Qual Quant (2009) 43:59–74
DOI 10.1007/s11135-007-9077-3

ORIGINAL PAPER

Linear versus logistic regression when the dependent variable is a dichotomy

Ottar Hellevik

Published online: 16 February 2007

Abstract The article argues against the popular belief that linear regression should not be used when the dependent variable is a dichotomy. The relevance of the statistical arguments against linear analyses, that the tests of significance are inappropriate and that one risk getting meaningless results, are disputed. Violating the homoscedasticity assumption seems to be of little practical importance, as an empirical comparison of results shows nearly identical outcomes for the two kinds of significance tests. When linear analysis of dichotomous dependent variables is seen as acceptable, there in many situations exist compelling arguments of a substantive nature for preferring this approach to logistic regression. Of special importance is the intuitive meaningfulness of the linear measures as differences in probabilities, and their applicability in causal (path) analysis, in contrast to the logistic measures.

Keywords Logistic regression · Binary variables · Significance tests

1 Introduction

In analyses of survey data it is not unusual that the dependent variable is a dichotomy. When the research problem requires a multivariate solution, regression analysis is very convenient for handling large numbers of independent variables.[1] Today it seems to be a common belief that with a binary dependent variable (dichotomy coded

[1] This opportunity is sometimes overexploited, however. A regression analysis of binary variables does not have access to information that is lacking in the corresponding tabular analysis. When more variables may be included in the regression analysis, this is due to the distributional assumptions on which the regression analysis is based. With a large number of variables one runs the risk that an estimate reflects the model more than the data. (Rubin 1997; Rothman and Greenland 1998).

O. Hellevik (✉)
Department of Political Science, University of Oslo, P.O. Box 1097, 0317, Blindern, Oslo, Norway
e-mail: ottar.hellevik@stv.uio.no

0–1) ordinary linear regression cannot be used.[2] Logistic regression is a prerequisite for an article to be considered for publication in a serious scientific journal, one hears.

Two statistical arguments are given for this rejection of linear regression. One is that with linear coefficients we risk meaningless results, since a predicted probability may fall outside the range 0–1. The other, which seems to be the main objection, is that the statistical tests for linear analyses are inappropriate with a binary dependent variable. These arguments are not as decisive as many seem to think, however. And when results from logistic regression are difficult to understand,[3] and in many cases may seem counterintuitive from a substantive point of view, there is every reason to ask for a more open minded discussion of the fruitfulness of different approaches to the statistical analysis of survey data. Hopefully the outcome of such a discussion would be that the choice of technique is guided more by considerations of what is meaningful in relation to the research problem, and less by a desire to demonstrate mastery of complicated statistical tools, than what may sometimes seem to be the case today.

2 Outline of the article

The article starts out by examining the two arguments of a statistical nature that are given as reasons why one should avoid a linear model when the dependent variable is a dichotomy. It is claimed that the risk for "impossible predictions" may be avoided by better specifications of the model. In the case of causal (path) analysis this objection is not relevant. With regard to inappropriate significance tests, I show by means of a simulation that the theoretical objection against the linear test is of little practical importance. The significance probabilities from linear and logistic regression analyses turn out to be nearly identical.

Having thus established that we actually have a choice when deciding how to analyze an empirical problem involving a binary dependent variable, arguments relevant to this choice is discussed. In order to reach also those unfamiliar with loglinear analysis, we start with a brief presentation of some of the central measures of statistical association used within this approach. A comparison between the results of linear and logistic measures is made, and reasons for preferring one or the other discussed.

The conclusion is that there in many situations exist compelling substantive reasons for preferring linear measures over the logistical. It would therefore be most unfortunate if we in fact were restricted to the latter measures when the dependent

[2] "Use of logistic regression — a statistical technique suited for the analysis of dichotomous categorical dependent variables — has become increasingly common in public opinion research. Outcomes involving voting, internet use, survey non-response, and opinions on a variety of controversial issues often take the form of yes/no, agree/disagree, favor/oppose, or present/absent. These outcomes do not meet the requirement of linear regression for an interval or ratio scale and often warrant use of logistic regression." From page 30 in the program for the AAPOR/WAPOR (American/World Association of Public Opinion Research) annual conference in May 2004, advertising the course "Making Sense of Logistic Regression: An Introduction".

[3] We cite once more from the ad for a course in logistic regression mentioned in note 2: "… for those who haven't had a graduate-level statistics course on the topic, making sense of logistic regression is daunting. The technique involves unfamiliar terms such as logged odds, odds, predicted probabilities, marginal effects, log likelihood values, pseudo-variance explained, and maximum likelihood estimation that can make meaningful interpretation of results difficult".

variable is a dichotomy, as it is often claimed. We will now take a closer look at the validity of the statistical arguments against the use of linear measures.

3 "Impossible" results of linear analyses?

Results outside of the interval 0–1 may occur when we determine the probability for a high value on the dependent variable for various combinations of values on the independent variables by summing the products of variable values and coefficients in the regression equation. The problem of "impossible results" cannot arise in a loglinear analysis, since these measures do not have limits to their variation. With a range from plus to minus infinity "anything goes".

But are predictions outside the range of meaningful variation really such a serious problem? "In practice, estimates of cumulative incidence would seldom be much below zero or above 1, and then only for extreme combinations of values of the independent variables. In place of such inadmissible estimates, it is reasonable to substitute the theoretical minimum or maximum value, as appropriate, for the inadmissible value." (Rothman 1986: 291).

