Average predictive effects for models with nonlinearity, interactions, and variance components (presentation handout)*

Andrew Gelman† Iain Pardoe‡

8 September, 2004

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

*We thank John Carlin for several long discussions, Donald Rubin 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

1 Application: sentencing convicted felons in the U.S.

Response, Y, 1: prison sentence, 0: jail/non-custodial sentence. $n = 8{,}446$ convictions, 39 counties, 17 states (Bureau of Justice Statistics' State Court Processing Statistics, May 1998).

Figure 1: *Binary individual predictors: sample means, descriptions, and percent missing data (in parentheses). "Most serious conviction charge" (*ICVIOL1*,* ICVIOL2*,* ICTRAF*,* ICDRUG*, and* ICPROP*) is relative to a reference category of weapons, driving-related, and other public order offenses.*

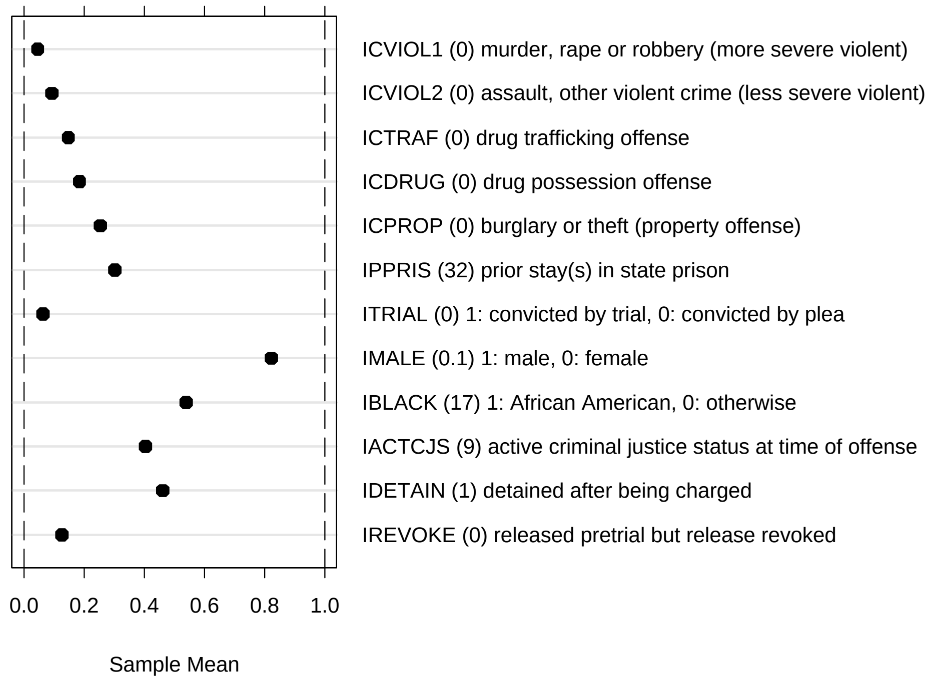


Table 1: *County-level predictors and summary statistics. Means and standard deviations are raw statistics (i.e. not population weighted) for 39 counties representing 24% of the U.S. population.*

Predictor	Description	Mean	S.D.	Min.	Max.
CCRIME	Index* (known to police) crime rate per 10,000 residents	587	220	214	1,095
CUNEMP	Unemployment rate (%)	4.4	1.8	2.3	10.0
CPCTAA	Census estimate of African American population (%)	18.9	12.4	1.8	45.9
CCONS	Share of vote for Bush in 2000 (%)	38.2	13.3	11.8	55.7
CSOUTH	1: located in a Southern state, 0: otherwise	0.28	-	0	1
CGUIDE	1: voluntary or mandatory state sentencing guidelines, 0: otherwise	0.23	-	0	1

* Index crimes include murder, rape, robbery, aggravated assault, burglary, larceny/theft, motor vehicle theft, and arson.

