(5)

INTERPRETING MULTIPLE LOGISTIC REGRESSION COEFFICIENTS IN PROSPECTIVE OBSERVATIONAL STUDIES*

Robert D. Abbott, Ph.D.

Biometrics Research Branch
National Heart, Lung, and Blood Institute
Bethesda, MD 20205

1-301-496-5826

Raymond J. Carroll, Ph.D.§

Department of Statistics
University of North Carolina
Chapel Hill, North Carolina 27514

1-919-962-1279

*Presented in part at a workshop on the relationship of hypertriglyceridemia to to atherosclerosis, June 8, 1981, sponsored by the National Heart, Lung and Blood Institute, Bethesda, Maryland.

§Raymond J. Carroll's work is supported by the Air Force Office of Scientific Research, Contract AFOSR F49620 82 C 0009.

INTERPRETING MULTIPLE LOGISTIC REGRESSION COEFFICIENTS IN PROSPECTIVE OBSERVATIONAL STUDIES

Summary

Multiple logistic models are frequently used in observational studies to assess the contribution of risk factors to disease. In the presence of correlation among risk factors, the estimated magnitude of a multiple logistic coefficient can become uncertain or meaningless. This paper highlights the problem of interpreting a multiple logistic coefficient and suggests a procedure for examining the total contribution of a risk factor to disease that includes a direct association and associations that exist through relationships with other antecedent characteristics. Examples are given, along with results that are not immediately obvious when considering the multiple logistic coefficient alone. Conclusions that are presented are important in biological studies if isolating the effect of an antecedent characteristic is unreasonable in the presence of confounding influences.

Running head: Interpreting Multiple Logistic Regression Coefficients

Keywords: multiple logistic regression, prospective observational studies, correlation, projected slope.

Introduction

The multiple logistic model is a common statistical tool used to analyze data from prospective observational studies when the endpoint is a dichotomous variable (1,2). For example, in the Framingham Heart Study (3), the endpoint of interest is often whether or not an individual develops coronary heart disease (CHD). If one is interested in the effect of triglyceride (TG) on the probability of developing CHD, the first step might be to model this effect by a univariate logistic analysis:

$$\text{logit}[p=\text{probability of CHD}] = \log[p/(1-p)] = \beta_0+\beta_1(TG)$$

As reported in more detail later in this presentation, for Framingham males, an estimate of β_1 is 0.437 with $p<0.05$, indicating that TG is a significant univariate predictor of CHD. One can now easily estimate the probability of developing CHD given an individuals TG value. Furthermore, given two different values of TG, TG_1 and TG_2, one can also compute the odds ratio of developing CHD based on the value of TG_1 relative to the value of TG_2. That is,

$$\text{odds ratio} = \exp[\beta_1(TG_1-TG_2)]$$

Note that when β_1 is significant, the odds ratio will be significantly different from one.

At this point, most investigators would then consider a more complete analysis, attempting to uncover the relationship between CHD and TG controlling for covariables such a high density lipoprotein cholesterol (HDL-C), total cholesterol (T-C), and Metropolitan relative weight (MRW). The investigator would then fit the logistic model

$$\text{logit}[p=\text{probability of CHD}] = \beta_0+\beta_1(TG)+\beta_2(HDL\text{-}C)+\beta_3(T\text{-}C)+\beta_4(MRW)$$

For the Framingham males, the estimate for the multivariate coefficient for TG,

β_1, is -0.183 which is not statistically significant ($p>.10$). On the other hand, the coefficient for HDL-C, β_2, is -0.048 which is statistically significant ($p<.05$). The coefficient for T-C, β_3, is 0.005 which is significant at the 0.10 level and the coefficient for MRW, β_4, is 0.002 which is not significant.

Having performed the above analysis, it is quite natural for the investigator to conclude that for Framingham males, while TG is a significant univariate predictor of CHD, most of its predictive ability can be explained through HDL-C, T-C and MRW. This is often phrased as something like, "TG is not a significant independent predictor of CHD." The usual implied set of conclusions then follows:

a. Most of the effects of TG on CHD are explainable by HDL-C and to a lesser degree the other covariates.
b. TG is an unimportant variable in the study of atherogenesis.
c. Altering TG to reduce CHD risk may be ineffective.

One purpose of this article is to show that in prospective observational studies, the three conclusions outlined above can result in a misleading understanding of the relationship of TG to CHD. This dilemma is often encountered and discussed in terms of confounding or multicollinearity (4-7). Our attempts in this presentation will be to introduce a different perspective which will be of use to epidemiologists in explaining the consequence of these misleading conclusions and what important information can be salvaged. The problem of course is that in prospective observational studies, the predictor variables such as TG and HDL-C are likely to be h_ghly correlated. For Framingham males the correlation is -0.451 ($p<0.05$). This means that for a given level of HDL-C, the variation of TG may be small, so that it may be unlikely to expect that for a fixed value of HDL-C that TG should have much of a relationship with CHD because information on TG is insufficient. As a result, the evidence is not

available to support the three conclusions given above. In this instance and in other examples that will be presented, it will be shown that from prospective observational studies it is often difficult to investigate the three conclusions if the studies are not specifically designed to do so. If, for example, all levels of TG could be cross classified with all levels of HDL-C then such conclusions are possible to consider. Cross classification of this type, however, is usually a goal of controlled clincial trials and is not commonly experienced in observational type studies.

