

Average predictive effects for models with nonlinearity, interactions, and variance components*

Andrew Gelman[†] Iain Pardoe[‡]

February 16, 2005

Abstract

In a predictive model, what is the expected change in the outcome associated with a unit change in one of the inputs? In a linear regression model without interactions, this *average predictive effect* is simply a regression coefficient (with associated uncertainty). In a model with nonlinearity or interactions, however, the average predictive effect in general depends on the values of the predictors. We consider various definitions based on averages over a population distribution of the predictors, and we compute standard errors based on uncertainty in model parameters. We illustrate with a study of criminal justice data for urban counties in the United States. The outcome of interest measures whether a convicted felon received a prison sentence rather than a jail or non-custodial sentence, with predictors available at both individual and county levels. We fit three models: a hierarchical logistic regression with varying coefficients for the within-county intercepts as well as for each individual predictor; a hierarchical model with varying intercepts only; and a non-hierarchical model that ignores the multilevel nature of the data. The regression coefficients have different interpretations for the different models; in contrast, the models can be compared directly using predictive effects. Furthermore, predictive effects clarify the interplay between the individual and county predictors for the hierarchical models, as well as illustrating the relative size of varying county effects.

Keywords: attributable risk, generalized linear model, hierarchical model, interactions, logistic regression, marginal model

1 Introduction

1.1 Inputs, predictors, and predictive effects

Interpreting coefficients in nonlinear models can be tricky. This is a problem since standard practice in applied statistics is to fit a model and then try to understand it by looking directly at parameter estimates. For example, logistic regression coefficients are interpretable as changes in “logit space” or “odds ratios,” but their more direct interpretation in terms of changes in probability depends on where you stand in the space of predictors. This is an issue even if the model is correct and arises even more dramatically when comparing models with different parametric forms.

*We thank John Carlin for several long discussions, Donald Rubin, Ross Stolzenberg and an anonymous reviewer for helpful comments, and the National Science Foundation for support through grants SBR-9708424, SES-9987748, and Young Investigator Award DMS-9796129.

[†]Department of Statistics, Columbia University, New York, USA, gelman@stat.columbia.edu

[‡]Charles H. Lundquist College of Business, University of Oregon, Eugene, USA, ipardoe@lcbmail.uoregon.edu

Assume you have fit a probability model $p(y|x, \theta)$ predicting a numerical outcome y based on vectors x of inputs and θ of parameters. We denote the separate sources of information in x that are used in the predictive model as *inputs* or *input variables*, as distinguished from the *linear predictors* that form the columns of the design matrix of a linear or generalized linear model. For example, consider a logistic regression of some individual outcome on age, sex, an age \times sex interaction, and age². There are two inputs (age and sex) but five linear predictors (including the constant term).

We often look to the regression coefficients in a predictive model, $p(y|x, \theta)$, to see how changing one of the inputs, x , (while holding the other inputs constant) affects the outcome, y . We call the resulting expected change in y a *predictive effect*. However, in a model with interactions, changing one of the inputs, with all others held constant, can alter several of the terms in the model. Another setting where coefficients can be difficult to interpret is multilevel modeling. In a linear multilevel model, one might use the standard deviation of a batch of random effects as a measure of their predictive effect on the outcome variable. But in a nonlinear model, once again this does not directly transfer to the scale of measurement.

Finally, when considering different models, sometimes using different transformations, it can be hard to directly compare coefficients between models so as to understand the true range of uncertainty indicated by the different possible model fits (see, e.g., Carroll and Ruppert, 1981, Hinkley and Runger, 1984, and Siqueira and Taylor, 1999).

1.2 Notation and basic definition of predictive effects

In this paper we develop a systematic approach for defining and estimating average predictive effects that should be appropriate in a wide range of applications.

In any model, $p(y|x, \theta)$, we consider the scalar inputs one at a time, using the notation,

u : the input of interest,
 v : all the other inputs.

Thus, $x = (u, v)$. We focus on the expected change in y per unit change in the input of interest, u , with v (the other components of x) held constant, i.e., the predictive effect of u changing from $u^{(1)}$ to $u^{(2)}$,

$$\text{predictive effect: } \delta_u(u^{(1)} \rightarrow u^{(2)}, v, \theta) = \frac{\text{E}(y|u^{(2)}, v, \theta) - \text{E}(y|u^{(1)}, v, \theta)}{u^{(2)} - u^{(1)}}. \quad (1)$$

(We use superscripts here to avoid confusion with the subscripting in the data $(x, y)_i, i = 1, \dots, n$.) We assume that $\text{E}(y|x, \theta)$ is a known function (e.g., the inverse logit) that can be computed directly. As we shall discuss in Section 2, this is *not* the same as the first-order partial derivative of the mean function with respect to u since there is no limit taken as $u^{(2)} - u^{(1)} \rightarrow 0$. For the logistic regression

model when $u^{(2)} - u^{(1)} = 1$, the predictive effect (1) is a “predicted change in probability” (see Petersen, 1985, Roncek, 1991, and Kaufman, 1996). For a linear model with no interactions, δ_u does not depend on $u^{(1)}$, $u^{(2)}$, or v , and is simply the regression coefficient associated with u . More generally, however, δ_u varies as a function of these inputs, and it can be useful to summarize the predicted difference with some sort of weighted average. In practice, one must also average over θ (or plug in a point estimate).

The approach recommended in this paper, of averaging $E(y|x, \theta)$ over a distribution for x , has been used in a number of applications and has been discussed in the statistical literature as well as in applied fields such as economics, medicine, psychology, and sociology; for example, Lee (1981) considers predictive effects for binary inputs to logistic regression models in epidemiology, and Graubard and Korn (1999) estimate population-average model predictions in a sample survey context. Lane and Nelder (1982) calculate average predicted values for generalized linear models (McCullagh and Nelder, 1989) in the context of “standardization,” which in our notation refers to different possible choices of the population distribution of v .

We go beyond this existing work by attempting to set up a general framework for predictive comparisons, including for models with interactions and variance components, and accounting for uncertainty in parameter estimates.

1.3 Outline of this paper

In Section 2 we discuss the strengths and weaknesses of some existing and alternative methods to defining and estimating average predictive effects. Section 3 defines our approach in general, considering averages of δ_u in (1) for numerical input variables, unordered categorical inputs, variance components (random or mixed effects), interactions, and other models. Section 4 discusses how to estimate average predictive effects from data, and Section 5 covers standard errors for the proposed estimates. Section 6 illustrates with an application to a study of criminal justice data for urban counties in the United States, for which several models were fit. We find average predictive effects to be a helpful adjunct to regression coefficients, both for understanding and for comparing models. Finally, we conclude with a brief discussion in Section 7.

2 Existing methods for summarizing predictive effects

2.1 Direct examination of regression coefficients

The most common summary of predictive effects is simply the set of estimated coefficients. These can be useful when directly interpretable; for example, the coefficients of a linear regression model without interactions are simply additive effects. In other cases the coefficients are harder to under-

stand. For example, logistic regression coefficients do not have direct probability interpretations, even when the model is appropriate to the data at hand. Sometimes a simple approximation can be useful (for example, if the predicted values are near $1/2$, then the logistic regression coefficient, divided by 4, is the approximate effect on the probability scale), but this technique can only work in a narrow range of the space of input variables. Experienced users of logistic regression, especially in epidemiology, can be comfortable enough with logits to interpret these coefficients directly as effects on log-odds, but many users—including ourselves—cannot so easily visualize odds ratios and their logarithms.

A completely different problem with regression coefficients is that, when interactions are present, the individual coefficients cannot be interpreted as predictive effects for individual inputs, *holding all other inputs constant*. With interactions, a single input variable can enter into several columns of the design matrix. The coefficients in a model with interactions may be of interest substantively and for understanding the model, but for predictive effects we are interested in changes in the inputs. For a very simple example, if age and age² are both included in a linear model, then the predictive effect for age will involve altering both the linear and quadratic terms. This sort of problem motivates a new approach, defining a predictive effect for each *input* rather than for each linear *predictor* (see Section 1.1).