For the ith individual in county j, $\mathrm{Y}_{ij}|p_{ij} \sim \text{Bernoulli}(p_{ij})$, where $p_{ij} = \Pr(\mathrm{Y}_{ij} = 1)$, and

$$\text{logit}(p_{ij}) = \log\left(\frac{p_{ij}}{1 - p_{ij}}\right) = \boldsymbol{X}_i^T \boldsymbol{\beta}_j$$

where $\boldsymbol{X}_i$ is a vector of K individual predictors and $\boldsymbol{\beta}_j$ is a vector of K regression parameters (specific to the jth county). Across counties,

$$\boldsymbol{\beta}_j = \boldsymbol{G}_j \, \boldsymbol{\eta} + \boldsymbol{\alpha}_j$$

where $\boldsymbol{G}_j$ is a $K \times M$ block-diagonal matrix for L county-level predictors, $\boldsymbol{\eta}$ is a vector of M regression parameters, and $\boldsymbol{\alpha}_j$ is a $K \times 1$ vector of county-level errors. Combining,

$$\text{logit}(p_{ij}) = \boldsymbol{X}_i^T \boldsymbol{G}_j \, \boldsymbol{\eta} + \boldsymbol{X}_i^T \boldsymbol{\alpha}_j$$

Table 2: *Posterior summaries for $\boldsymbol{\eta}$: means (standard deviations). The first row contains the county-level main effects, the first column contains the individual-level main effects, while the remainder of the table contains interactions. Bold indicates that the absolute value of the posterior mean is larger than the posterior standard deviation.*

		County					
Individual		CCRIME	CUNEMP	CPCTAA	CCONS	CSOUTH	CGUIDE
	−5.9	0.2	**−0.6**	0.0	0.4	−0.6	0.4
	(0.5)	(0.5)	(0.5)	(0.4)	(0.5)	(1.0)	(0.7)
ICVIOL1	**2.9**	−0.3	−0.2	**0.8**	0.2	−0.2	−0.1
	(0.5)	(0.4)	(0.5)	(0.4)	(0.5)	(0.8)	(0.8)
ICVIOL2	**1.8**	**−0.6**	0.3	**0.5**	−0.1	**0.9**	−0.1
	(0.4)	(0.3)	(0.4)	(0.3)	(0.4)	(0.7)	(0.7)
ICTRAF	**1.6**	−0.2	−0.1	0.1	0.0	0.1	**0.8**
	(0.4)	(0.4)	(0.5)	(0.4)	(0.4)	(0.7)	(0.7)
ICDRUG	**0.5**	**−0.6**	0.3	**0.6**	0.3	0.3	0.4
	(0.4)	(0.4)	(0.5)	(0.4)	(0.4)	(0.8)	(0.7)
ICPROP	**1.7**	**−0.4**	0.1	**0.4**	−0.1	−0.4	**−1.0**
	(0.4)	(0.3)	(0.4)	(0.3)	(0.4)	(0.7)	(0.6)
IPPRIS	**1.9**	0.2	0.2	**−0.3**	0.1	**−0.7**	−0.2
	(0.3)	(0.3)	(0.3)	(0.2)	(0.3)	(0.5)	(0.5)
ITRIAL	**0.6**	−0.4	0.2	−0.3	0.1	0.7	0.5
	(0.5)	(0.4)	(0.6)	(0.4)	(0.5)	(0.7)	(0.8)
IMALE	**0.5**	−0.0	−0.1	−0.1	−0.0	0.4	
	(0.3)	(0.3)	(0.3)	(0.2)	(0.3)	(0.5)	
IBLACK	−0.1	**0.6**	−0.2	0.0	0.1	−0.1	
	(0.2)	(0.2)	(0.3)	(0.2)	(0.3)	(0.5)	
IACTCJS	**0.9**	0.1	−0.1	−0.1	0.2	−0.1	
	(0.2)	(0.2)	(0.3)	(0.2)	(0.3)	(0.5)	
IDETAIN	**2.2**	0.2	−0.2	−0.1	0.2	**−1.0**	
	(0.3)	(0.3)	(0.3)	(0.2)	(0.3)	(0.5)	
IREVOKE	**1.4**	0.3	0.3	−0.1	**0.5**	−0.6	
	(0.3)	(0.3)	(0.4)	(0.3)	(0.3)	(0.6)	

Further background is available in Pardoe and Weidner (2004).