The second purpose of this article is to provide a simple method for better explaining the association of TG with CHD. We will define a statistic called the projected slope which measures the effect of changing TG levels on the probability of developing CHD, while at the same time considering the effect of the other covariates on CHD and the relationship between TG and these covariates. The projected slope is not new and has appeared elsewhere (8). It has been shown to be a useful statistic based on the same ideas used in path analysis (9) for linear regression, and can be easily used in many analyses. We also provide some additional examples from the Framingham data on the use of the projected slope involving sets of risk factors for predicting CHD other than those already introduced.

Along with the purposes of the paper described above, we acknowledge that the limitations of multiple logistic regression mirror those that are exhibited in the usual least squares regression situation. The multiple logistic model receives special emphasis in this paper, not because it is characterized by any unique statistical feature used in estimating parameters, but because it has appeared in so many investigations linking risk factors to disease (10-15). Furthermore, its recent introduction into widely used statistical packages (16, 17) has encouraged its use and the attendant need for the cautions and caveats

that are given here.

Interpreting the Meaning of a Multiple Logistic Coefficient

Suppose there is interest in estimating a logistic expression for the probability of developing disease conditional on knowing two characteristics. For every individual examined, a bivariate observation of antecedent characteristics can be plotted as in Figure 1 along the x_1 x_2 plane. As an exmaple, x_1 might be TG, while x_2 might be HDL-C. The point (x_{1i},x_{2i}) represents the observations for the ith person: i.e., $(TG_i,HDL\text{-}C_i)$.

These data points are observed at the beginning of a study, and when the study has terminated a tally of healthy and diseased individuals is made. All of the data are used to provide estimates of coefficients for a logistic equation. The resulting equation describes a response surface represented in Figure 1 as a plane. The height of the response surface reflects the estimated risk of disease for individuals who possess characteristics directly below the plane. For the ith person with characteristics (x_{1i},x_{2i}), the probability of developing disease can be estimated from the logistic equation and is represented by the height of the arrow (the height falling somewhere between zero and one).

Suppose that the multivariate logistic regression coefficient associated with x_1(TG) is zero, while that associated with x_2 (HDL-C) is negative. Thus, x_2 is said to be inversely associated with disease while there is no association between disease and x_1. This interpretation can be easily described geometrically by considering Figure 1. Note that changes made parallel to the x_1 axis, when x_2 is held fixed, do not affect the chance of developing disease, corresponding to a coefficient of zero that is associated with x_1. In contrast, changes made parallel to the x_2 axis do affect the chance of developing disease. In fact, for a

given value of x_1, increases in x_2 will reduce the height of the response surface and reduce the estimated chance of developing disease, consistent with the inverse association between x_2 and disease that is implied by the corresponding negative coefficient associated with x_2.

The scatter of points in Figure 1 in which x_1 and x_2 are unrelated might well be observed in a controlled clinical trial. In this instance, it makes sense to discuss the impact of holding one characteristic fixed and interpreting the importance of another characteristic through the multiple logistic coefficient, because all combinations of characteristics have been observed and are reasonable to consider. In this instance, unlikely combinations of characteristics are not being created by holding one characteristic fixed and then adjusting the other.

The data typical of prospective observational studies rarely result in predictors x_1 and x_2 which are unrelated. Correlations between risk factors are the rule rather than the exception. Such an instance is described in Figure 2. One can see that the data points represented in the x_1 x_2 plane tend to fall along a line. For example, small values of x_1 are related to large values of x_2. In contrast, there are no data points in which both x_1 and x_2 are near zero. Clearly, it is not meaningful to discuss the effect of changing x_1 while holding x_2 fixed, but it is just this assumption which is at the heart of the reasoning used to support the three conclusions given in the Introduction; i.e., for given levels of x_2 we force unobserved differences in x_1 that enable us to imply that x_1 is not independently related to disease. We are basing this decision on insufficient data that is observed for fixed values of x_2.

In Figure 2, unlike the example in Figure 1, the response surface no longer rests above a sample of all combinations of values of x_1 and x_2, but behaves like a teeter-totter resting on a locus of points projected up from the observed

data. Note that the only area of the response surface that has any meaning is that area directly above the observed data. This is true because the data are insufficient to suggest that other areas of the surface adequately represent the chance of developing disease. As in the usual linear regression problem, the variances of the estimated multiple logistic coefficients are potentially inflated by the correlation between x_1 and x_2, and the instability of the response surface, which may result in an uncertain indication of the importance of x_1 and x_2 in predicting disease.

From Figure 2, two components that relate x_1 with disease can be envisaged. The first is a direct or independent effect or association. The second relates x_1 with disease indirectly via an association with x_2 and the association x_2 has with disease. Figure 2 illustrates that it is not clear how to interpret the magnitude of the multiple logistic coefficient associated with the slope of the lines in the response surface appearing in the same planes as the x_1 and x_2 axes, because levels of x_1 are related to levels of x_2. In such an instance, assessing the effect on disease by changing x_1 is not realistic unless values of x_2 are also changed in a way that is observed in nature. To change x_1 while holding x_2 fixed may exceed the limits of the data and may be contrary to what is possible. This is where interpretation of the multiple logistic coefficient of x_1 becomes misleading because the independent component associated with changing x_1 alone cannot be realistically separated from the component represented by the indirect association that exists between x_1 and x_2 and the relationship x_2 has with disease.