Finally, regression coefficients can drastically change their meaning when a model is changed, whether by altering the functional form or by adding or removing predictors. Carlin et al. (2001) illustrate with a comparison of a semiparametric marginal model and two hierarchical logistic regression models fit to binary data: their “estimated marginal differences” (predictive effects defined at a point value; see below) are similar for the three models, while the estimated regression coefficients vary greatly, reflecting changes in parameterization more than real differences in the implications of the models.

Coefficient estimates are an important way to understand regression models, and we do not advocate abandoning them. Rather, we propose augmenting them with displays of a new measure—*average predictive effects*—which we define in Section 3.

2.2 Defining predictive effects at a central value

A rough approach that is sometimes used is to evaluate the regression mean function, $E(y|x)$, at a central value x_0 —perhaps the mean or the median of the data—and then estimate predictive effects by perturbing x_0 by altering the inputs one at a time. For binary inputs, this means evaluating the function at 0 and 1 (for example, Roncek, 1991, illustrates this method with a logistic regression of opinion differences among ethnic groups). In general, one must define some rule for choosing two

values, for example, the mean plus or minus one standard deviation (that is, ± 1 on a standardized scale).

Evaluating changes about a central value can work well in practice (for example, Gelman and King, 1993, and Gelman, King, and Liu, 1998, use this approach to summarize and compare logistic regression models with interactions) but one can run into problems when the space of inputs is very spread out (in which case no single central value can be representative) or if many of the inputs are binary or bimodal (in which case the concept of a “central value” is less meaningful, as is noted by Chang, Gelman, and Pagano, 1981). In addition, this approach is hard to automate since it requires choices about how to set up the range for each input variable. In fact, our work in this area was motivated by practical difficulties that can arise in trying to implement this central-value approach.

We illustrate some of the challenges in defining predictive effects with a simple hypothetical example of a logistic regression model of data y on a binary input of interest, u , and a continuous input variable, v .

Display Figure 1 here

The curves in each plot of Figure 1 show the assumed predictive relationship. In this example, u has a constant effect on the logit scale but, on the scale of $E(y)$, the predictive effect as u increases from 0 to 1 is high for v in the middle of the plotted range and low at the extremes. As a result, a “typical” (or average) predictive effect should depend on the distribution of the other input, v .

The two plots in Figure 1 show two examples in which an average predictive effect differs from the predictive effect evaluated at a central value. In the first plot, the data are at the extremes, so the average predictive effect is small—but the predictive effect evaluated at $E(v)$ is misleadingly large. The predictive effect on y corresponding to a realistic change in u is small, because switching u from 0 to 1 typically has the effect of switching $E(y|u, v)$ from, say, 0.02 to 0.05 or from 0.96 to 0.99. In contrast, the predictive effect if evaluated at the mean of the data is large, where switching u from 0 to 1 switches $E(y|u, v)$ from 0.36 to 0.60.

The second plot in Figure 1 shows a similar example, but where the centrally-computed predictive effect is too low compared to the population-average predictive effect. Here, the centrally-located value of v is already near the edge of the curve, at which point a change in u has little effect on $E(y|u, v)$, changing it by a mere 0.03. In comparison, the average predictive effect has the larger value of 0.11, which appropriately reflects that many of the sample data are in the range where a change in u can have a large effect on $E(y)$.

2.3 Partial derivatives

Partial derivatives provide another approach to calculating an expected change in outcome as a given input changes for models with transformed outcome variables, interactions, polynomials, and other nonlinearities and nonadditivities (see Stolzenberg, 1979, and Roncek, 1991). However, as noted by Petersen (1985) and DeMaris (1993) among others, the first-order partial derivative of a nonlinear function (such as a logistic regression mean function) with respect to a particular input variable is not the same as the predictive effect (as defined in Section 1.1) for a one-unit change in that input. To illustrate, consider predictive effects for v (rather than u) for the situation represented by the right plot of Figure 1. The predictive effect on y when v increases from 0 to 1 (and u is held constant at $u = 0$) is approximately $0.3 - 0.1 = 0.2$. In contrast, the partial derivative (i.e., the tangent line at $v = 0$) is much less than this (approximately 0.1). Even if the tangent line is moved to midway between $v = 0$ and $v = 1$, the slope of this tangent is still smaller than 0.2 due to the concavity of the logistic function here.

As this example illustrates, partial derivatives can be useful summaries of predictive effects for small changes in an input but less so with larger changes. Also, partial derivatives can run into trouble with discrete inputs where they may not be well-defined. Further, partial derivatives suffer from the same difficulty as the differencing discussed in Section 2.2, which is that there is no general way to pick a single point at which to evaluate the predictive effects.

2.4 Transformed coefficients

Transformations can sometimes clarify model interpretation; for example, the exponentiated coefficients of a log regression model are simply multiplicative effects. Similarly, as mentioned above, users of logistic regression models can transform the model coefficients to produce odds ratios that provide a measure of the multiplicative impact on the odds of a particular outcome for each unit increase in a given input (see also Long, 1987, and DeMaris, 1993). In contrast, the method we propose in this article automatically leads to interpretation on the original scale of the response variable, which in the case of logistic regression is the probability scale rather than odds. Without wishing to get too deep into the debate about which scale is easier to interpret, we merely note that in our experience probabilities are both more familiar and intuitive to work with than odds for many social scientists. One can always choose to summarize predictive effects on the transformed scale if that is deemed to be more interpretable.

2.5 Standardized coefficients

There is a long history of standardizing coefficients in regression models in an attempt to determine the “relative importance” of each input in affecting the outcome variable. This goal is controversial, however, and Firth (1998) provides a useful annotated bibliography of the scattered literature in this area. Common criticisms of comparing standardized coefficients include their dependence on the sample variation in the inputs, lack of inherent meaning for categorical inputs, and difficulties in dealing with input transformations and interactions. In practice, standardized coefficients can be useful in particular settings (for example, Long (1987) and Kaufman (1996) apply them to multinomial logit and logistic regression models) and to provide an automatic starting point for making coefficients roughly comparable. In any case, standardized coefficients do not directly address the problem of estimating predictive effects in the presence of nonlinearity and interactions.

3 General approach to defining population predictive effects

The basic predictive effect δ_u defined in (1) depends in general on $u^{(1)}$ and $u^{(2)}$ (the beginning and end points of the hypothesized change in the input of interest), v (the values of the other inputs), and θ (the parameters of the model). We define the *average predictive effect* Δ_u as the mean value of δ_u over some specified distribution of the inputs and parameters. Section 3.1 describes where these distributions come from, and then in Sections 3.2–3.7 we define predictive effects for various sorts of inputs u , starting with numerical input variables, including continuous and binary inputs as special cases, and moving to unordered categorical variables, random effects, interactions, and constraints.

It turns out that the form of the input of interest u , not the form of data y or other predictors v , is crucial in deciding how to define average predictive effects. We reiterate that in any application, we would compute the average predictive effect of each of the inputs to a model one at a time—that is, treating each component of x in turn as the “input of interest.” This is often a goal of regression modeling: estimating the predictive change in the outcome while changing one input variable, with all other inputs held constant.

3.1 Assumed existing data and model fit

We assume that a model $p(y|x, \theta)$ has already been fit to a data set $(x, y)_i, i = 1, \dots, n$, and our goal is numerically to summarize the predictive effect on y of each input in x , with inputs and predictive effects as defined in Section 1.2.

We further assume that inference about the parameters θ can be summarized by a set of simulation draws, $\theta^l, l = 1, \dots, L$ (with L set to some large value such as 100 or 1000) from a posterior

distribution or, in a non-Bayesian framework, from a distribution representing the estimate of θ and its uncertainty. If, as in the output to a typical generalized linear model program, only a point estimate and covariance matrix for θ are available, then we suppose that posterior draws have been obtained using the multivariate normal distribution with that mean and variance. Another option is to obtain the simulation draws from a bootstrap procedure applied to an estimator of θ (Efron and Tibshirani, 1993). For the purposes of this paper, it is not important where the simulation draws come from, only that they represent inferential uncertainty about θ .