2 Predictive effects

Gelman and Pardoe (2004) define the expected change in y per unit change in the input of interest, u, with v (the other components of x) held constant, as the predictive effect (PE) of u changing from $u^{(1)}$ to $u^{(2)}$:

$$\delta_u(u^{(1)} \to u^{(2)}, v, \theta) = \frac{\mathrm{E}(y|u^{(2)}, v, \theta) - \mathrm{E}(y|u^{(1)}, v, \theta)}{u^{(2)} - u^{(1)}}. \tag{1}$$

Related work includes Graubard and Korn (1999), Lane and Nelder (1992), Lee (1981), and McCullagh and Nelder (1989).

Consider $\delta_u(u_i, u_j, v_i, \theta^l)$, which represents the PE as u changes from u_i to u_j, with v held constant at v_i, evaluated at simulation draw θ^l. These PEs can be averaged over the model parameters θ and a distribution for x to define an average predictive effect (APE). Averaging over θ is straightforward using the set of L simulation draws, $\theta^l, l = 1, \dots, L$. To average over x we assign each pair of points in $\delta_u(u_i, u_j, v_i, \theta^l)$ 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. If v lies in a continuous Euclidean space, we suggest, as a default, the following weighting function based on Mahalanobis distances:

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

We can then use these weights in a weighted average of PEs calculated from the data and parameter simulations. The exact form of the resulting APE depends on the nature of the input u. We consider separately continuous, binary, and unordered categorical inputs.

2.1 Continuous inputs

If u is a continuous variable, then the APE is a weighted average of PEs:

$$\widehat{\Delta}_u = \frac{\sum_{i=1}^n \sum_{j=1}^n \sum_{l=1}^L w_{ij}\, \delta_u(u_i, u_j, v_i, \theta^l)\, |u_j - u_i|}{L \sum_{i=1}^n \sum_{j=1}^n w_{ij}\, |u_j - u_i|}, \tag{2}$$

where the PEs have been weighted by w_{ij} as described above, and also by $|u_j - u_i|$, since PEs with small values of $|u_j - u_i|$ will in general be more unstable (and less reliable as estimates) than PEs with larger values of $|u_j - u_i|$.

We can also define various intermediate averages, which can be useful in understanding how PEs vary with θ and with x. In particular, simulation predictive effects (SPEs) average over i and j only:

$$\widehat{\Delta}_u^l = \frac{\sum_{i=1}^n \sum_{j=1}^n w_{ij}\, \delta_u(u_i, u_j, v_i, \theta^l)\, |u_j - u_i|}{\sum_{i=1}^n \sum_{j=1}^n w_{ij}\, |u_j - u_i|}, \tag{3}$$

and represent an average predictive effect for each simulation draw $\theta^l, l = 1, \dots, L$.

We can also define $n \times n$ transition predictive effects (TPEs):

$$\widehat{\Delta}_u^{ij} = \frac{1}{L} \sum_{l=1}^L \delta_u(u_i, u_j, v_i, \theta^l), \tag{4}$$

which represent the (simulation averaged) expected change in y per unit change in u when u changes from u_i to u_j (and v stays constant). These TPEs can be averaged over j to give n IPEs:

$$\widehat{\Delta}_u^i = \frac{\sum_{j=1}^n \sum_{l=1}^L w_{ij}\, \delta_u(u_i, u_j, v_i, \theta^l)\, |u_j - u_i|}{L \sum_{j=1}^n w_{ij}\, |u_j - u_i|}, \tag{5}$$

which represent the (simulation averaged) expected change in y per unit change in u when u changes from u_i to u_j (and v stays constant), averaged over all possible transitions u_j (excluding to u_i itself).