Thus, in the many practical situations in which the predicting characteristics are highly correlated, interpreting the multiple logistic coefficient by considering one characteristic held fixed while changing the other may be unreasonable. One may be artificially producing unlikely combinations of character-

istics and formulating extrapolations that exceed the limitations of the observed data. We feel that a more useful analysis of the predictive importance of a variable should not hold constant the level of another variable to which it is physiologically related, but rather, allow the characteristics to vary simultaneously as they would be expected to biologically. As illustrated in Figure 2, it would be important to consider not only the multiple logistic coefficients of x_1 and x_2, but also the slope of the line connecting the points P and Q that lies above most of the data that are observed and the regression line between x_1 and x_2. Consideration of the slope of the line designated by P and Q is appealling because it is a function of the relationship between x_1 and x_2 as well as their relationships with the disease. If, for example, the characteristic x_1 is altered, on the average, x_2 will also be altered, and the chance of developing disease will move along the line marked by P and Q. For lack of a better term, the slope of the line connecting P and Q when written in the logit scale will be called the projected slope.

The Projected Slope

The preceeding discussion has focused on the effects on logistic regression due to correlation between predictor variables. This is, of course, a special circumstance of what has been called multicollinearity or confounding, which is a general issue affecting all nonrandomized studies. It is not our purpose here to become involved in the controversies surrounding the problem of confounding. Rather than trying to discuss the independent effect of a predictor such as TG, we will use the idea of the projected slope to try to see if a particular variable has any predictive effect on the probability of disease. As mentioned before, the development is based on the ideas of path analysis, which is often

used in linear regression but not multiple logistic regression.

First, we suspect that there may be a linear relationship as in Figure 2 between x_1 and x_2. Specifically, we may think that x_1 can be used to predict x_2; e.g., TG predicting HDL-C. This is written symbolically as

[1] $$x_2 = \gamma_1+\gamma_1 x_1+\varepsilon$$

Conditionally, once we have observed x_1 and x_2, we hypothesize a multiple logistic regression model:

[2] $$\text{logit[probability of CHD]} = \beta_0+\beta_1 x_1+\beta_2 x_2 \ .$$

Informally, we could substitute the expected value of [1] into [2] obtaining as an approximation

$$\text{logit[probability of CHD]} = (\beta_0+\gamma_0\beta_2) + (\beta_1+\gamma_1\beta_2)x_1$$

It turns out, that $\beta_1+\gamma_1\beta_2$ is the projected slope associated with x_1.

The projected slope can be derived more formally as follows. If we take two people exhibiting predictors (x_{1i},x_{2i}) and (x_{1j},x_{2j}) that appear along the regression line between x_1 and x_2 in Figure 2, the log odds ratio of developing disease for these two people has expectation

$$(\beta_1+\gamma_1\beta_2)(x_{1i}-x_{1j})$$

It is clear that if $\beta_1+\gamma_1\beta_2=0$, then the slope of the line connecting P and Q shown in Figure 2 (that is projected up from the regression between x_1 and x_2) will be zero.

One way to better understand the meaning of the projected slope is through consideration of Figures 1 and 2. In Figure 1 there is no effect of x_1 on the probability of disease, as we have seen geometrically. Since x_1 and x_2 are unrelated in this figure, $\gamma_1=0$. Further, the multiple logistic coefficient for x_1 is $\beta_1=0$. This means the projected slope is $\beta_1+\gamma_1\beta_2=0$, indicating no effect

on the risk of disease due to differences in x_1. In Figure 2, changing x_1 should change x_2 which in turn will change the risk of disease. Here, $\gamma_1 \neq 0$, $\beta_2 \neq 0$, $\beta_1 = 0$ and the projected slope is $\gamma_1 \beta_2 \neq 0$ as expected.

Details of estimating the projected slope from data as well as its definition when there are more than two predictors are provided in the Appendix. A test of significance of the projected slope is equivalent to testing H_o: $\beta_1 + \gamma_1 \beta_2 = 0$. This tests whether or not x_1 has any predictive effect on the risk of disease. A discussion of the mechanics of making this hypothesis test is also provided in the Appendix.

It can also be shown that in certain situations the estimated projected slope for a risk factor is asymptotically equal to the univariate logistic regression coefficient that relates the risk factor to disease (7). The asymptotic convergence of the estimated projected slope to the univariate coefficient, however, is not guaranteed. Nevertheless, the consequence is that it can emphasize the importance of the univariate coefficient. The advantage of considering the projected slope is that in most situations the variance is smaller than the variance of the univariate coefficient derived from a simple regression of disease on the risk factor. Also, we are assuming a multivariate model and it makes more sense to refer to estimates from such a model. An additional advantage is that the projected slope provides a descriptive partitioning of the univariate coefficient into explanatory segments that describe how a risk factor is related to disease both directly and through relationships with other covariables. Notice in one of the examples above that the projected slope for x_1 was $\gamma_1 \beta_2 \neq 0$. This suggests that the magnitude of the univariate coefficient is solely attributed to the association of x_1 with x_2 and the relationship x_2 has with disease.

Example 1

In Table 1, the first example using Framingham data is given because it is instructive and indicates a very desirable property of the projected slope for potentially protecting against the overemphasis of a statistically significant multiple logistic coefficient. Individuals in this example are followed for 26 years beginning around 1950 for the development of CHD. The predicting variables of interest are height and weight. The significance of the multivariate coefficient for height suggests that for a given weight, tall people have a reduced chance of developing disease. If nothing were known about the relationship between height and weight, one might conclude that height is an independent contributor to CHD. If height and weight were unrelated this would be true. Height and weight, however, have a correlation of 0.276 ($p<0.05$) so that for a given weight taller people are leaner, and it is not height that effects CHD but the whole concept of leaness; i.e., height and weight considered together. In this instance, one should be interested in the total contribution of height to CHD; i.e., a direct association, as well as well as the association of height to weight. Here, the multiple logistic coefficient for height is $\beta_1 = -0.098$, for weight the coefficient is $\beta_2 = 0.012$, and the slope of the regression between height and weight is $\gamma_1 = 2.892$. Thus, the projected slope is $\beta_1+\gamma_1\beta_2 = -0.063$ and it is not significant. It is clear from the form of the projected slope that the benefit of being tall (indicated by $\beta_1 = -0.098$) is reduced by the liability of increased weight that is associated with being tall (indicated by $\gamma_1\beta_2 = 0.035$), and that height is not a meaningful contributor of CHD by itself.