3.2 Numerical inputs

For the purpose of estimation, it does not matter whether inputs are continuous or discrete (since the likelihood function is the same). However, for understanding the model it can make a difference. We start with numerical input variables u that can take on multiple values, including continuous inputs as a special case. There are no restrictions or assumptions about the other inputs, v , or the response, y . We average over $u^{(1)}, u^{(2)}, v$, and θ in both the numerator and denominator of (1), for all increasing transitions of u (that is, $u^{(1)} < u^{(2)}$):

$$\Delta_u = \frac{\int \int_{u^{(1)} < u^{(2)}} du^{(1)} du^{(2)} \int dv \int d\theta (E(y|u^{(2)}, v, \theta) - E(y|u^{(1)}, v, \theta)) p(u^{(1)}|v) p(u^{(2)}|v) p(v) p(\theta)}{\int \int_{u^{(1)} < u^{(2)}} du^{(1)} du^{(2)} \int dv \int d\theta (u^{(2)} - u^{(1)}) p(u^{(1)}|v) p(u^{(2)}|v) p(v) p(\theta)}. \quad (2)$$

This is equivalent to taking a weighted average of the δ_u 's in (1) with weights $(u^{(2)} - u^{(1)})$ —this makes sense from an estimation perspective since the estimates of predictive effects considered in Section 4 may be unstable for small values of $(u^{(2)} - u^{(1)})$. We only consider increasing transitions of u in Δ_u since otherwise the numerator and denominator of (2) would reduce to zero.

Display Figure 2 here

Averaging over $u^{(1)}, u^{(2)}$, and v in this way is equivalent to counting all pairs of transitions of $(u^{(1)}, v^{(1)})$ to $(u^{(2)}, v^{(2)})$ in which $v^{(1)} = v^{(2)}$ —that is, changes in u with v held constant. Figure 2 illustrates. The integral of θ averages over its distribution, which in a Bayesian context would be a posterior distribution, or classically could be a point estimate or an uncertainty distribution defined by simulations (as discussed at the end of the previous section). The distributions of (u, v) and θ are independent because we are working in a regression context in which θ represents the parameters of the model for y conditional on u, v . More generally, one could replace $p(u^{(1)}|v) p(u^{(2)}|v) p(v) p(\theta)$ by $p(u^{(1)}|v, \theta) p(u^{(2)}|v, \theta) p(v|\theta) p(\theta)$, in the multivariate modeling scenario in which θ is also relevant to the distribution of the regression inputs u, v .

In the special case in which u is a binary input, (2) reduces to a simple average of the differences $(E(y|u^{(2)}, v, \theta) - E(y|u^{(1)}, v, \theta))$. More generally, the average predictive effect has the form of a ratio of integrals.

3.3 Unordered categorical inputs

When an unordered categorical input variable u can take on more than two values, the average predictive effect—the expected change in y corresponding to a change in u —becomes more difficult to define. For instance, if u can take on the values “red,” “yellow,” and “blue,” then there are three possible changes in color, each with two possible directions. We here consider two natural ways to define predictive effects for general categorical input variables.

3.3.1 Separately considering each possible change in an input

One option is to define predictive effects separately for each possible pair of values $(u^{(1)}, u^{(2)})$ (in the above example, for the red-yellow, yellow-blue, and red-blue transitions). These are the average predictive effects (2), considering only two possible values of u at a time, and taking the convention that $u^{(2)} - u^{(1)} = 1$ in the denominator. Thus,

$$\Delta_u(u^{(1)} \rightarrow u^{(2)}) = \frac{\int \int_{u^{(1)} < u^{(2)}} du^{(1)} du^{(2)} \int dv \int d\theta (E(y|u^{(2)}, v, \theta) - E(y|u^{(1)}, v, \theta)) p(u^{(1)}|v) p(u^{(2)}|v) p(v) p(\theta)}{\int \int_{u^{(1)} < u^{(2)}} du^{(1)} du^{(2)} \int dv \int d\theta p(u^{(1)}|v) p(u^{(2)}|v) p(v) p(\theta)}. \quad (3)$$

3.3.2 Averaging over all possible transitions

Another option is to define a predictive effect that averages over all possible changes in the categorical input variable u . Such a definition may be more appealing in the context of this paper, where we seek to summarize the conditional impact of each variable in a predictive model. In averaging over changes in an unordered categorical input u , we must look at the magnitude, rather than the sign, of the effects. For example, if some input values have large positive effects and others have large negative effects, then we would want to say that this input has effects of large magnitude.

Once we decide to average over all possible transitions, we automatically lose all ordering (e.g., the sense that “red to yellow” is the opposite of “yellow to red”), so any averaging will have to take an absolute value. We shall follow common practice in statistics and work with the root mean square:

$$\Delta_u = \left(\frac{\sum_{u^{(1)}} \sum_{u^{(2)}} [\Delta_u(u^{(1)} \rightarrow u^{(2)})]^2 \int p(u^{(1)}|v) p(u^{(2)}|v) p(v) dv}{\sum_{u^{(1)}} \sum_{u^{(2)}} \int p(u^{(1)}|v) p(u^{(2)}|v) p(v) dv} \right)^{1/2}, \quad (4)$$

which weights each transition $u^{(1)} \rightarrow u^{(2)}$ in proportion to the probability of that pair in the distribution.

3.4 Variance components models

“Random effects” or “mixed effects” or “variance components” models are characterized by parameters in batches (see, e.g., Searle, Casella, and McCulloch, 1992, and Gelman, 2005). For example,

a model for longitudinal data, with several measurements on each individual, might have a varying intercept for each person, in which case the predictive effect is the expected change in y corresponding to a switch from one person to another. For the purpose of defining average predictive effects, we can treat a batch of K parameters $\phi_k, k = 1, \dots, K$, as an unordered categorical input variable with K levels $u^{(k)}$, in the sense of Section 3.3. The essence of a variance components model is that this batch of parameters ϕ_k are considered to have been drawn from a continuous population distribution. In our example in Section 6, we consider vector ϕ_k 's.

3.5 Models with interactions

The above definitions automatically apply to models with interactions. The key is that u represents a single input, and $x = (u, v)$ represents the vector of inputs to the predictive model. As discussed in Section 1.1, the vector of inputs (in the sense used in this paper) is not in general the same as the vector of linear predictors. For example, in the model in Section 1.1, sex is included on its own and also interacted with age. When defining the predictive effect for sex, we must alter this input wherever it occurs in the model—that is, both the “sex” predictor and the “sex \times age” predictor must be changed. For another example, the constant term in a regression is *not* an input in our sense and has no corresponding predictive effect, since it can take on only one possible value.

From a computational perspective, it is important that the model be coded in terms of its separate inputs. Thus, to compute predictive effects, it is not enough simply to specify the design matrix of a regression model; one must be able to evaluate $E(y)$ as a function of the original inputs.

3.6 Inputs that are not always active

A model will sometimes have inputs that are involved in prediction for only some of the data. For example, consider an experiment in which some units are given the control (no treatment) and others are given the treatment, in doses ranging from 10 to 20 (on some meaningful scale). Suppose the data are fit by a generalized linear model with treatment indicator, dose, and some pre-treatment measurements as predictors.

Now consider how to define the average predictive effect for dose. One approach is to consider treatment and dose to be a single input with value 0 for control units and the dose for treated units. This will not be appropriate, however, if we are particularly interested in the effect of dose in the range 10 to 20, conditional on treatment. We can define the predictive effect for dose as in Section 3.2, restricting all integrals over v to the subspace in which dose is defined (in this case, the treated units).

We can formally define the average predictive effect for a partially-active input u by introducing

a function $\zeta_u(v)$ that equals 1 wherever u is defined and 0 elsewhere. Then all the earlier definitions hold, as long as we insert the factor $\zeta_u(v)$ in all the integrals.