In a linear regression analysis with binary independent variables "impossible" results may only occur when there is interaction in the data. In case of interaction, we will, by including the relevant interaction terms in the model, always be guaranteed that the predicted result will lie within the interval 0–1.

If some of the independent variables are continuous instead of binary, predictions outside the permitted range will still be possible, however. This may occur if the association between such a variable and the dependent variable is non-linear. When there is reason to question the assumption of linearity, one alternative is to transform the continuous variable into a set of dummy variables. Then we would still be protected against predictions outside the 0–1 interval, by including the relevant interaction terms in the analysis.

If the purpose of the analysis is not prediction but causal decomposition of bivariate associations (as in path analysis), the problem of "impossible predictions" is no longer relevant. Whether or not the conditional effects of an independent variable are different (interaction) or similar, we may calculate an average which tells us what impact on the dependent variable these conditional effects add up to within a population with a given composition (the distributions for the independent variables, which determine the weights in a tabular analysis (Hellevik 1983, 1984)). Such causal effects should not be used to predict probabilities, which means that whether a prediction falls within the range of 0–1 or not is irrelevant. What matters for the results of a causal analysis is whether the sum of components of direct, indirect and spurious effects is identical to the bivariate association. This requirement is met by the linear regression coefficients, but not by the logistic (Hellevik 1984).

4 Inappropriate linear significance tests?

When the danger of misleading significance tests are used as a decisive argument against linear regression with binary dependent variables, this first of all indicates that uncertainty arising from sampling is made into a bigger problem than it usually is in survey analysis. Random sampling error is just one of many possible methodological explanations of the results from surveys. With a reasonably large sample, random sampling error may often be rejected off hand as a possible cause of a tendency, as long as we are not interested in very weak effects.

Secondly it is pertinent to ask how important the objection based on statistical theory against the linear test is in practice. The problem with a binary dependent variable is that the homoscedasticity assumption (similar variation on the dependent variable for units with different values on the independent variable) is not satisfied. This is of no consequence for the regression coefficient. But the uncertainty estimate for the coefficient, and thus the test of significance, is affected. This seems to be the main basis for the widely held opinion that linear regression is inappropriate with a binary dependent variable.

How far are the results from linear tests from those of logistic tests? One possibility is that the divergence between the two is systematic, in the sense that the significance probability from one of the tests always is larger than the other. If it were the case that the linear tests consistently showed larger *P*-values than the logistic ones for the same empirical data, this would mean that these tests give a conservative estimate of the significance of a regression coefficient.

To look at the correspondence between the two tests a series of parallel linear and logistic regression analyses involving two independent variables (the binary variable gender and the continuous variable age) and a series of binary dependent variables (preferences for various parties and political attitudes) were carried out. The percentage with a high value on the dependent variable varied between 1 and 52, with an approximately even distribution on the ranges 0–9, 10–19, 20–29, and 30+, within each of four categories of random samples with approximately 200, 500, 1250 and 2500 respondents. A total of 320 comparisons of significance probabilities for the corresponding linear and logistic regression coefficients were made, 20 in each cell of the 4*4 matrix defined by sample size and distribution on the dependent variable.

It turned out that there was no systematic tendency for one of the two sets of *P*-values to be larger than the other. Even so the results were very interesting with regard to the applicability of linear tests of significance with a binary dependent variable.

Quite surprisingly in view of the dire warnings against the use of linear tests, the differences between the corresponding significance probabilities were extremely modest. The correlation between the two sets of *P*-values was as high as 0.9998, which means an explained variance of 99.96% when one *P*-value is regressed on the other. The average difference in probability (disregarding the sign) is 0.002. In just one case out of the 320 did a 5% level significance test result in conflicting conclusions for the two tests.[4]

The correspondence between the results is illustrated by the diagonal pattern of a scatterplot for the *P*-values from the two kinds of regression analysis. This is an agreement so perfect that it borders on complete identity (Fig. 1).

When we look at the *P*-values from the linear and logistic significance tests, they in 20% of the 320 cases were completely identical with four digit probabilities. In 43% of the comparisons the difference was less than 0.001, in 90% less than 0.005 (Fig. 2).

[4] The *P*-values were 0.0497 for the linear and 0.0505 for the loglinear coefficient. Of the remaining 319 coefficients 136 (43%) were found significant by both tests, and 183 (57%) were found not significant. The one case of disagreement constitutes 0.3% of the results. In comparison, we can in 16 of the tests expect to make a type I error by rejecting a true null hypothesis of b = 0 in the population when the significance level is 5%.

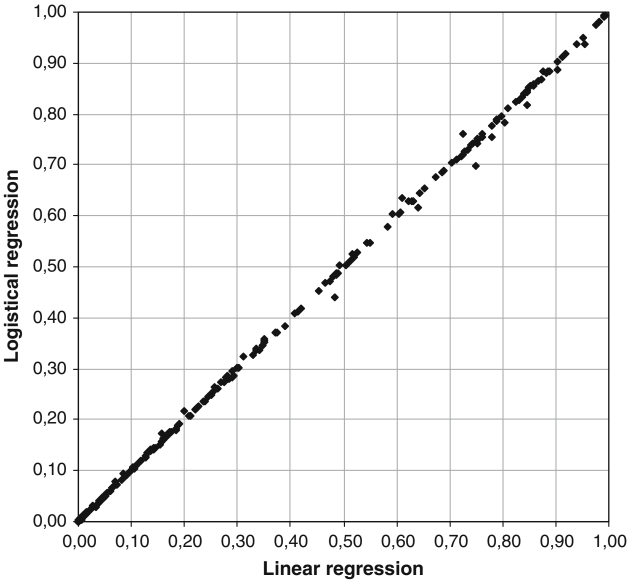

Fig. 1 The correspondence between significance probabilities for 320 linear and logistic regression coefficients (sample sizes varying between 200 and 2500, percentage high on dependent binary variable between 1 and 52)