Example 2

Example 2 is similar to example 1 in terms of conclusions but is based on a more realistic application of multiple logistic analysis. Here, T-C, TG, HDL-C, MRW, systolic blood pressure (SEP), smoking, and age are examined in our Framingham sample as risk factors for CHD with follow-up of subjects beginning around 1972 and lasting about 6 years. There is some belief that in older age groups, such as that depicted by our sample, the relationship between T-C and CHD is weaker than it is among younger individuals (18). In our example, the univariate coefficient for T-C is consistent with this hypothesis since it is not significant. In contrast, the multiple logistic regression coefficient for T-C is significant. The latter implies that for given levels of the covariables, high levels of T-C significantly increase the chance of developing CHD. This interpretation, however, is misleading among our older sample because differences in T-C are commonly accompanied by differences in the other covariables.

The projected slope helps describe a more comprehensive relationship between T-C and CHD. From the Appendix, a general expression for the projected slope of a variable x_1 when there are p covariables is $\beta_1+\gamma_{21}\beta_2+\gamma_{31}\beta_3+\ldots\gamma_{p1}\beta_p$. Here, β_k is the multiple logistic regression coefficient for the kth variable. The coefficient γ_{k1} is the slope coefficient for x_k regressed on x_1. For this example, we take x_1=T-C, x_2=TG, x_3=HDL-C, x_4=MRW, x_5 = SBP, x_6 = smoking status, and x_7=age. The respective estimates for γ_{k1}, k=2,3,..., 7, are 0.003, 0.030, 0.023, -0.005, 0.000, and -0.014. The respective estimates for β_k, k=1,2,..., 7 are 0.006, -0.261, -0.047, 0.006, 0.008, 0.216, and 0.029. Thus, the projected slope is 0.003 and more in line with what is implied by the univariate coefficient and what is expected.

Among the covariables, it turns out that HDL-C is the most consistent predictor of CHD ($p<0.05$) and acts on CHD in a protective fashion. HDL-C is also correlated with T-C. The correlation is 0.092 ($p<0.05$). It would seem that since high levels of T-C are accompanied by elevated and protective levels of HDL-C that the effect of T-C on CHD should be diminished. If we look at the contribution to the projected slope by the relationship between T-C and HDL-C and the association HDL-C has with CHD, we see that the misleading magnitude of the multiple logistic regression coefficient associated with T-C (represented by $\beta_1=0.006$) is partially reduced by an amount equal to $\gamma_{31}\beta_3$. This reduction suggests that the liability of possessing higher levels of T-C are mitigated by the likely presence and beneficial effects of also possessing elevated levels of HDL-C. Here, it is the joint contribution between HDL-C and T-C that is important and clearly taken into account by the projected slope. Of course, relationships among the other covariables and CHD also influence interpretation of the projected slope. These relationships, however, exist to a much lesser degree and describing them would be superfluous.

Example 3

We have to this point given examples indicating a useful property of the projected slope in interpreting the predictive ability of a risk factor when its multiple logistic coefficient is statistically significant. The projected slope, however, also has the property of potentially protecting against the unwarranted underemphasis of a statistically insignificant multiple logistic coefficient as will be shown in example 3 using the Framingham data with a similar length of follow-up as example 2.

The third example was motivated by a paper (14) that questioned the

relationship of TG with CHD. The paper highlighted studies based on multiple logistic regression models that indicated that TG is an insignificant independent predictor of CHD. The paper concluded that the treatment of hypertriglyceridemia to alter the chance of developing CHD may be ineffective. The result was deemed important by the lay press and prompted close examination of the issue at a workshop on hypertriglyceridemia where some of the cautions and perspectives given in this paper were presented (19).

In the third example, the univariate coefficient for TG is significant, but when HDL-C, T-C, and MRW are included as covariates, the significance is reduced. In fact, the magnitude of the multivariate coefficient has become so distorted as to be negative. This finding, although enigmatic at first, is largely attributed to the strong correlation between TG and each of the covariables ($p<0.05$). The correlation of TG with HDL-C was given earlier and is -0.451. The correlations of TG with T-C and MRW are 0.276 and 0.227, respectively. The direct interpretation of the multiple logistic regression coefficient implies that for fixed levels of the covariables, changes in TG do not affect CHD. But, on the average, differences in TG are often accompanied by differences in all the covariables. At least one of these covariables, HDL-C, exhibits a strong relationship with CHD ($p<0.05$).

To improve our understanding of the relationship of TG with CHD we again compute the projected slope using the notation in the Appendix. We first need the slope coefficients of HDL-C, T-C, and MRW regressed on TG. These values are, respectively, $\gamma_{21}=-11.855$, $\gamma_{31}=22.062$, and $\gamma_{41}=7.267$. We also need the corresponding multiple logistic regression coefficients for TG, HDL-C, T-C, and MRW. These were given earlier in the Introduction. The estimate of the projected slope is then $\beta_1+\gamma_{21}\beta_2+\gamma_{31}\beta_3+\gamma_{41}\beta_4=0.511$, more in line with a positive association between TG and CHD that is commonly expected. The implication is

that the physiologic relationships between TG and the covariables have changed the magnitude of the importance of TG in a multivariate setting. Nevertheless the total contribution of TG to CHD that includes a direct association and an indirect relationship with CHD through the covariables, and especially HDL-C, may still be important. This is clearly represented by the projected slope. Here, the projected slope, which is significant ($p<0.05$), suggests that if observational data are useful for making clinical decisions that altering TG may be an effective means of changing the risk to CHD.