3.7 Nonmonotonic functions

It is possible in some applications for the predictive effects in (1) to be negative for some values of v and positive for other values of v such that the average predictive effect cancels out to zero. In such cases, it may be more appropriate to consider a root-mean-square predictive effect, as for categorical inputs in (4), or an average absolute predictive effect. Alternatively, variation in the predictive effects resulting from variation in v can be graphically displayed (see Pardoe and Shor, 2005).

4 Estimation

We now consider how to estimate average predictive effects from a finite sample of data $(u, v)_i, i = 1, \dots, n$, and a set of simulations $\theta^l, l = 1, \dots, L$. It is enough to figure out how to estimate (2) defined in Section 3.2 and (4) defined in Section 3.3.2; the average predictive effects defined in the rest of Section 3 are all special cases or combinations of these expressions.

4.1 Numerical inputs

The challenge in (2) is averaging over the product of densities $p(u^{(1)}|v)p(u^{(2)}|v)$. Averaging over $p(v)$ is simply attained by using the empirical data points v_1, \dots, v_n , and we can similarly average over $p(\theta)$ using $\theta^1, \dots, \theta^L$. The distribution of u given θ is necessary to appropriately average over all possible transitions, but it cannot be trivially estimated from a finite dataset.

We estimate (2) by the following ratio of sums:

$$\widehat{\Delta}_u = \frac{\sum_{i=1}^n \sum_{j=1}^n \sum_{l=1}^L w_{ij} (\mathbb{E}(y|u_j, v_i, \theta^l) - \mathbb{E}(y|u_i, v_i, \theta^l)) \text{sign}(u_j - u_i)}{\sum_{i=1}^n \sum_{j=1}^n \sum_{l=1}^L w_{ij} (u_j - u_i) \text{sign}(u_j - u_i)}, \quad (5)$$

where the factors w_{ij} are weights which we shall discuss shortly. The $\text{sign}(u_j - u_i)$ factors are used since we only consider increasing transitions of u in (2).

The summations over i, j , and l in (5) serve to average over the distributions of $(u^{(1)}, v)$, $u^{(2)}$, and θ . In the theoretical definition (2), transitions are from points $u^{(1)}$ to $u^{(2)}$ with a common v . For any given dataset, however, especially with continuous v , there may be few (if any) pairs of points with identical values of v . To approximate such exact transitions, we assign each pair of points with a weight,

$$w_{ij} = w(v_i, v_j),$$

which should reflect how likely it is for u to transition from u_i to u_j when $v = v_i$. Since, from the data, u_j occurs with v_j , w_{ij} should be maximized when $v_j = v_i$ and should in general have higher values when v_j is close to v_i . The goal is to approximate the distribution $p(u^{(2)}|v)$ in (2) by giving more weight to pairs of points in the data with values of v that are close to one another and less weight to pairs of points in the data with values of v that are far from one another.

As with any density estimation problem, the appropriate choice of weighting function w will depend on the space of v . Recall that u is scalar, but in general v will be a vector. If v lies in a continuous Euclidean space, we suggest, as a default, the following weighting function based on Mahalanobis distances:

$$w(v_i, v_j) = \frac{1}{1 + (v_i - v_j)^T \Sigma_v^{-1} (v_i - v_j)}.$$

This function will also work when v has one or more binary components. If v has some unordered discrete components, we would use a measure that penalizes components where v_i does not match v_j . If certain transitions are more likely than others, this could be included in the weighting function definition, although this is information that goes beyond the original predictive model.

For binary inputs u , (5) simplifies to:

$$\hat{\Delta}_u = \frac{\sum_{i=1}^n \sum_{l=1}^L [\sum_{j=1}^n w_{ij}] (\mathbb{E}(y|u = 1, v_i, \theta^l) - \mathbb{E}(y|u = 0, v_i, \theta^l))}{L \sum_{i=1}^n [\sum_{j=1}^n w_{ij}]}.$$
 (6)

Expression (6) is an estimate of the expected difference in y after changing u from 0 to 1 in a randomly-chosen unit in the population. This corresponds to an expected causal effect under a counterfactual assumption (Neyman, 1923, Rubin, 1974, 1990), if it makes sense to consider these inputs causally. We use the term *predictive effect* to emphasize that we are summarizing the structure of the predictive model and not necessarily estimating causal effects.

4.2 Unordered categorical inputs

For unordered categorical inputs u with K categories, we estimate (4) by the following:

$$\hat{\Delta}_u = \left(\frac{\sum_{i=1}^n \sum_{k=1}^K \sum_{l=1}^L [\sum_{j \in \{k\}} w_{ij}] (\mathbb{E}(y|u^{(k)}, v_i, \theta^l) - \mathbb{E}(y|u_i, v_i, \theta^l))^2}{L \sum_{i=1}^n \sum_{k=1}^K [\sum_{j \in \{k\}} w_{ij}]} \right)^{1/2},$$
 (7)

where $\sum_{j \in \{k\}} w_{ij}$ represents weights summed over the data points in category k . This estimate is also appropriate for variance components models, where for example there is a batch of K parameters $\phi_k, k = 1, \dots, K$ considered to have been drawn from a continuous population distribution, and we wish to estimate the predictive effect of switching from one group of measurements (represented by a value of k) to another group (represented by another value of k).

Finally, the actual evaluation of (5), (6), and (7) requires computing double and triple sums which could be overwhelming to estimate if n and L are large. Instead, we can approximate by replacing

each of the summations by a sum over a randomly-drawn subset of the indexes. For efficiency in simulation, it makes sense to use the same random draws for approximating the numerator and the denominator of (5). If there is a question about whether the subsets are large enough, the simulations can be repeated with different random subsets to see if the answers change substantially.

5 Standard errors

We recommend automatically computing average predictive effects for each of the inputs to a fitted model, and then displaying them in a table or graph along with their standard errors (for example, see Figure 3). This can provide a useful addition to the usual displays of regression coefficients. As discussed in Section 3.4, variance components in a multilevel model can be considered as unordered categorical inputs in this context.

In determining standard errors for estimated average predictive effects, we treat variation in the data points $x_i = (u, v)_i$ differently from the uncertainty expressed by the parameter simulations θ^l . Variation in x leads to variation in the predictive effects δ_u (except in the trivial case of linear models with no interactions), and we are averaging over this in estimating Δ_u . Pardoe and Shor (2005) consider graphical methods for displaying variation in δ_u that results from variation in x .

In contrast, we want uncertainty in θ to directly propagate to uncertainty in the average predictive effect. Thus, we are treating θ in the expressions for $\hat{\Delta}_u$ as random and the values of u and v as fixed. This is similar to inference for regression, where the standard errors derive from the distribution of the outcomes y , conditional on the inputs x . We can compute the standard errors using standard methods from sampling theory.

5.1 Numerical inputs

For general numerical inputs, estimate (5) can be expressed as $\hat{\Delta}_u = \frac{1}{L} \sum_{l=1}^L \hat{\Delta}_u^l$, where

$$\hat{\Delta}_u^l = \frac{\sum_{i=1}^n \sum_{j=1}^n w_{ij} (\mathbb{E}(y|u_j, v_i, \theta^l) - \mathbb{E}(y|u_i, v_i, \theta^l)) \text{sign}(u_j - u_i)}{\sum_{i=1}^n \sum_{j=1}^n w_{ij} (u_j - u_i) \text{sign}(u_j - u_i)}.$$

Then

$$\text{s.e.}(\hat{\Delta}_u) = \left(\frac{1}{L-1} \sum_{l=1}^L (\hat{\Delta}_u^l - \hat{\Delta}_u)^2 \right)^{1/2}. \quad (8)$$

Estimate (6) for binary inputs also takes this form, with

$$\hat{\Delta}_u^l = \frac{\sum_{i=1}^n [\sum_{j=1}^n w_{ij}] (\mathbb{E}(y|u = 1, v_i, \theta^l) - \mathbb{E}(y|u = 0, v_i, \theta^l))}{\sum_{i=1}^n [\sum_{j=1}^n w_{ij}]}$$