The APE, (2), can then be considered as an average of L SPEs:

$$\widehat{\Delta}_u = \frac{1}{L} \sum_{l=1}^{L} \widehat{\Delta}_u^l,$$

or as a weighted average of n IPEs:

$$\widehat{\Delta}_u = \frac{\sum_{i=1}^{n} [\sum_{j=1}^{n} w_{ij} \, |u_j - u_i|] \, \widehat{\Delta}_u^i}{\sum_{i=1}^{n} [\sum_{j=1}^{n} w_{ij} \, |u_j - u_i|]}.$$

2.2 Binary inputs

If u is a binary variable, then the APE simplifies to:

$$\widehat{\Delta}_u = \frac{\sum_{i=1}^{n} \sum_{l=1}^{L} [\sum_{j=1}^{n} w_{ij}] \, \delta_u(0, 1, v_i, \theta^l)}{L \sum_{i=1}^{n} [\sum_{j=1}^{n} w_{ij}]}, \tag{6}$$

while the SPEs simplify to:

$$\widehat{\Delta}_u^l = \frac{\sum_{i=1}^{n} [\sum_{j=1}^{n} w_{ij}] \, \delta_u(0, 1, v_i, \theta^l)}{\sum_{i=1}^{n} [\sum_{j=1}^{n} w_{ij}]}, \tag{7}$$

and the TPEs and IPEs are the same:

$$\widehat{\Delta}_u^{ij} = \widehat{\Delta}_u^i = \frac{1}{L} \sum_{l=1}^{L} \delta_u(0, 1, v_i, \theta^l). \tag{8}$$

2.3 Unordered categorical inputs

If some regression parameters are allowed to vary by clusters or groups in the dataset, such as in "random effects" or "variance components" models, then we can consider the predictive effect on y of switching from one group to another. In this case u represents all inputs specific to a group, and can be thought of as an unordered categorical input, while v represents all inputs that vary within groups. For example, with a multilevel or hierarchical model, u represents inputs measured at the group-level together with group-level error terms, while v represents inputs measured on individuals within groups—such an example is considered in detail in Section 5.

In averaging over changes in an unordered categorical input u, it is the magnitude, rather than the sign, of the effects that is of interest. 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. We shall follow common practice in statistics and work with the root mean square, so that for unordered categorical inputs u with K categories, the APE of switching from one category to another is:

$$\widehat{\Delta}_u = \left(\frac{\sum_{i=1}^{n} \sum_{k=1}^{K} \sum_{l=1}^{L} [\sum_{j \in \{k\}} w_{ij}] \, (\delta_u(u_i, u^{(k)}, v_i, \theta^l))^2}{L \sum_{i=1}^{n} \sum_{k=1}^{K} [\sum_{j \in \{k\}} w_{ij}]} \right)^{1/2}, \tag{9}$$

where $\sum_{j \in \{k\}} w_{ij}$ represents weights summed over the data points in category k, and the denominator in (1) is taken to be 1. Similarly, the SPEs are:

$$\widehat{\Delta}_u^l = \left(\frac{\sum_{i=1}^{n} \sum_{k=1}^{K} [\sum_{j \in \{k\}} w_{ij}] \, (\delta_u(u_i, u^{(k)}, v_i, \theta^l))^2}{\sum_{i=1}^{n} \sum_{k=1}^{K} [\sum_{j \in \{k\}} w_{ij}]} \right)^{1/2}. \tag{10}$$