Conclusion

In the investigation of an association between a characteristic and disease, it is important to consider not just significance of a multiple logistic regression coefficient, but the total contribution that a characteristic has on the development of disease. These contributions include those that are direct and those that are shared among relationships with other characteristics. If this is not the interest, then to isolate and understand the effect of a characteristic on CHD when it could be one of several interacting components participating in a biological mechanism may be difficult.

The projected slope is used as a means to help show that the magnitude of the multiple logistic coefficient is often difficult to interpret. The projected slope is meant to offer explanation and insight into the importance of a significant univariate coefficient and why a multivariate coefficient has or has not achieved significance by way of relationships through the covariates included in a logistic expression.

In example 1, the projected slope provided a comprehensive perspective that

helped explain an important relationship between height and CHD. In the second example, we discovered how the multiple logistic coefficient for T-C can be reduced, when among older individuals, elevated T-C may increase the capacity to carry cholesterol in the high density lipoprotein class resulting in a diminished association between T-C and CHD. In both of these examples, the projected slope has not only improved our perspective of disease causality, but it has also protected us against the overemphasis of a statistically significant multiple logistic regression coefficient. Furthermore, in example 3, the projected slope has also shown how it can protect against the unwarranted underemphasis of a statistically insignificant multiple logistic regression coefficient. Here, TG has the potential for being thought of as an innocuous lipid marginally related to disease. TG, however, is related to HDL-C, the latter of which strongly influences the chance of developing CHD. Unless this relationship is taken into account as it is by the projected slope, the effect of TG on CHD will not be understood and the benefits of reducing elevated levels of TG will not be appreciated.

This presentation has shown that the magnitude of the multiple logistic regression coefficient is uncertain when other variables with a close physiologic relationship are included in the multiple logistic expression and that awareness of this possibility is important. Furthermore, attempting to isolate independent contributions to disease by examining the magnitude of the multiple logistic coefficient may be misleading because of the confounding influences shared among covariates. It may also be the case that these latter influences and their relationships with a risk factor may define a metabolic system that should not be broken down into components, but instead, considered in its entirety.

It is apparent that even if it were realistic to isolate risk factors, that

to properly assess their independent contribution to disease would require that enough observations on the risk factor be observed across all levels of the other risk factors. This is often the goal of controlled clinical trials but rarely ever occurs in nonrandomized or observational studies. It is clear that if we have insufficient data on a variable for all levels of the other variables that we will lack the evidence to investigate the first two conclusions of the Introduction. Indeed, the relationship of TG to CHD may be partially explainable by HDL-C, but we lack the data to say that TG has an unimportant direct relationship with CHD. Furthermore, if we do mistakenly assume that the first two conclusions are true, we certainly cannot assume that the last conclusion is also true. This is most evident in our example on TG where changes in TG affect the chance of developing CHD.

The examples we have presented show clearly that the projected slope is a useful device when used as a supplement for multiple logistic regression in prospective observational studies. With standard computer packages, it is easy to calculate and test. We believe the projected slope, similar as it is to the well known area of path analysis, is intuitively easy to understand. While it is certainly not the only way to deal with confounding and multicollinearity, the projected slope is a useful tool for understanding important causal relationships between risk factors and disease.

APPENDIX

When independent variables are correlated in linear regression, estimating parameters and reducing the error associated with the estimates is often accomplished by using Stein estimates or ridge regression (20). Such methods to date are not readily available for logistic regression and our interest is in estimating not one parameter but a function of parameters. In this paper, the test statistic for the significance of the projected slope is approximated by hypothesizing a joint model for the ith class of individuals that share the same values of x_{1i} and x_{2i};

[3] $\quad x_{2i} = \gamma_0+\gamma_1 x_{1i}+\varepsilon_i$, and conditionally on x_{2i}

[4] $\quad \text{logit}[p_i(\text{disease})] = \beta_0+\beta_1 x_{1i}+\beta_2 x_{2i}$.

Model [3] represents the linear relationship between x_1 and x_2 illustrated in Figure 2. The unconditional expectation of the log odds ratio for any two individuals in classes i and j falling on the regression line between x_1 and x_2 is:

$$\text{logit}[p_i(\text{disease})] - \text{logit}\,[p_j(\text{disease})] = (\beta_1+\gamma_1\beta_2)(x_{1i}-x_{1j})$$

In this expression, x_{2i} is replaced by its expectation, $\gamma_0+\gamma_1 x_{1i}$.

A test of significance of the projected slope is equivalent to testing H_o: $\beta_1+\gamma_1\beta_2=0$. While the obvious estimates of β_1, β_2, and γ_1 can be used to estimate $\beta_1+\gamma_1\beta_2$, the following informal analysis is useful for computing a test statistic for H_o. In order to test H_o, model [4] is rewritten with x_{2i} replaced with $\gamma_0+\gamma_1 x_{1i}+\varepsilon_1$ to give the following model.