5.2 Unordered categorical inputs

When averaging over all possible transitions for unordered categorical inputs or considering variance components, expression (7) can be written as $\widehat{\Delta}_u = \sqrt{\frac{1}{L} \sum_{l=1}^L (\widehat{\Delta}_u^l)^2}$, where

$$(\widehat{\Delta}_u^l)^2 = \frac{\sum_{i=1}^n \sum_{k=1}^K [\sum_{j \in \{k\}} w_{ij}] (\mathbb{E}(y|u^{(k)}, v_i, \theta^l) - \mathbb{E}(y|u_i, v_i, \theta^l))^2}{\sum_{i=1}^n \sum_{k=1}^K [\sum_{j \in \{k\}} w_{ij}]}$$

We then estimate the standard error of the entire expression (7), including the square root, using a simple Taylor expansion: for any Z ,

$$\text{s.e.}(\sqrt{Z}) \approx \frac{1}{2} \text{s.e.}(Z) / \sqrt{Z}.$$

Then

$$\text{s.e.}(\widehat{\Delta}_u) \approx \frac{1}{2\widehat{\Delta}_u} \left(\frac{1}{L-1} \sum_{l=1}^L ((\widehat{\Delta}_u^l)^2 - \widehat{\Delta}_u^2)^2 \right)^{1/2}. \quad (9)$$

6 Application to a multilevel logistic regression of prison sentences

Pardoe and Weidner (2005) analyze sentencing of 8,446 convicted felons in 39 of the 75 most populous counties in the United States during May, 1998. They use a Bayesian multilevel logistic regression model with twelve individual-level variables from the State Court Processing Statistics (SCPS) program of the Bureau of Justice Statistics, linked to six county-level variables using the Federal Information Processing Standards code. Information collected in the SCPS program includes demographic characteristics, criminal history, details of pretrial processing, disposition, and sentencing of felony defendants.

The response variable for this study was “sentence severity,” defined as $y_{ij} = 1$ if the offender i in county j received a prison sentence, or 0 for a jail or non-custodial sentence (considered to be much less severe than prison). Under the model, the outcomes are independent with probabilities,

$$\Pr(Y_{ij} = 1) = \text{logit}^{-1}(\mathbf{X}_i^T \mathbf{G}_j \eta + \mathbf{X}_i^T \alpha_j), \quad (10)$$

where \mathbf{X}_i represents measurements on K individual-level variables and \mathbf{G}_j is a $K \times M$ block-diagonal matrix of measurements on L county-level variables. In particular, interactions between individual and county-level variables are used to account for dependence of individual-level effects across counties, so that M is $K \times L$ if all county-level variables are used to explain these individual-level effects, or of smaller dimension otherwise. The coefficients η in (10) represent main effects and interactions of the predictors and are constant across counties, while the coefficients α have a j -subscript and represent varying effects across counties, or can be viewed as interactions between the predictors X and the county indicators j .

6.1 Applying average predictive effects to a single model

The Bayesian model was fit with vague priors for η and α_j and using the software package Bugs (Spiegelhalter et al., 1994, 2004) to generate posterior samples; this free software enables Bayesian analysis of complex statistical models using Markov chain Monte Carlo techniques. Since nearly half the cases had some missing data, additional steps in the algorithm were used to impute missing values (Little and Rubin, 1987). Model checking diagnostics (see Pardoe, 2004) suggest that the multilevel model provides a much improved fit over a conventional non-hierarchical model. However, the presence of individual-county interactions and varying county effects complicates the interpretation of the η and α_j parameters. In contrast, average predictive effects are relatively straightforward to compute and provide a clear indication of the overall effects of each variable on the probability of receiving a prison sentence (rather than a jail or non-custodial sentence).

Display Table 1 here

Table 1 displays the estimated regression coefficients and standard errors, with additional entries giving the estimates and standard errors of the average predictive effects. As discussed at the beginning of Section 1.2, average predictive effects are defined for input variables rather than predictors and thus are not presented for the constant term, transformed variables, or interactions. The table also displays estimated standard deviations for variance components.

The regression coefficients and predictive effects in Table 1 serve two different purposes: the coefficients allow direct use of the model and can be interpreted on the logit scale, and the predictive effects summarize the importance of each input variable on the probability scale.

Display Figure 3 here

Figure 3 displays the average predictive effect for each variable in the model, together with a “random county” average predictive effect. Horizontal bars indicate ± 1 standard error for each average predictive effect, as calculated using the methods of Section 5. Due to computational limitations, we based all calculations on a randomly drawn subset of $L = 100$ posterior samples, with $n = 4500$ data points for the binary inputs, $n = 450$ for the continuous inputs, and $n = 4000$ for the varying county effects. Results varied little on repeating the calculations with different random subsets.

Individual-level variables in Figure 3 are denoted with an initial “I” and county-level variables are denoted with an initial “C.” The five individual-level variables for “most serious conviction charge” (ICVIOL1, ICTRAF, ICVIOL2, ICPROP, and ICDRUG) are relative to a reference category of weapons, driving-related, and other public order offenses. The twelve individual-level variables and

two of the county-level variables are binary, and so their average predictive effects were calculated using expression (6), with standard errors calculated using expression (8). The remaining four county-level variables are continuous, and so their average predictive effects were calculated using expression (5), with standard errors based on expression (8). Finally, the random county average predictive effect was calculated using expression (7), with standard error derived using expression (9).

The individual-level predictor with the largest effect on the probability of receiving a prison sentence is ICVIOL1 (murder, rape, or robbery), with an estimated average predictive effect of 0.36 (and standard error 0.02). That is, the expected difference in the probability of receiving a prison sentence between a randomly-chosen individual in the population charged with murder, rape, or robbery and a similar individual charged with a reference category offense is 0.36. Other charges appear less likely to result in a prison sentence, with decreasing probability: drug trafficking (with an estimated average predictive effect of 0.21), then assault (0.19), then property offenses (0.14), and finally drug possession offenses (0.05). Other individual-level variables can be interpreted similarly, as can the two binary county-level variables (which compare individuals in Southern and non-Southern counties, and individuals in counties with and without state sentencing guidelines).

Predictive effects for the continuous input variables correspond to one standard deviation changes. Standard deviations for the four variables (CCONS, CCRIME, CBLPCT, and CUNEMP) are 13.3%, 220 per 10,000, 12.4%, and 1.8% respectively. The positive average predictive effects for CCONS and CCRIME suggest that, comparing otherwise-similar cases, those in counties with higher conservative populations or higher crime rates have slightly higher probabilities of receiving a prison sentence. Conversely, the negative average predictive effect for CUNEMP suggests lessened sentence severity in high-crime counties, with all other inputs fixed. Taking into account all other factors, the proportion of the county's population that is African American (CBLPCT) has little bearing by itself on sentence severity. (However, this factor does play a role in reducing or increasing the effects of various individual-level variables; see Pardoe and Weidner, 2005.)

The random county average predictive effect differs from the other average predictive effects in this example in that it considers just the magnitude, rather than the sign, of the effects. This is because "county" is the only unordered categorical variable in the model. To understand its average predictive effect, consider the expected probabilities of receiving a prison sentence for two individuals who are identical in all respects except that they are in different counties. So, the two individuals will share the same values for individual-level variables but have different values for county-level variables (and county random effects). The random county average predictive effect of 0.32 represents the root mean square of the difference in the probability of receiving a prison

sentence between a randomly-chosen individual in one county and a similar individual in another county.

6.2 Applying average predictive effects to compare models

The hierarchical logistic regression model (10) has varying coefficients for the within-county intercepts as well as for each individual predictor. We also fit a hierarchical model with varying intercepts only, as well as a non-hierarchical model that ignores the multilevel nature of the data and excludes random effects. Whereas the regression coefficients have different interpretations for the different models, predictive effects allow for direct comparison.