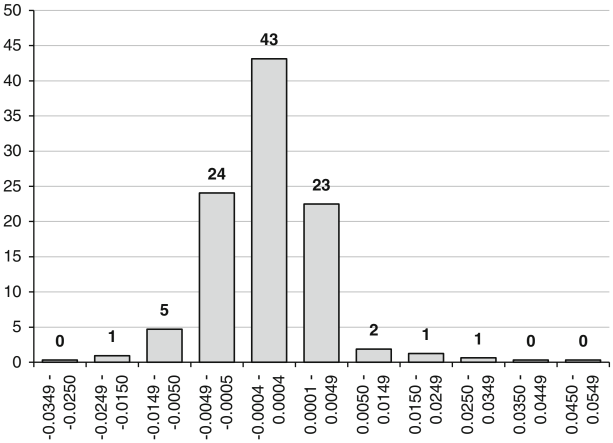

Fig. 2 The percent distribution of differences between significance probabilities for 320 corresponding linear and logistic regression coefficients (linear–logistic)

The size of the difference between the corresponding P-values is primarily affected by the distribution on the dependent variable in this series of tests. Size of sample also matters, with the exception of tests with a very skewed distribution on the dependent variable (percentages below 10). In these particular cases the pattern actually is in the opposite direction of what is to be expected (Table 1). When we compare the results for the binary and the continuous independent variables, there is a tendency towards slightly larger differences in P-value for the binary variable (in average 0.002).

Table 1 Average difference in size of significance probability for linear and logistic test (320 comparisons, evenly distributed between the 16 cells of the table)

Sample size	Percentage with high value on dependent variable				All
	0–9	10–19	20–29	30–	
200	0.0034	0.0033	0.0020	0.0017	0.0027
500	0.0062	0.0024	0.0010	0.0005	0.0027
1200	0.0060	0.0013	0.0009	0.0004	0.0020
2500	0.0068	0.0004	0.0010	0.0001	0.0021
All	0.0055	0.0019	0.0011	0.0008	0.0024

Table 2 Experiencing similarities between own work in the government office and that of a judge, depending on own education (Proportions)

Education	Law	Other	PD
Feels that own role is similar to that of a judge	0.75	0.20	0.55
Does not feel own role as similar to that of a judge	0.25	0.80	–0.55
Sum	1.00	1.00	0.00

The high agreement in test results must come as a surprise for all those who firmly believe that linear regression cannot be used with a binary dependent variable due to inadequacy of the significance test. The outcome is, however, in line with the statement we find in some textbooks on regression analysis that linear tests will give misleading results for significance tests only with small samples (Fox 1997: 451; Skog 1998: 301). There also have been similar results in empirical analyses reported earlier (Kanagy et al. 1994). The simulation reported in this article indicates that the linear test is so robust that even sample sizes unusually small for survey research and very skewed distributions on the dependent variable do not create problems.

It is our claim that the statistical objections to applying linear regression analysis with a dichotomous dependent variable may be put to rest. This means that we do have a choice between this approach and logistic regression. The discussion now turns to reasons for preferring one method over the other. But first a quick review of the two families of statistical measures.

5 Linear and loglinear measures

Table 2 shows the bivariate association between education and how own occupational role is experienced among government office employees (from Lægreid and Olsen 1978).

The role perception differs between those administrators with an degree in law and those with a different educational background. The association may be described by the *proportion difference* PD, which corresponds to the unstandardized *linear regression coefficient* b when the two variables are coded 0 or 1 (binary variables).[5]

[5] The average for a 0–1 variable is the same as the proportion of units with value 1. The constant is the proportion with this value among units with value 0 on the binary independent variable, and b the difference in this proportion when we move to units with value 1 on the independent variable. For a discussion of how regression coefficients in an analysis with binary variables correspond to proportions and proportion differences, see Hellevik 1984.

$$\mathrm{PD} = \mathrm{b} = \mathrm{P}_{\mathrm{law}} - \mathrm{P}_{\mathrm{other}} = 0.75 - 0.20 = 0.55$$

Besides looking at the absolute difference or gap between the proportions, one may calculate the ratio between them. We can call this relative difference for the *proportion ratio* (PR):

$$\mathrm{PR} = \mathrm{P}_{\mathrm{law}}/\mathrm{P}_{\mathrm{other}} = 0.75/0.20 = 3.75$$

Relative measures are rarely used in comparisons of groups, but are common in analyses of change over time. If the proportion with a higher education increases from 40 to 60%, this absolute change of 20 percentage points may be described as a change of 50% relative to the starting point.