[5] $\quad \text{logit}[p_i(\text{disease})] = \delta_0+\delta_1 x_{1i}+\delta_2\varepsilon_i$

Here, $\quad \delta_0 = \beta_0+\gamma_0\beta_2$,

$$\delta_1 = \beta_1 + \gamma_1\beta_2 \quad , \text{ and}$$

$$\delta_2 = \beta_2 \ .$$

Thus, an equivalent hypothesis is H_o: $\delta_1 = 0$.

To test H_o, the coefficients of model [5] are estimated by regressing the estimated class logits, $y_i = \text{logit}[\hat{p}_i(\text{disease})]$, on x_{1i} and e_i, where $e_i = x_{2i} - \hat{\gamma}_0 - \hat{\gamma}_1 x_{1i}$, and $\hat{\gamma}_0$ and $\hat{\gamma}_1$ are the ordinary least squares estimates of γ_0 and γ_1. Here, iterative weighted least squares (2) is used to estimate δ_1, δ_2, and δ_3.

Exact computations based on the usual linear model suggest that the estimate, $\hat{\delta}_1$, of δ_1 is approximately unbiased (i.e., consistent and asymptotically normal) for $\beta_1 + \gamma_1\beta_2$, but the estimated variance of $\hat{\delta}_1$ underestimates the true variance by a factor of $\beta_2^2\sigma_\varepsilon^2$. Here, β_2 is from model [4] and σ_ε^2 is the variance of ε_i from model [3].

The magnitude of $\beta_2^2\sigma_\varepsilon^2$ is negligible, however, in the examples considered in this paper, primarily because the value of β_2 is frequently much less than one making the contribution of β_2^2 small. Comparisons were made with bootstrap methods of estimation (21), however, which give improved estimates of the variance of $\hat{\delta}_1$ and indicate that ignoring $\beta_2^2\sigma_\varepsilon^2$ does not appreciably alter statistical results provided by the simpler estimation procedure given above.

The technique used here is also easily extended to the case when several independent variables are modeled in a multiple logistic equation. In this instance, if $x_1, x_2, \ldots, x_p$ are antecedent characteristics, the test of the projected slope (which becomes projected in a hyperplane), can be written as H_o: $\beta_1 + \gamma_{21}\beta_2 + \gamma_{31}\beta_3 + \ldots, \gamma_{p1}\beta_p = 0$. Here, γ_{k1} is the slope coefficient appearing in the following model for the ith class of individuals.

$$x_{ki} = \gamma_{k0} + \gamma_{k1}x_{1i} + \varepsilon_{ki}$$

The test of H_o is then extended by regressing the estimated class logits, $y_i=\text{logit}[p_i(\text{disease})]$, on x_{1i}, e_{2i}, e_{3i},..., e_{pi}, where $e_{ki}=x_{ki}-\hat{\gamma}_{k0}-\hat{\gamma}_{k1}x_{1i}$, and $\hat{\gamma}_{k0}$ and $\hat{\gamma}_{k1}$ are the least squares estimates of γ_{k0} and γ_{k1}. As when only two antecedent characteristics are included in the logistic model, the test for H_o is equivalent to testing the logistic coefficient associated with x_1 in the reparameterized analog to model [5].

REFERENCES

1. Truett, J., Cornfield, J., and Kannel, W. "A multivariate analysis of risk of coronary heart disease in Framingham," Journal of Chronic Diseases, 20, 511-524 (1967).

2. Walker, S. and Duncan, D. "Estimation of the probability of an event as a function of several independent variables," Biometrika, 54, 167-179 (1967).

3. Gordon, T. and Kannel, W. Introduction and General Background in the Framingham Study - The Framingham Study, Sections 1 and 2, Bethesda, Maryland, National Heart, Lung, and Blood Institute, 1968.

4. Brownlee, K. Statistical Theory and Methodology in Science and Engineering, John Wiley and Sons, New York, 1965.

5. Gordon, T. "Hazards in the use of the logistic function with special reference to data from prospective cardiovascular studies," Journal of Chronic Diseases, 27, 97-102, 1974.

6. Mosteller, F. and Tukey, J. Data Analysis and Regression, Addison-Wesley, Reading, Massachusetts, 1977.

7. Neter, J. and Wasserman, W. Applied Linear Statistical Models, Irwin, Homewood, IL, 1974.

8. Abbott, R., Garrison, R., Wilson, F., and Castelli, W. "Coronary heart disease risk: The importance of joint relationships among cholesterol levels in individual lipoprotein classes, "Preventive Medicine, 11, 109-119, (1982).

9. Kempthorne, O., An Introduction to Genetic Statistics, Iowa State University Press, Ames, Iowa, 1969.

10. Gordon, T., Castelli, W., Hjortland, M., Kannel, W., and Dawber, T. "High density lipoprotein as a protective factor against coronary heart disease," American Journal of Medicine, 2, 707-714, (1977).

11. Gordon, T., Castelli, W., Hjortland, M., and Kannel, W., and Dawber, T. "Predicting coronary heart disease in middle-aged and older persons - The Framingham Study," Journal of the American Medical Association, 238, 497-499, (1977).

12. Gordon, T., Castelli, W., Hjortland, M., and Kannel, W. "The prediction of coronary heart disease by high density and other lipoproteins - An historical perspective," in Hyperlipidemia: Diagnosis and Therapy, Rifkind, E. and Levy, R. (Eds.), Grune and Stratton, New York, 1977.

13. Costas, R., Garcia-Palmieri, M., Nazario, E., and Sorlie, P. "Relation of lipids, weight and physical activity to incidence of coronary heart disease - The Puerto Rico Heart Study," American Journal of Cardiology, 42, 653-658, (1978).