Display Figure 4 here

Figure 4 displays the average predictive effect for each variable across all three models. The average predictive effects and standard errors are very similar across the two hierarchical models, perhaps suggesting that the additional variation for each individual predictor may be redundant. The individual-level effects are also very similar for the non-hierarchical model. However, the county-level effects tend to be smaller in magnitude and have smaller standard errors for the non-hierarchical model; the average predictive effect for unemployment, CUNEMP, even has the opposite sign. As demonstrated in Pardoe (2004), the non-hierarchical model fits this dataset poorly, and the average predictive effects displayed here suggest that while individual-level effects may be robust to model misspecification of this nature, higher-level effects can clearly be adversely affected.

7 Discussion

We recommend that predictive effects automatically be supplied with all fits of regression-type models that use a vector of inputs. A natural display would be a table or graph of average predictive effects, along with standard errors, for each predictive input and each batch of varying coefficients. For linear models with no interactions, this is identical to the coefficient estimates (for standardized predictors) and estimated variance components. For generalized linear models, there would seem to be no conceptual barrier to automating the computation and display of average predictive effects. For all-positive outcomes, it might also be appropriate to examine predictive effects on the logarithmic scale. For models with interactions, an additional step is required to isolate the vector of inputs from the vector of linear predictors. This could be done in parallel with summaries of marginal predictive effects (Pardoe, 2001, Pardoe and Cook, 2002).

It might be argued that predictive effects, like any other automatic summary of a model, cannot be universally applicable, because the best approach in any problem must be tailored to the specifics

of the application. We agree with this point, of course, but note that the overwhelming current practice in applied statistics of regression models is simply to report coefficient estimates (and standard errors), with no sense of their implications on the original scale of the data. We do not intend our approach to be a replacement for regression coefficients but rather a summary of predictive effects that can complement the coefficient estimates in order to make their scale more interpretable. Thus, we agree that there is no such thing as a “one size fits all” method—but that is what the current standard approach implicitly assumes. The “automaticity” of our approach has the important virtue that it can be used as an option in all sorts of problems, and thus it has a chance at being automatically implemented and used alongside regression coefficients to allow better understanding of predictive models.

Some applications naturally lend themselves to calculation of an average population effect, for example, in quantifying the average effect of a policy on a group of people (see Dehejia and Wahba, 1999, for a study of the effectiveness of job training programs). At the same time, one might be interested in explicit consideration of the functional relationship between predictive effects and values of inputs, where questions are often about input effects and how other inputs impact those effects. While this distinction lies beyond the scope of this article (which merely aims to complement the usual list of regression coefficients with a more meaningful summary of predictor effects), a companion article (Pardoe and Shor, 2005) considers this issue from a graphical perspective.

The example in Section 6 illustrates the effectiveness and convenience of predictive effects. In this multilevel dataset with a binary outcome measure, they clarify the overall role of each individual and group-level predictor in the presence of multiple interactions, as well as illustrate the relative size of the varying coefficients. They can also be used to understand and compare models directly, in a way that is difficult to do using logistic regression coefficients.

As discussed in Section 3, to define the average predictive effect for an input u , one must specify distributions for the values $u^{(1)}$ and $u^{(2)}$ for the comparison and the values of the other inputs v that will be held constant. This paper sets up a default structure based on using the data $\{x_1, \dots, x_n\}$ as an empirical distribution. But the idea of a predictive effect can be applied in settings where one does not wish to consider the observed data to be a simple random sample from a population. For example, the sample can be weighted to adjust for stratification or poststratification (see, e.g., Kish, 1965, or Gelman and Carlin, 2001, for a model-based perspective). Or an entirely different distribution can be chosen. Graubard and Korn (1999) discuss these possibilities in detail in the sample survey context. When a regression model has nonlinearity or interactions, predictive effects depend on the distribution of inputs for which the model will be applied.

References

- Carlin, J. B., Brown, C. H., Wolfe, R., and Gelman, A. (2001). A case study on the choice, interpretation and checking of multilevel models for longitudinal binary outcomes. *Biostatistics* **2**, 397–416.
- Carroll, R. J., and Ruppert, D. (1981). On prediction and the power transformation family. *Biometrika* **68**, 609–615.
- Chang, I. M., Gelman, R., and Pagano, M. (1982). Corrected group prognostic curves and summary statistics. *Journal of Chronic Diseases* **35**, 669–674.
- Dehejia, R., and Wahba, S. (1999). Causal effects in non-experimental studies: re-evaluating the evaluation of training programs. *Journal of the American Statistical Association* **94**, 1053–1062.
- DeMaris, A. (1993). Odds versus probabilities in logit equations: a reply to Roncek. *Social Forces* **71**, 1057–1065.
- Efron, B., and Tibshirani, R. (1993). *An Introduction to the Bootstrap*. New York: Chapman and Hall.
- Firth, D (1998). Annotated bibliography: Relative importance of explanatory variables. Maintained online at <http://www.nuff.ox.ac.uk/sociology/alcd/relimp.pdf> (By-product of a research project at Nuffield College, University of Oxford as part of the ESRC’s Analysis of Large and Complex Datasets programme - ESRC award number H519255044.)
- Gelman, A. (2005). Analysis of variance: why it is more important than ever (with discussion). *Annals of Statistics*, to appear.
- Gelman, A., and Carlin, J. B. (2001). Poststratification and weighting adjustments. In *Survey Nonresponse*, ed. R. Groves, D. Dillman, J. Eltinge, and R. Little.
- Gelman, A., and King, G. (1993). Why are American Presidential election campaign polls so variable when votes are so predictable? *British Journal of Political Science* **23**, 409–451.
- Gelman, A., King, G., and Liu, C. (1998). Not asked and not answered: multiple imputation for multiple surveys (with discussion). *Journal of the American Statistical Association* **93**, 846–874.
- Graubard, B. I., and Korn, E. L. (1999). Predictive margins with survey data. *Biometrics* **55**, 652–659.
- Hinkley, D. V., and Runger, G. (1984). The analysis of transformed data (with discussion). *Journal of the American Statistical Association* **79**, 302–320.
- Kaufman, R. L. (1996). Comparing effects in dichotomous logistic regression: a variety of standardized coefficients. *Social Science Quarterly* **77**, 90–109.

- Kish, L. (1965). *Survey Sampling*. New York: Wiley.
- Lane, P. W., and Nelder, J. A. (1982). Analysis of covariance and standardization as instances of prediction. *Biometrics* **38**, 613–621.
- Lee, J. (1981). Covariance adjustment of rates based on the multiple logistic regression model. *Journal of Chronic Diseases* **34**, 415–426.
- Little, R. J. A., and Rubin, D. B. (1987). *Statistical Analysis with Missing Data*. New York: Wiley.
- Long, J. S. (1987). A graphical method for the interpretation of multinomial logit analysis. *Sociological Methods and Research* **15**, 420–446.
- McCullagh, P., and Nelder, J. A. (1989). *Generalized Linear Models*, second edition. New York: Chapman & Hall.
- Neyman, J. (1923). On the application of probability theory to agricultural experiments. Essay on principles. Section 9. Translated and edited by D. M. Dabrowska and T. P. Speed. *Statistical Science* **5**, 463–480 (1990).
- Pardoe, I. (2001). A Bayesian sampling approach to regression model checking. *Journal of Computational and Graphical Statistics* **10**, 617–627.
- Pardoe, I. (2004). Model assessment plots for multilevel logistic regression. *Computational Statistics and Data Analysis* **46**, 295–307.
- Pardoe, I., and Cook, R. D. (2002). A graphical method for assessing the fit of a logistic regression model. *American Statistician* **56**, 263–272.
- Pardoe, I., and Shor, B. (2005). Graphical displays of predictive effects. Technical report, Lundquist College of Business, University of Oregon.
- Pardoe, I., and Weidner, R. R. (2005). Sentencing convicted felons in the United States: a Bayesian analysis using multilevel covariates (with discussion). *Journal of Statistical Planning and Inference*, to appear.
- Petersen, T. (1985). A comment on presenting results from logit and probit models. *American Sociological Review* **50**, 130–131.
- Roncek, D. W. (1991). Using logit coefficients to obtain the effects of independent variables in changes in probabilities. *Social Forces* **70**, 509–518.
- Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* **66**, 688–701.
- Rubin, D. B. (1990). Discussion of “On the application of probability theory to agricultural experiments. Essay on principles. Section 9,” by J. Neyman. *Statistical Science* **5**, 472–480.