In a loglinear analysis *odds* are calculated instead of proportions. The odds is a ratio which equals the proportion (or number) of units with a certain value on a variable divided by the proportion (number) of units without this value. Below the odds for experiencing own role as similar to that of a judge is calculated for those educated in law as well as for other educations among government employees. An odds larger than 1 shows that in the first group the number who feel similarities is higher than the number who does not. The odds for other educations is far below 1, telling us that here only a minority feels like a judge in their work.

$$\begin{aligned} \text{Odds:} \quad & \mathrm{O} = \mathrm{P}/(1-\mathrm{P}) \\ \text{Odds}_{\text{law}}: \ & \mathrm{O}_{\mathrm{l}} = 75/25 = 3.00 \\ \text{Odds}_{\text{other}}: \ & \mathrm{O}_{\mathrm{o}} = 20/80 = 0.25 \end{aligned}$$

To measure how different the two groups of bureaucrats are, we calculate the *odds ratio* (OR), which as the name suggests equals the ratio between the two odds, in other words the ratio between two ratios. The result of the calculation tells us that the odds for those educated in law is twelve times the odds for employees with other kinds of educations.

$$\text{Odds Ratio:} \quad \mathrm{OR} = \mathrm{O}_{\mathrm{l}}/\mathrm{O}_{\mathrm{o}} = 3.00/0.25 = 12.0$$

From the OR other measures of statistical association are derived. One is the natural logarithm of OR, called the logit, which corresponds to the *logistic regression coefficient* we find in a regression analysis of the variation in the odds for having a high value on the dichotomous dependent variable (here denoted B to distinguish it from the linear b).

$$\mathrm{B} = \ln(\mathrm{OR}) = \ln(12.0) = 2.48$$

Another loglinear measure is lambda, which is $\frac{1}{4}$ of B.

$$\text{Lambda} = 1/4\ \mathrm{B} = 1/4 \ln(\mathrm{OR}) = 1/4 \ln(12.0) = 0.62$$

Lambda in many cases will be identical to or only slightly larger than the corresponding proportion difference (or b) for a given empirical association. In our example lambda is 0.62 as compared to 0.55 for PD. In certain situations, however, lambda will take on much larger values. While PD or b varies between plus or minus 1, B or lambda has plus/minus infinity as extreme values.

6 What is measured?

The definition and calculation of the linear, relative and loglinear measures differ, and the results may be contradictory with regard to whether concrete empirical associations are strong or weak, and whether there is statistical interaction or not. To choose between them, one ought to have an idea of what aspects of reality the different measures capture. Only then it is possible to decide what measure is relevant for the purpose of the investigation.

The linear measures have interpretations that are intuitively easy to comprehend. A proportion may be understood as an estimate for the probability that the units studied have a specific value on a variable. A proportion difference or linear regression coefficient accordingly is the difference in this probability between groups of units defined by an independent variable. In a causal analysis we may talk about how the probability for having a high value on the effect variable is influenced by substituting units with different values on the causal variable, e.g. how the outcome of a recruitment process is influenced when we replace persons with a low education with persons with a high education (Hellevik 1984).

The relative measure may correspondingly be interpreted as the ratio of probabilities. In epidemiology the term relative risk is used when comparing the incidence of a medical condition within two groups by calculating the ratio. In studies of recruitment to attractive positions such as higher education the term participation ratio is sometimes used (Kivinen, Ahoa and Hedman 2001; Hellevik 2002). PR here is interpreted as a measure of equality–inequality in the distribution of a good, with 1 (identical probabilities) indicating complete equality.

When the largest proportion is used as numerator, as in our illustrative example, PR approaches infinity as inequality increases, and may thus be called a measure of inequality. If instead the largest proportion is used as denominator, we have 0 when inequality is extreme (one group gets nothing of the good). In this case increasing values for PR towards the maximum of 1 means increasing equality.

An odds is the ratio between the probability for having a certain value on a variable and the probability of not having this value. It is easily grasped what is meant by for instance an odds of 3 (or 3 to 1, as it is often said). This means that 75% of the units have a certain characteristic, which 25% lack. It becomes more difficult, however, when we look at the odds ratio. The OR is sometimes described as the ratio between probabilities, and it is stated that something is so and so many times more common in one group than in another.[6] This is incorrect, what OR shows is the ratio between odds, not between proportions as PR does. But when the proportions are small the results for PR and OR are similar. Within epidemiological research one for this reason sometimes uses such an interpretation of OR.[7] In analyses of survey data, however,

[6] A research report gave the following explanation for OR (translated from Norwegian): “the odds ratio expresses how many times more often failing health occurs in the group in question as compared to the reference group”. Comment to an OR of 3.8: “failing health occurs 3–4 times as frequently for those who …”. The ratio between the proportions involved was much lower, however: 0.39/0.15 = 2.6. Referred to in Hellevik (1996).

[7] “The odds ratio … approximates how much more likely (or unlikely) it is for the outcome to be present among those with x= 1 than among those with x= 0. For example, if y denotes the presence or absence of lung cancer and x denotes whether or not the person is a smoker, then ψ [OR] = 2 indicates that lung cancer occurs twice as often among smokers than among nonsmokers in the study population”. Hosmer and Lemeshow (1989), *Applied Logistic Regression* (New York: John Wiley & Sons).

where the phenomena studied usually occur more frequently, this interpretation often becomes misleading.

It does not get any easier to give a meaningful description of what is captured by a loglinear measure, on the contrary, when we turn our attention to B (the logistic regression coefficient). “The natural logarithm of the ratio between two odds” is probably not very helpful for many readers. This may explain why sometimes presentations of results from logistic regression analysis only comment on whether the effects are significant or not, and not on their magnitude. Or effects are converted to linear differences in probabilities given certain assumptions.

7 Differences in results

When the definition does not contribute much to an understanding of what aspects of social reality is captured by a stastical measure, it may be useful to look at how the measure “behaves” in concrete situations. We will mainly be comparing the more commonly used linear and loglinear measures, and comment briefly on the relative measures towards the end.

If as an example 60% of the men in a sample are in favor of Norway joining the European Union, while the corresponding figure is 45% for the women, we have a PD of $0.60 - 0.45 = 0.15$, which is also the result for the linear regression coefficient b.