14. Hulley, S., Rosenman, R., Bawol, R., and Brand, R. "Epidemiology as a guide to clinical decisions - The association between triglyceride and coronary heart disease," New England Journal of Medicine, 302, 1383-1389, (1980).

15. Wilson, P., Garrison, R., Castelli, W., Feinleib, M., McNamara, P., and Kannel, W. "Prevalence of coronary heart disease in the Framingham Offspring Study - Role of lipoprotein cholesterols," American Journal of Cardiology, 46, 649-654, (1980).

16. SAS Supplemental Library User's Guide, SAS Institute, Inc., Raleigh, N.C., 1980.

17. Dixon, W. and Brown, M. (Eds.), BMDP Biomedical Computer Programs P-Series, University of California Press, Los Angeles, 1979.

18. Gofman, J., Young, W., and Tandy, R. "Ischemic heart disease, atherosclerosis and longevity," Circulation, 34, 679-697, (1966).

19. Lippel, K., Tyroler, H., Eder, H., Gotto, A., and Vahouny, G. "Relationship of hypertriglyceridemia to atherosclerosis," Arteriosclerosis, 1, 406-417, (1982).

20. Dempster, A., Schatzoff, M., and Wermuth, N. "A simulation study of alternatives to ordinary least squares," Journal of the American Statistical Association, 72, 77-91, (1977).

21. Efron, B. "Bootstrap methods - Another look at the jackknife," The Annals of Statistics, 7, 1-26, (1979).

FIGURE LEGEND

Figure 1: The Probability of Disease, P(disease), Predicted from Uncorrelated Risk Factors, x_1 and x_2.

Figure 2: The Probability of Disease, P(disease), Predicted from Correlated Risk Factors, x_1 and x_2.

Table 1. Logistic Regression Coefficients and Projected Slopes for Selected Variables Used to Predict Coronary Heart Disease

Example	Coefficient Univariate	Coefficient Multivariate	Projected Slope	Characteristics[1] Variable	Characteristics[1] Covariance	Group[2]
1	-0.006	-0.098*	-0.063	Height	Weight	Females 35-44
2	0.003	0.006§	0.003	T-C	TG HDL-C MRW SBP Smoking Age	Males 50-80
3	0.437*	-0.183	0.511*	TG	HDL-C T-C MRW	Males 50-80

*$p<0.05$ §$p<0.10$

[1]HDL-C = high density lipoprotein cholesterol
MRW = Metropolitan relative weight
SBP = Systolic blood pressure
TG = Tryglyceride
T-C = Total cholesterol
Smoking status = yes or no

[2]Intervals denote observed range of ages

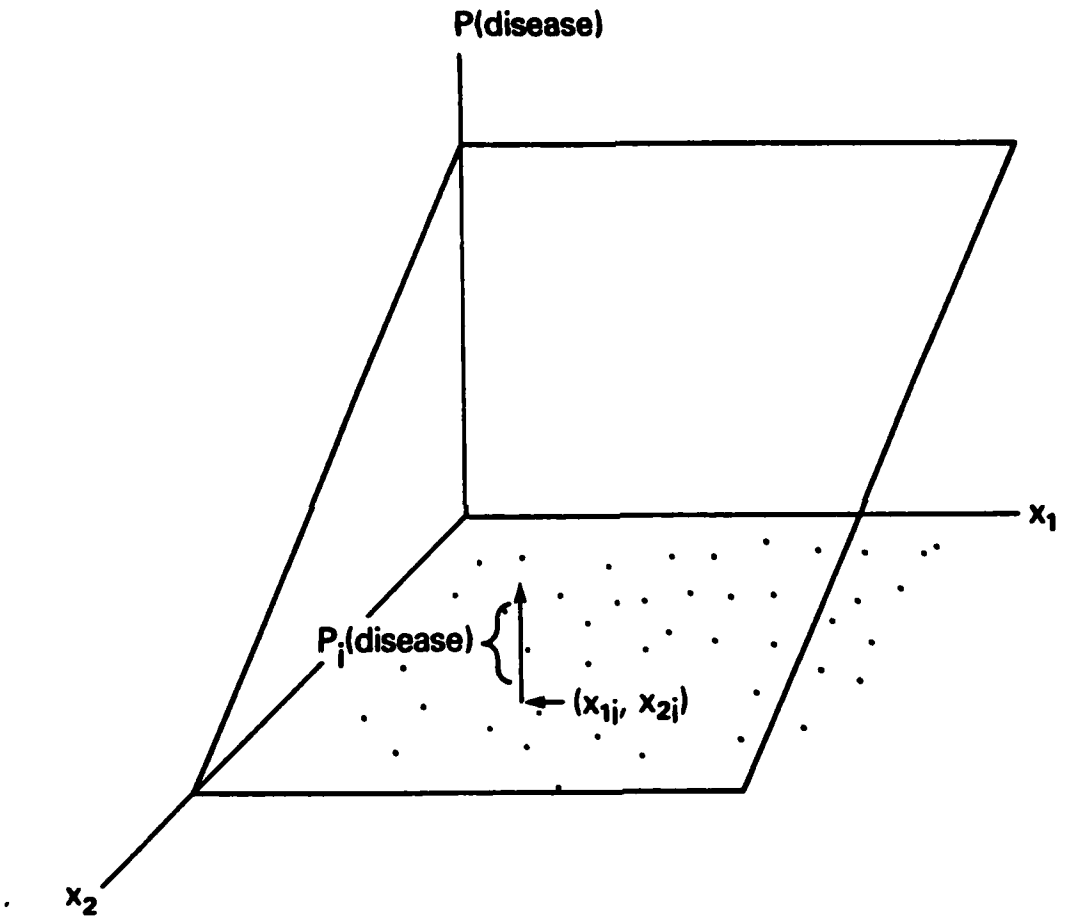
P(disease)
x_1
P_i(disease)
(x_{1i}, x_{2i})
x_2