- Searle, S. R., Casella, G., and McCulloch, C. E. (1992). *Variance Components*. New York: Wiley.
- Siqueira, A. L., and Taylor, J. M. G. (1999). Treatment effects in a logistic model involving the Box-Cox transformation. *Journal of the American Statistical Association* **94**, 240–246.
- Spiegelhalter, D., Thomas, A., Best, N., Gilks, W., and Lunn, D. (1994, 2004). BUGS: Bayesian inference using Gibbs sampling. MRC Biostatistics Unit, Cambridge, England.
www.mrc-bsu.cam.ac.uk/bugs/
- Stolzenberg, R. M. (1979). The measurement and decomposition of causal effects in nonlinear and nonadditive models. *Sociological Methodology* **11**, 459-488.

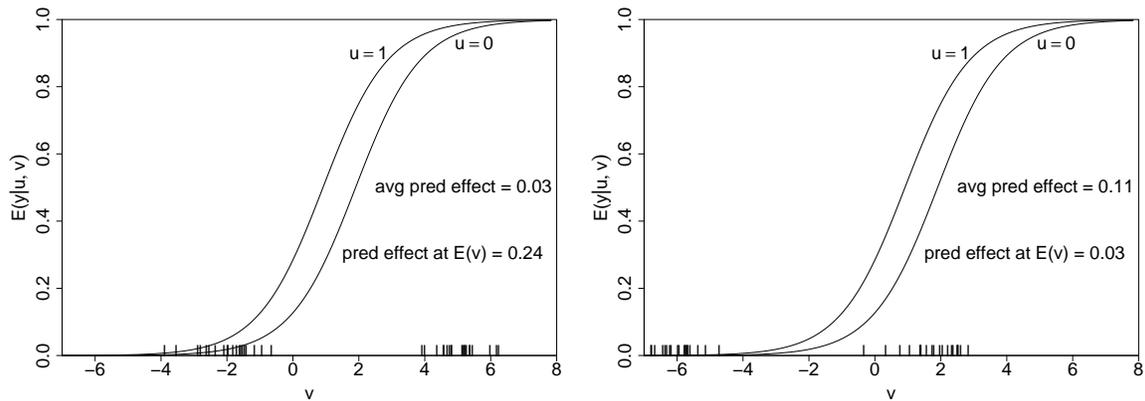


Figure 1: Hypothetical examples of predictive effects for an input in a model illustrating the need for averaging over the distribution of other inputs in the model (rather than simply working with a central value). Each graph shows a hypothesized logistic regression model for an outcome y given a binary input of interest, u , and a continuous input variable, v . The vertical lines on the x -axis indicate the values of v in the hypothetical dataset. (a) In the left plot, data v are concentrated near the ends of the predictive range. Hence, for each data value of v , the predictive effect on y as u increases from 0 to 1 is small; averaging over v produces a relatively small “average predictive effect.” In contrast, $E(v)$ is near the center of the range; hence the predictive effect at this average value is large, even though this is not appropriate for any of the data points individually. (b) Conversely, in the right plot, the average predictive effect is reasonably large, but this would not be seen if the predictive effect were evaluated at the average value of v .

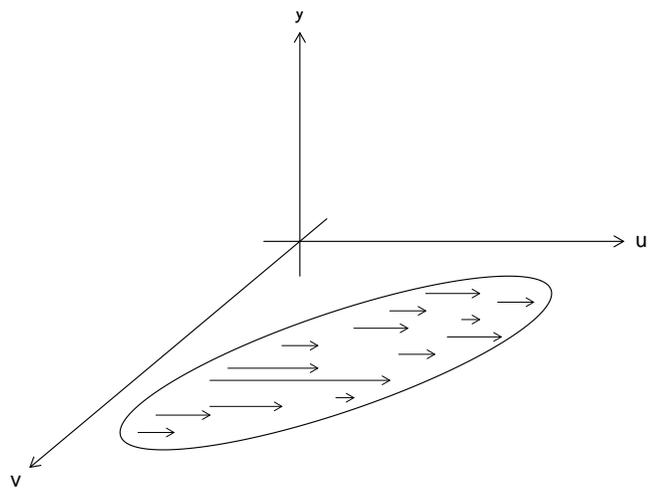


Figure 2: Diagram illustrating changes in the input of interest u , with the other inputs v held constant. The ellipse in (u, v) -space represents the joint distribution $p(u, v)$, and as the arrows indicate, we wish to consider changes in u in the region of support of this distribution.

Individual	County						sd(α)	$\hat{\Delta}_u$	
	CCRIME	CUNEMP	CPCTAA	CCONS	CSOUTH	CGUIDE			
	-5.2	0.4	-0.6	0.0	0.6	-0.7	0.1	1.2	
	(0.4)	(0.3)	(0.4)	(0.3)	(0.3)	(0.6)	(0.6)	(0.2)	
ICVIOL1	2.6	-0.1	-0.1	0.5	0.2	-0.0	0.0	0.6	0.36
	(0.3)	(0.2)	(0.3)	(0.2)	(0.3)	(0.3)	(0.4)	(0.3)	(0.02)
ICVIOL2	1.6	-0.3	0.2	0.4	0.1	0.5	-0.1	0.4	0.19
	(0.2)	(0.2)	(0.2)	(0.2)	(0.2)	(0.3)	(0.3)	(0.1)	(0.02)
ICTRAF	1.5	-0.0	-0.2	0.0	-0.0	-0.1	0.5	0.8	0.21
	(0.2)	(0.2)	(0.2)	(0.2)	(0.2)	(0.3)	(0.4)	(0.2)	(0.02)
ICDRUG	0.4	-0.3	0.2	0.4	0.3	0.1	0.2	0.7	0.05
	(0.2)	(0.2)	(0.3)	(0.2)	(0.2)	(0.3)	(0.4)	(0.2)	(0.01)
ICPROP	1.4	-0.2	-0.0	0.3	-0.1	-0.3	-0.7	0.7	0.14
	(0.2)	(0.2)	(0.2)	(0.2)	(0.2)	(0.3)	(0.3)	(0.2)	(0.01)
IPPRIS	1.7	-0.1	0.3	-0.3	-0.1	-0.2	-0.1	0.3	0.29
	(0.1)	(0.1)	(0.1)	(0.1)	(0.1)	(0.2)	(0.2)	(0.1)	(0.01)
ITRIAL	0.7	-0.2	-0.0	-0.2	0.1	0.2	0.1	0.4	0.09
	(0.2)	(0.2)	(0.2)	(0.2)	(0.2)	(0.3)	(0.3)	(0.2)	(0.02)
IMALE	0.5	0.0	-0.0	0.0	0.0	0.1		0.1	0.05
	(0.1)	(0.1)	(0.1)	(0.1)	(0.1)	(0.1)		(0.1)	(0.01)
IBLACK	0.0	0.2	-0.0	0.1	0.0	-0.0		0.1	0.01
	(0.1)	(0.1)	(0.1)	(0.1)	(0.1)	(0.1)		(0.1)	(0.01)
IACTCJS	0.8	0.0	0.0	-0.1	0.1	0.1		0.2	0.10
	(0.1)	(0.1)	(0.1)	(0.1)	(0.1)	(0.2)		(0.1)	(0.01)
IDETAİN	1.9	0.1	-0.2	-0.1	0.1	-0.5		0.4	0.23
	(0.1)	(0.2)	(0.2)	(0.1)	(0.2)	(0.2)		(0.1)	(0.01)
IREVOKE	1.3	0.1	0.2	-0.1	0.2	-0.2		0.5	0.16
	(0.2)	(0.2)	(0.2)	(0.2)	(0.2)	(0.3)		(0.2)	(0.01)
$\hat{\Delta}_u$		0.04	-0.05	0.01	0.08	-0.12	0.00	0.32	
		(0.03)	(0.03)	(0.03)	(0.03)	(0.07)	(0.06)	(0.05)	

Definitions of the inputs are provided in Figure 3.