The odds is $60/40 = 1.5$ for men, and $45/55 = 0.82$ for women, rendering the OR $1.5/0.82 = 1.83$. We find the logistic regression coefficient B as $\ln(1.83) = 0.60$. Lambda is $1/4$ B $= 1/4$ $0.60 = 0.15$. The results in other words are identical for b and lambda. In general it holds that within the interval from 0.30 to 0.70 lambda will be identical to or just slightly larger than the PD, and the logistic B will be four times the size of the linear b.

When we move outside of this range, and towards the extremes of 0 or 1 for the proportion with a high value on the dependent variable, lambda will take on much higher values than PD. If for instance 3% of all men are candidates in local elections, as compared to 1.5% of all women, we have a modest PD of 0.015, one tenth of the effect on EU opinion mentioned above. With odds of $3/97 = 0.031$ for men and $1.5/98.5 = 0.015$ for women, we have an OR of $0.031/0.015 = 2.1$. This means that the association according to lambda is higher in the second than in the first example: $1/4 \ln(2.1) = 1/4$ $0.74 = 0.185$. In this case B is not four but nearly fifty times larger than b ($B / b = 0.74 / 0.015 = 49.3$).

The same contrast between the two kinds of measures is found at the opposite extreme of the scale. If for instance 99.99% of the men and 99.90% of the women took part in the “elections” of a dictator, the difference between the sexes according to the linear regression coefficient b is microscopic: 0.0009. The logistic coefficient B is 2.30, or 2559 times the size of b. According to the logistic coefficient this difference in turnout between men and women of 0.09% actually is larger than what we have when the turnout is 25% for women and 75% for men (B in this case is 2.20).

This shows that statistical association must mean quite different things within the two approaches. In the last two examples the linear b tells us that the sex of a person makes very little difference when it comes to being a candidate or vote. In the first case because very few are nominated, irrespectively of whether they are men or women, while in the second case almost all vote regardless of sex. Knowing the sex of a person accordingly has little value when we want to predict his or her position on

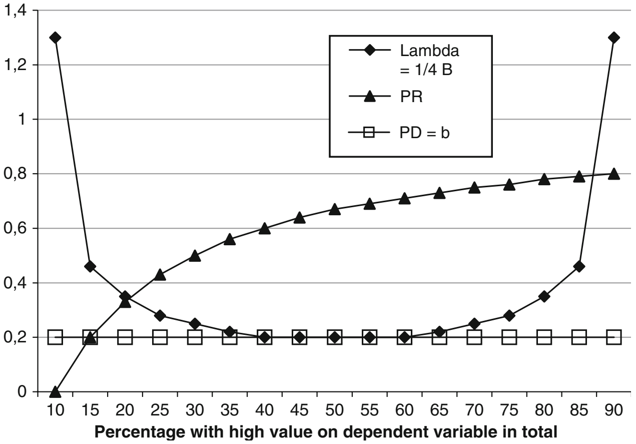

Fig. 3 Lambda, proportion difference (linear b) and proportion ratio for comparisons of two groups of equal size (with proportions for the binary dependent variable 0.10 above and below the total level)

the dependent variable. In a causal perspective we would say that the sex of a person has very little effect on the probability of being a candidate or a voter.

Figure 3 gives a more general picture of the difference between the two kinds of measures by comparing two groups (which for simplicity of calculations are of the same size, e.g. men and women), at various levels for the dependent variable. The proportion is 0.10 higher than the total level for one of the groups, and 0.10 lower for the other group, which means that we have a PD of 0.20 at all levels (0.20 vs. 0.00; 0.25 vs. 0.05, and so on). The linear regression coefficient b thus is 0.2 at all levels for the distribution on the dependent variable. The loglinear lambda takes on the same values at the medium levels, but becomes much higher near the extremes. The same pattern holds for the logistic regression coefficient B (not shown), which is four times the size of lambda.

Figure 3 also includes the relative measure PR, with the largest proportion used as the denominator. To the left in the figure, where the proportions in the two groups are 0 and 0.2, respectively, we have the minimum value for equality (0). As the proportion in each group rises, PR moves closer to maximum equality (1). In other words, the higher the level for the dependent variable, the less inequality resulting from the constant absolute difference, according to the results for the relative measure. This pattern of monotonous increasing values for PR is different from that of both the linear and the loglinear measures. For this reason it has been argued that neither of these two classes of measures are suited for analyses of inequality (Hellevik 1997, 2000, 2002).

Some of the examples discussed above indicate that loglinear measures may give results that seem counterintuitive. We shall now look at situations where the special properties of loglinear measures appear substantively meaningful.

8 Variations in rare but important phenomena

When studying a dependent variable where the value of interest only rarely occurs, the effects of independent variables may be so small that they are hardly noticeable when analyzed by means of linear measures. Even so the tiny differences between groups in the incidence of a fatal illnesses, suicide rates, etc. will be important enough to warrant closer scrutiny. For this reason logistic measures, which may manifest strong effects in such cases, are often used in epidemiology. Another example is quantitative studies of conflicts between states, where the probability for war fortunately is small.

