i\right)\left(\sum (1 - y_i)\right)} \\
&= \frac{n \sum (y_i - \bar{y})(\hat{\pi}_i - \bar{y})}{\left(\sum y_i\right)\left(n - \sum y_i\right)} = \frac{\mathrm{SPD} + \mathrm{SSD}_{\mathrm{mod}}}{\sum y_i^2 - \frac{1}{n}\left(\sum y_i\right)^2} = \frac{\mathrm{SPD} + \mathrm{SSD}_{\mathrm{mod}}}{\mathrm{SSD}_{\mathrm{tot}}}
\end{aligned}$$

which, according to the previous computations, equals $\sqrt{R^2_{\mathrm{cor}} R^2_{\mathrm{mod}}}$.

6. References

Cameron, A. C. and Windmeijer, F. A. G. (1997).
An R-squared measure of goodness of fit for som common nonlinear regression models.
Journal of Econometrics **77**, 329–342.

Cox, D. R. and Snell, E. J. (1989).
The Analysis of Binary Data (Second Edition).
Chapman and Hall.

Cox, D. R. and Wermuth, N. (1992).
A Comment on the Coefficient of Determination for Binary Responses.
The American Statistician **46**, 1–4.

Hosmer, D. W. and Lemeshow, S. (2000).
Applied Logistic Regression (Second Edition).
Wiley Series in Probability and Statistics.

Hosmer, D. W., Hosmer T., Le Cessie, S. and Lemeshow, S. (1997).
A Comparison of Goodness–of–fit Tests for The Logistic Regression Model.
Statistics in Medicine **16**, 965–980.

Kvålseth, T. O. (1985).
Note about R^2.
The American Statistician **39**, 279–285.

Menard, S. (2000)
Coefficients of Determination for Multiple Logistic Regression Analysis.
The American Statistician. **54**, 17–24.

Nagelkerke, N. J. D. (1991).
A note on a general definition of the coefficient of determination.
Biometrika **78**, 691–692.

Raveh, S. (1986)
On Measures of Monotone Association.
The American Statistician. **40**, 117–123.

Zheng, B. and Agresti, A. (2000).
Summarizing the predictive power of a generalized linear model.
Statistics in Medicine **19**, 1771–1781.

7. Acknowledgement. Thanks to the editor and an associate editor for critical comments, constructive proposals of improval and a very careful proof–reading.

Tue Tjur
Center for Statistics, Dept. of Finance,
Copenhagen Business School
Solbjerg Plads 3
2000 Frederiksberg
DENMARK
tuetjur@cbs.dk