Figure 1. Top

Robert D. Abbott and
Raymond J. Carroll

Interpreting Multiple Logistic
Regression Coefficients in Prospective
Observational Studies

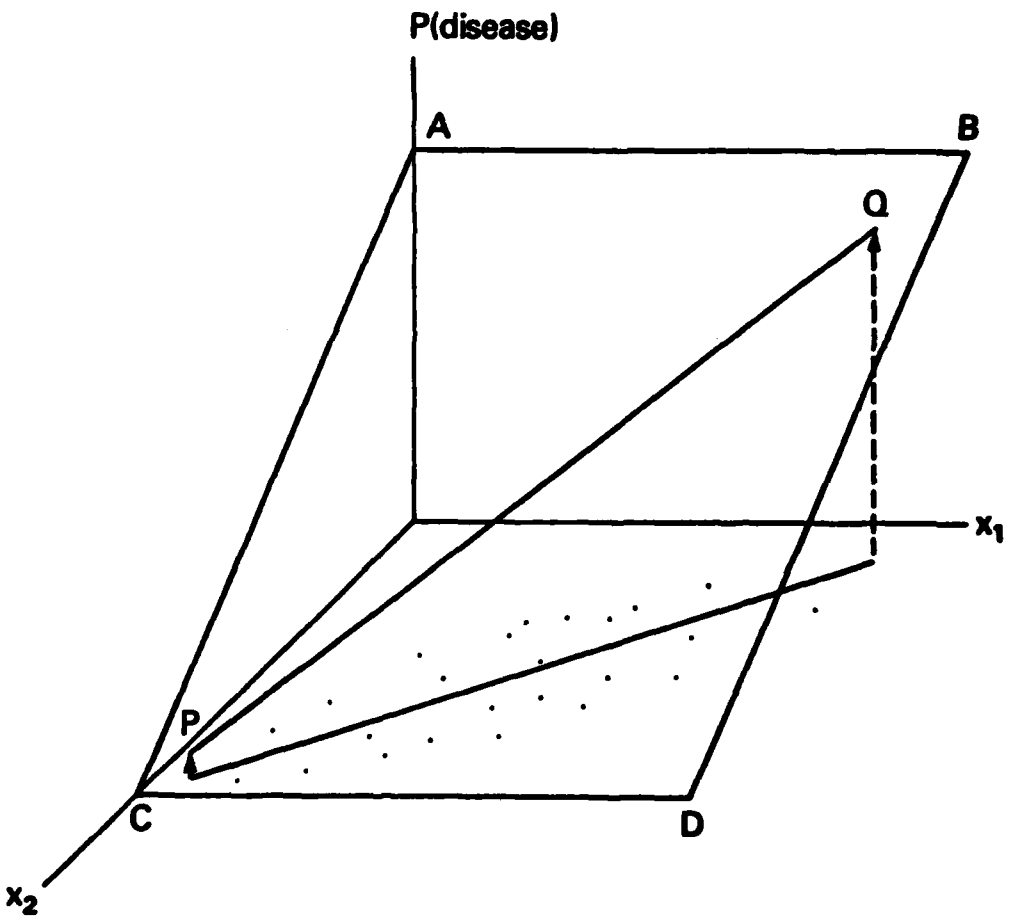
P(disease)
A
B
Q
x_1
P
C
D
x_2

Figure 2. Top

Robert D. Abbott and
Raymond J. Carroll

Interpreting Multiple Logistic
Regression Coefficients in Prospective
Observational Studies

SECURITY CLASSIFICATION [redacted] (When Data Entered) UNCLASSIFIED

REPORT DOCUMENTATION PAGE	READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER: AFOSR-TR- 82-0886 / 2. GOVT ACCESSION NO.: AD-A122694	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle): INTERPRETING MULTIPLE LOGISTIC REGRESSION COEFFICIENTS IN PROSPECTIVE OBSERVATIONAL STUDIES	5. TYPE OF REPORT & PERIOD COVERED: TECHNICAL
	6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s): R.D. Abbott and R.J. Carroll	8. CONTRACT OR GRANT NUMBER(s): F49620 82 C 0009
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Dept. of Statistics, University of N. Carolina, Chapel Hill, N.C. 27514	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: 61102F, 2304/A5
11. CONTROLLING OFFICE NAME AND ADDRESS: AFOSR, Bolling AFB, Washington, DC 20332	12. REPORT DATE: November 1982
	13. NUMBER OF PAGES: 27
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)	15. SECURITY CLASS. (of this report): UNCLASSIFIED
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release--distribution unlimited

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Multiple logistic regression, prospective observational studies, correlation, projected slope.

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Multiple logistic models are frequently used in observational studies to assess the contribution of risk factors to disease. In the presence of correlation among risk factors, the estimated magnitude of a multiple logistic coefficient can become uncertain or meaningless. This paper highlights the problem of interpreting a multiple logistic coefficient and suggests a procedure for examining the total contribuiton of a risk factor to disease that includes a direct association and associations that exist through relationships with other antecedent characteristics. Examples are given, along with results that are not immediately obvious

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

when considering the multiple logistic coefficient alone. Conclusions that are presented are important in biological studies if isolating the effect of an antecedent characteristic is unreasonable in the presence of confounding influences.