In analyses of such rare occurrences also the linear effects (or the relevant proportions) should be reported. Otherwise there is a danger that misleading impressions of large absolute differences are created. A newspaper reporting from the results of a skin cancer research project (Veierød et al. 2003) used the headline "55% higher chance of getting skin cancer if you take solarium".[8] The headline for interviews with persons who take solarium was "Frightening". They would probably have been less alarmed if the newspaper instead had reported that the absolute difference in probability was less than 0.001.[9]

9 Changes over time

With trend data the loglinear measures may be interpreted in relation to a model of logistical growth (Heath et al. 1992). We may take as an example a process of social diffusion, such as the dissemination of a rumor by word of mouth within a population, from the starting point where no one has heard the rumor to the end point where all have been reached. At the outset the proportion having heard the rumor will grow very slowly, since so few are spreading it. Then the dissemination speeds up as more people become active, until the growth once more flattens out as the ceiling of 100% knowing the rumor approaches, making it difficult finding someone who has not already heard it.

A logistical growth curve depicts such a development. In this case lambda will have the same size for all time intervals of equal width, no matter if we look at the beginning, the middle period or the end of the diffusion process. The corresponding proportion differences for changes in knowledge will vary considerably, however. They are much smaller at the outset (start effect), and towards the end (saturation or ceiling effect), but similar to lambda in the middle of the process. A stable lambda over time thus tells us that the growth follows a logistical pattern (Fig. 4).

A logistical growth model does not seem appropriate for all developmental processes, however. With regard to for instance attainment of higher education, we have the possibility of decline as well as growth. This creates a problem for the substantive justification of the loglinear measures. In a situation of growth, it may with reference to the ceiling effect seem reasonable to regard an *increase* from 85 to 95% as stronger or more "demanding" than an increase from 45 to 55%. This is what lambda indicates, taking on the value of 0.3 in the first and 0.1 in the last case.

According to the loglinear measures also the *decrease* from 95 to 85% should be regarded as three times as strong as the decrease from 55 to 45, since lambda is $-0,3$ in the first and $-0,1$ in the second case. Intuitively it is not easy to understand why it should be regarded as "harder" to drop 10 percentage points from 95 than from 55%.

[8] Headline in the newspaper Aftenposten 17.10.2003, referring to those who take solarium more than once a month in a study of more than 100,000 women in Norway and Sweden.

[9] The probabilities were 0.0024 and 0.0017, giving an absolute difference of 0.0007. This can be calculated from information in table 5 in Veierød et al. (2003: 1536). The relative risk according to these figures is 1.41, which after multivariate control for relevant variables increases to 1.55.

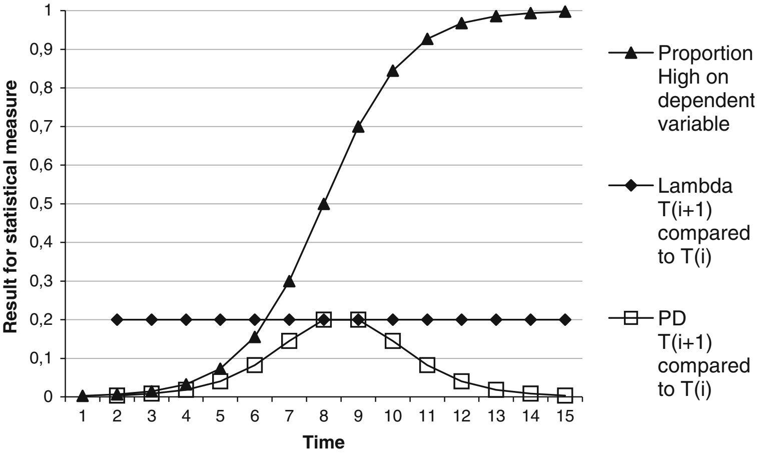

Fig. 4 The behavior of lambda and the proportion difference in a situation with logistical growth over time

It no longer makes sense to talk about a ceiling effect, when we are moving away from instead of closer to 100%.

If the variable has a set of values that may be regarded as symmetrical (e.g. being in favor of or opposed to a certain policy), a decline in support for one alternative may instead be seen as an increase for the other. In this way all changes may be defined as increases. The above discussion thus applies to variables with asymmetrical value sets, such as level of education, income, voting, holding political office, etc., where some values are regarded as more attractive or rewarding than others.

10 Dependent variable with a fixed overall distribution

In processes of change the level of the dependent variable at a certain point in time among other things may depend upon what the level was at prior points in the process. This is an argument for letting the level influence the result, as is the case for loglinear measures. What then when we compare groups at the same point in time? In some cases one may argue that there exists some form of coupling between the levels of the dependent variable within the various groups, namely when the total distribution for the dependent variable is fixed.

An example of a fixed distribution might be the proportion of a population obtaining higher education, in a situation where a limited number of positions is provided, and this number does not meet the demand. If we were told the number of educational positions obtained by men, we would also know the number available to women, given that the total number is fixed. In this situation there will be limits to the size of the difference in educational attainment between the sexes.

Let us assume that there are positions of higher education for 40% of those seeking such a position. Then the maximum possible value for PD would be 0.80.[10] This is

[10] For simplicity we assume the distribution on sex is exactly 50–50 for those seeking a higher education.