Further, as in the continuous input case, we can define $n \times K$ TPEs:

$$\widehat{\Delta}_u^{ik} = \frac{1}{L} \sum_{l=1}^{L} \delta_u(u_i, u^{(k)}, v_i, \theta^l), \tag{11}$$

which represent the (simulation averaged) expected change in y when switching from the category of individual i to category k (and v stays constant). These TPEs can be averaged over i to give K category predictive effects (CPEs):

$$\widehat{\Delta}_u^k = \frac{\sum_{i=1}^n [\sum_{j\in\{k\}} w_{ij}]\, \widehat{\Delta}_u^{ik}}{\sum_{i=1}^n [\sum_{j\in\{k\}} w_{ij}]} = \frac{\sum_{i=1}^n \sum_{l=1}^L [\sum_{j\in\{k\}} w_{ij}]\, \delta_u(u_i, u^{(k)}, v_i, \theta^l)}{L\sum_{i=1}^n [\sum_{j\in\{k\}} w_{ij}]}, \tag{12}$$

which represent the (simulation averaged) expected change in y per unit change in u when switching from the category of individual i to category k (and v stays constant), averaged over all individuals i (excluding those individuals already in category k).

3 Multiple linear regression

To see how predictive effects can be used and displayed graphically to aid understanding of regression models, we simulated $n = 180$ data-points for the multiple linear regression model with mean function

$$f(x) = \beta_1 + \beta_2 x_1 + \beta_3 x_2 + \beta_4 x_3 + \beta_5 x_2 x_3 + \beta_6 x_4 + \beta_7 x_5 + \beta_8 x_4 x_5 + \beta_9 x_6 + \beta_{10} x_6^2 + \beta_{11}\log(x_7), \tag{13}$$

where $(x_1, x_3, x_4, x_5, x_6, x_7)$ are independent standard normal, x_2 is Bernoulli with probability 0.5, $(\beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6, \beta_7, \beta_8, \beta_9, \beta_{10}, \beta_{11})$ are set at $(-0.5, 1.0, 1.0, 1.0, 0.5, 1.0, 1.0, 0.5, 1.0, 0.5, 1.0)$ and the error standard deviation is 0.5. We then obtained L =100 posterior simulation draws for the β-parameters (under standard noninformative prior distributions). For this linear model, we can derive simple algebraic expressions for many of the quantities defined earlier.

Table 3: *Algebraic derivations of predictive effects in model (13).*

Input	PE	SPE	TPE	IPE	APE
x_1	β_2^l	β_2^l	$\hat\beta_2$	$\hat\beta_2$	$\hat\beta_2$
x_2	$\beta_3^l + \beta_5^l x_{3i}$	$\beta_3^l + \beta_5^l \bar{x}_3^w$	$\hat\beta_3 + \hat\beta_5 x_{3i}$	$\hat\beta_3 + \hat\beta_5 x_{3i}$	$\hat\beta_3 + \hat\beta_5 \bar{x}_3^w$
x_3	$\beta_4^l + \beta_5^l x_{2i}$	$\beta_4^l + \beta_5^l \bar{x}_2^w$	$\hat\beta_4 + \hat\beta_5 x_{2i}$	$\hat\beta_4 + \hat\beta_5 x_{2i}$	$\hat\beta_4 + \hat\beta_5 \bar{x}_2^w$
x_4	$\beta_6^l + \beta_8^l x_{5i}$	$\beta_6^l + \beta_8^l \bar{x}_5^w$	$\hat\beta_6 + \hat\beta_8 x_{5i}$	$\hat\beta_6 + \hat\beta_8 x_{5i}$	$\hat\beta_6 + \hat\beta_8 \bar{x}_5^w$
x_5	$\beta_7^l + \beta_8^l x_{4i}$	$\beta_7^l + \beta_8^l \bar{x}_4^w$	$\hat\beta_7 + \hat\beta_8 x_{4i}$	$\hat\beta_7 + \hat\beta_8 x_{4i}$	$\hat\beta_7 + \hat\beta_8 \bar{x}_4^w$
x_6	$\beta_9^l + \beta_{10}^l (x_{6i} + x_{6j})$	–	$\hat\beta_9 + \hat\beta_{10}(x_{6i} + x_{6j})$	–	–
x_7	$\beta_{11}^l \frac{\log(x_{7j}/x_{7i})}{(x_{7j}-x_{7i})}$	–	$\hat\beta_{11} \frac{\log(x_{7j}/x_{7i})}{(x_{7j}-x_{7i})}$	–	–