Table 1: Estimated regression coefficients for the prison example, together with estimated standard deviations for variance components and estimates of the average predictive effects. (Standard errors are provided in parentheses.) The first row contains the regression coefficients for the county-level main effects, the first column contains the regression coefficients for the individual-level main effects, while the entries at intersections between county-level inputs and individual-level inputs are interactions. The second-to-last column contains estimated standard deviations for the variance components. The last row contains the average predictive effects for the county-level inputs, while the last column contains the average predictive effects for the individual-level inputs. Bold indicates that the absolute value of the estimate is larger than the standard error.

We do not in general recommend displaying the results in this sort of table, preferring graphs such as in Figures 3 and 4. However, we find this table helpful in explaining our method, by showing the information that is being summarized and compressed in defining average predictive effects in this example of a multilevel model with interactions.

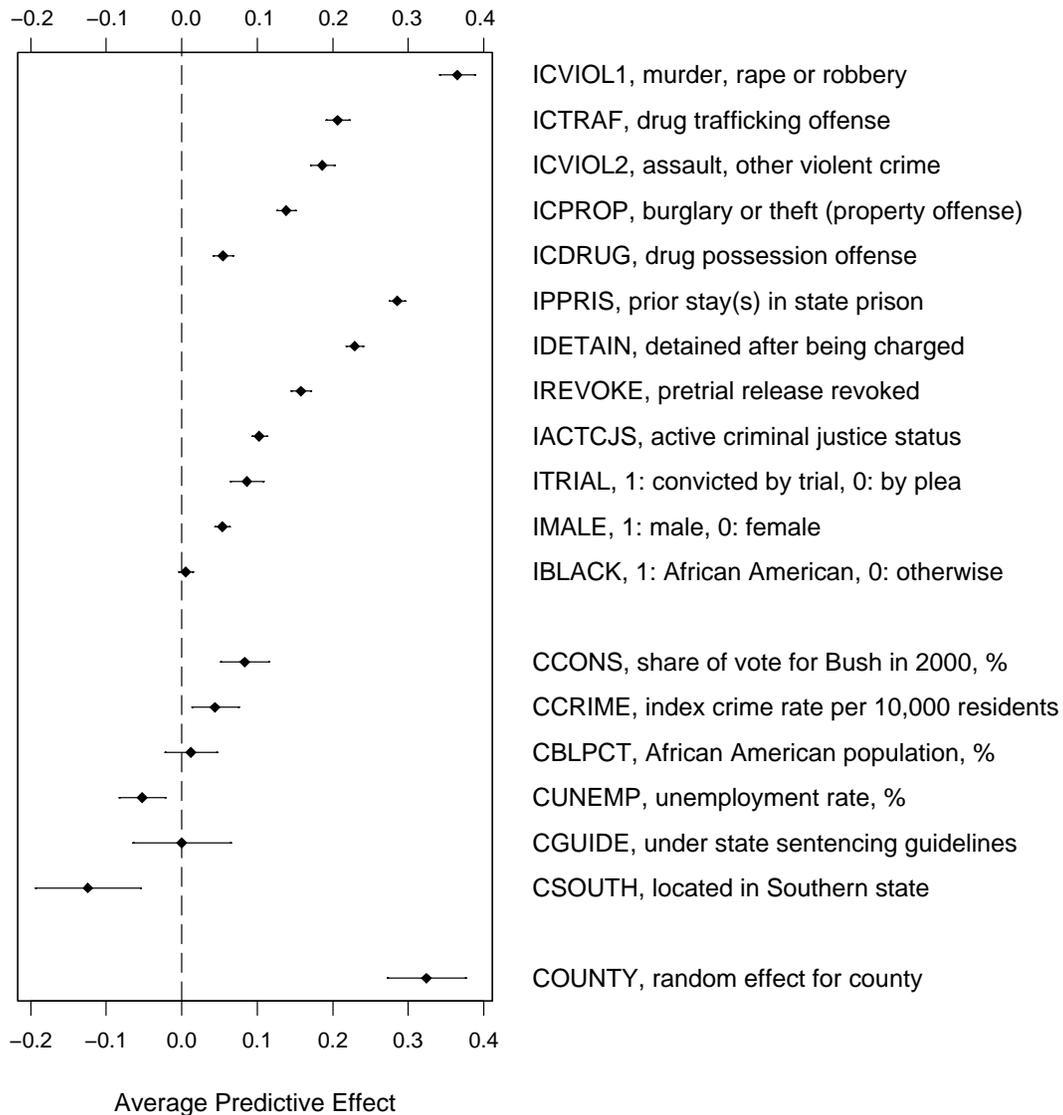


Figure 3: Estimated average predictive effects on the probability of a prison sentence (rather than a jail or non-custodial sentence), for each input variable in the prison example. Horizontal lines show ± 1 standard-error bounds. The first set of inputs, with initial letters I, are at the level of the individual offender; the second set, with initial letters C, are county-level inputs; and the last corresponds to the effect of varying the county random effect, keeping all other inputs constant. Many of the individual predictors have large effects and county itself predicts a fair amount of variability, but the county-level variables have relatively small effects (recall that these average predictive effects correspond to one standard deviation changes).

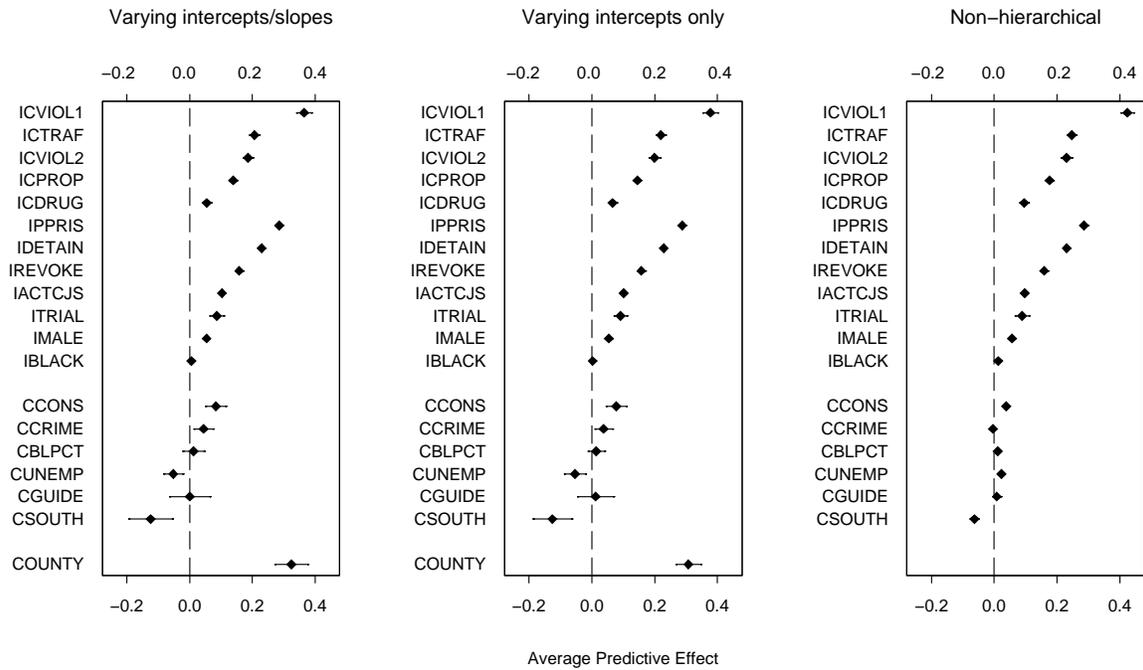


Figure 4: Estimated average predictive effects on the probability of a prison sentence for each input variable across three models in the prison example. Horizontal lines show ± 1 standard-error bounds. Effects and standard errors are very similar across the two hierarchical models. However, although the individual-level effects are similar for the non-hierarchical model, the county-level effects tend to be smaller in magnitude and have smaller standard errors.