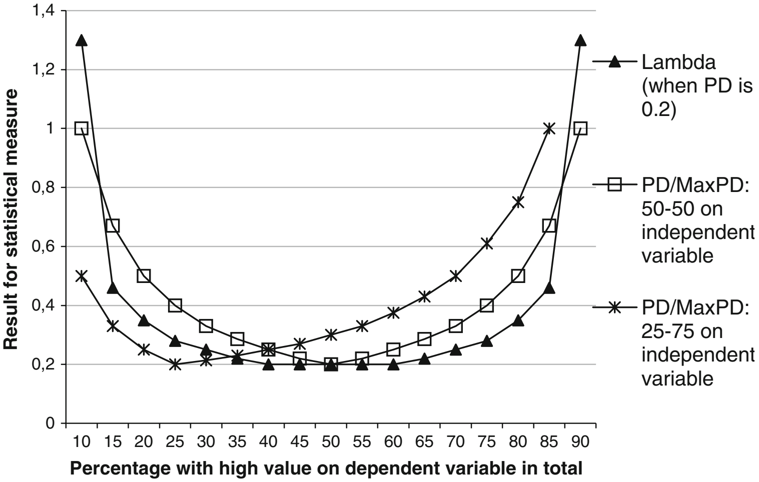

Fig. 5 Comparison of lambda and the proportion difference standardized by taking maximum value into account (for two different distributions on the independent variable)

seen if we "give" all positions to one of the groups, for example men. The proportion with a higher education then becomes 0.80 for men, and 0.00 for women.

If the provision of positions of higher education equals 70% of the applicants, the maximum sex difference is 0.60 (in this case with 100% attainment among men, and 40% among women, when they are given the remaining positions).

In this situation it makes sense to see the actual difference in relation to the maximum possible, to obtain a measure standardized to vary between 0 and ±1. This maximum-corrected PD is found by dividing the actual PD with the maximum possible given the fixed distribution on the dependent variable. If for instance we have an actual sex difference of 0.40, the maximum-corrected PD would be 0.4/0.8 = 0.50 in the first example above, and 0.4/0.6 = 0.67 in the second.

It has been suggested that loglinear measures of association provide such a correction for variations in the marginal distribution on the dependent variable (Heath et al. 1987: 270). It turns out, however, that the results for lambda deviate somewhat from the maximum-corrected PD (Hellevik 2002), as shown in Fig. 5.

The deviation is especially large when the distribution on the independent variable is uneven. The size of lambda is not influenced by this distribution, while the maximum possible difference is. If we want to correct our measure of association for limits to its maximum value given the fixed distribution on a dependent variable, other measures are better suited than the loglinear ones.

When the distribution on the dependent variable is not fixed, the proportions of other groups do not restrict the proportion with a certain value on this variable within a group. Whatever the percentage among men who are pro Norwegian EU membership, this does not in any way restrict the level possible for such an attitude among women. The maximum difference possible in this case would be 100%. In such a situation it is hard to see what the substantive justification could be for regarding the difference between 60 and 40% as a much smaller association than the difference between 90 and 70%, which is what the loglinear measures suggest (lambda equal to 0.20 and 0.34, respectively). But if we choose to apply these measures, including the logistic regression coefficient, we implicitly assume that it is meaningful to distinguish between these situations that according to linear measures are identical.

To accentuate this point: We would have to argue why it makes sense to regard a difference in the tendency to vote between 35% for women and 65% for men as a much smaller association than the difference between participations rates of 99.90 and 99.99%, since lambda is 0.31 in the first and 0.58 in the second case.

11 Other perspectives on the interpretation of loglinear measures

This discussion does not exhaust the list of possible substantive justifications for using non-linear models in analyses with dichotomous dependent variables. In studies of choice situations economists use the so-called probit model, which is similar to the logistic model.[11] The underlying assumption is that choices result from cost-benefit calculations which are not immediately observable, but which leads to a shift in value on the dependent variable when benefits surpasses costs. Using the probit model implies the assumption that the unobserved variation in preferences is normally distributed (se for instance Greene 1993: 642).[12]

This perspective is applicable in other situations where change of value on a dichotomous dependent variable may be seen as the result of unobserved characteristics or propensities of the units under study. If it seems reasonable to assume that these variables are normally distributed, the independent variables will have a diminishing effect on the probability for a change in the dependent variable the closer we move towards the extreme values of 0 and 1.

12 The purpose of the analysis

When deciding which statistical technique to use the purpose of the analysis ought to be an important consideration. A decisive distinction is the one between prediction and causal analysis (Hellevik 1983, 1984). Table 3 gives a brief overview over the differences between the two approaches to multivariate analysis.

The use of loglinear measures is restricted to prediction analysis; they cannot be used for causal purposes (path analysis). The effects found will not give a sum of components identical to the bivariate association one wishes to interpret causally (Davis and Schooler 1974; Reynolds 1977). A correct decomposition in this sense can only be obtained by using linear measures. With linear regression coefficients the sum of components will always correspond to the bivariate association, whether the variables are binary or continuous. Another possibility is to use weighted averages of proportion differences from tabular analyses (Boyle 1966; Hellevik 1983, 1984).[13]

[11] While the logistic model is based on the logistical distribution the probit model is based on the better-known normal distribution. Both distributions have approximately the same bell shaped form. The logistical growth curve discussed earlier is a cumulative logistical distribution.

[12] Point communicated to me by Axel West Pedersen.

[13] With more than three variables in the model there is, however, a possibility for minor deviations from the bivariate association of the sum of components (Hellevik 1984: 168–170).

Table 3 Prediction and causal analysis

	Prediction analysis	Causal analysis
Purpose	Predict the value on Y (dependent variable) from knowledge of value on set of X'es (independent variables)	Explain bivariate association between pair of variables by decomposing it into direct, indirect and spurious effect
Model	Two tiered: Two or more X, one Y	Hierarchical: Causal ordering X1 > X2 > X3 > X4
Design for analysis	Multivariate table / regression equation with all X'es and Y	Set of tables / regression equations (two alternative designs)
Calculation of effects	Variable and interaction effects. (In tabular analysis: unweighted average of conditional effects and differences between conditional effects. In regression analysis: equation with interaction terms)	Variable effects. (In tabular analysis: weighted average of conditional effects. In regression analysis: equation without interaction terms)
Requirement for result	Predicted value close to actual value for dependent variable	Sum of components identical to bivariate association
Statistical techniques	Percentage/proportion table analysis Linear regression Loglinear analysis of tables. Logistic regression	Percentage/proportion table analysis Linear regression

13 Conclusion

The statistical arguments against the use of linear regression with a binary dependent variable are not as decisive as it is often claimed. Even if the homoscedasticity assumption is violated, this in practice has little effect on the outcome of significance tests. The results for linear and logistic significance probabilities as we have seen turn out to be nearly identical, even with small samples and skewed distributions on the dependent variable. The problem of impossible predictions may be avoided by including interactions terms in the regression equation and/or by transforming continuous variables into dummy variables.

This means that we are not restricted to use logistic regression with a binary dependent variable. This is fortunate, since there in many cases are compelling substantive arguments for preferring the linear approach.

Proportion differences and coefficients from a linear regression analysis of binary variables may be interpreted as the difference in probability for having a certain value on the dependent variable for units with different values on an independent variable (in a multivariate analysis controlled for other independent variables). This is an interpretation that is substantively meaningful and easy to comprehend. The same cannot be said for the loglinear measures of association.

The intuitively meaningful interpretation makes it easier to communicate research results to broader audiences of interested citizens. Even without statistical training they may be able to comprehend and critically evaluate the conclusions drawn by researchers. For the natural sciences it may not be a problem that results are accessible only for specialists. But for the social sciences where the phenomena under study often will be of great concern for many outside the scientific milieu, it must be seen as an important consideration that the statistical techniques should not be more complicated than absolutely necessary.

The linear measures can be used to answer causal questions, which often is what the researcher wants to do. Loglinear measures are not applicable in causal analyses, since they do not provide an accurate decomposition of bivariate associations.

With proportions near 0 or 1 on the dependent variable, where the two kinds of measures show conflicting results, those from loglinear analyses in some cases appear to give little substantive meaning.

References

Boyle, R.P.: Causal theory and statistical measures of effect: a convergence. Am. Sociol. Rev. **31**, 843–851 (1966)

Davis, J.A., Schooler, S.R.: Nonparametric path analysis — the multivariate structure of dichotomous data when using the odds ratio or Yule's Q. Soc. Sci. Res. **3**, 267–297 (1974)

Fox, J.: Applied Regression Analysis, Linear Models, and Related Methods. Sage Publications, Thousand Oaks, CA (1997)

Greene, W.H.: Econometric Analysis, 2nd edn. Macmillan Publishing Company, New York (1993)

Heath, A., Jowell, R., Curtice, J.: Trendless fluctuation: a reply to Crewe. Polit. Stud. **35**, 256–277 (1987)

Heath A., Mills C., Roberts J.: Towards meritocracy? Recent evidence on an old problem. In: Crouch C., Heath A. (eds) Social Research and Social Reform: Essays in Honour of A.H. Halsey. Clarendon Press, Oxford (1992)

Hellevik, O.: Decomposing proportions and differences in proportions: approaches to contingency table analysis. Qual. Quant. **40**, 79–111 (1983)

Hellevik, O.: Introduction to Causal Analysis. Exploring Survey Data by Crosstabulation. George Allen & Unwin, London (1988. Oslo: Norwegian University Press) (1984)

Hellevik, O.: Fagkritikk av oppdragsforskning. Sosiologisk tidsskrift **4**, 219–228 (1996)

Hellevik, O.: Class inequality and egalitarian reform. Acta Sociologica **40**, 377–397 (1997)

Hellevik, O.: A less biased allocation mechanism. Acta Sociologica **43**, 81–83 (2000)

Hellevik, O.: Inequality versus association in educational attainment research:. comment on Kivinen, Ahola and Hedman. Acta Sociologica **45**, 151–158 (2002)

Hosmer, D.W., Lemeshow, S.: Applied Logistic Regression. John Wiley & Sons, New York (1989)

Kanagy, C.L., Humphrey, C.R., Firebaugh, G.: Surging environmentalism: changing public opinion or changing public? Soc. Sci. Quart. **75**, 804–819 (1994)

Kivinen, O., Ahola, S., Hedman, J.: Expanding education and improving odds. Participation in higher education in Finland in the 1980s and 1990s. Acta Sociologica **44**, 171–181 (2001)

Lægreid, P., Olsen, J.P.: Byråkrati og beslutninger: En studie av norske departement. Universitetsforlaget, Oslo (1978)

Reynolds, H.T.: The Analysis of Cross-Classifications. Free Press, New York (1977)

Rothman, K.J.: Modern Epidemiology. Little, Brown and Company, Boston (1986)

Rothman, K.J., Greenland, S. (eds.): Modern Epidemiology, 2nd edn. Lippincott-Raven Publishers, Philadelphia (1998)

Rubin, D.B.: Estimating causal effects from large data sets using propensity scores. Ann. Internal Med. **127**, 757–763 (1997)

Skog, O-J.: Å forklare sosiale fenomener. En regresjonsbasert tilnærming. Ad Notam Gyldendal, Oslo (1998)

Veierød, M.B., Weiderpass, E., Thörn, M., Hansson, J., Lund, E., Armstrong, B., Adami, H-O.: A prospective study of pigmentation, sun exposure, and risk of cutaneous malignant melanoma in women. J. Nat. Cancer Inst. **95**, 1530–1538 (2003)