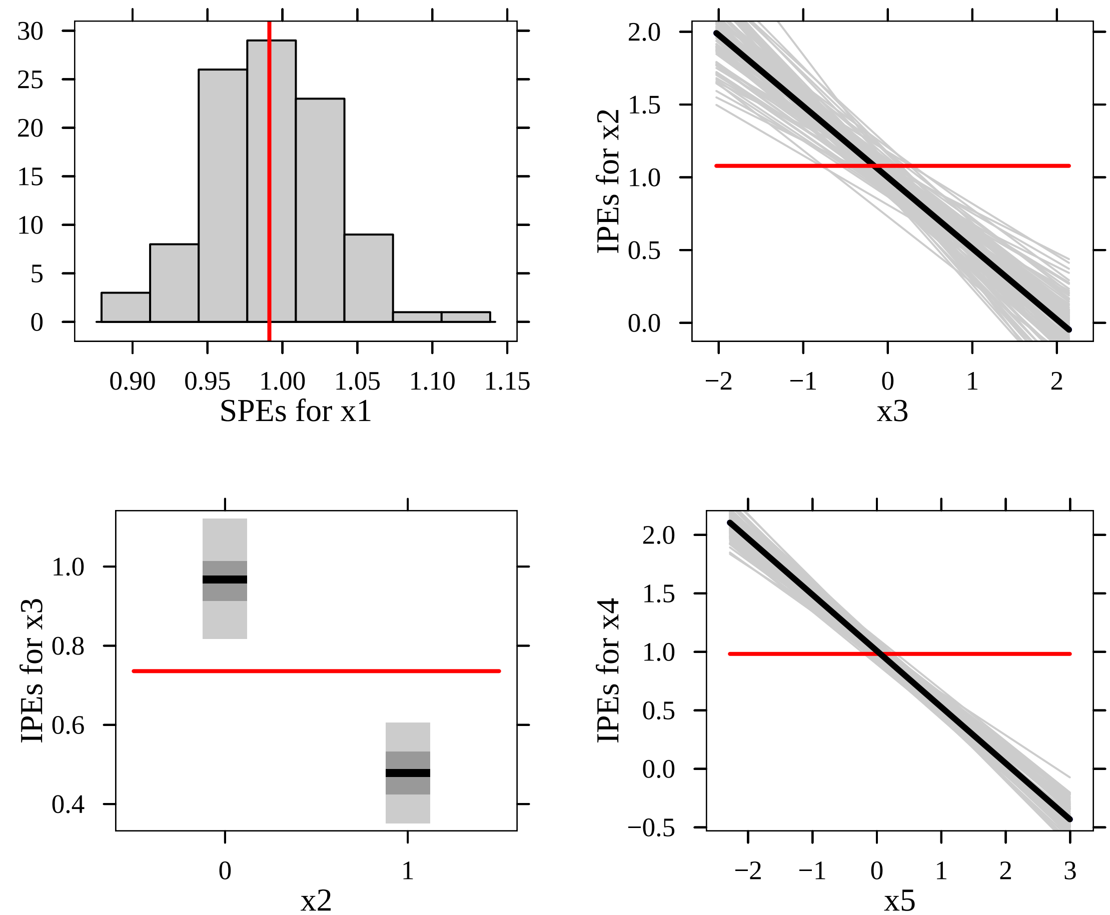
Here, $\bar{x}_p^w$ is the weighted average of x_p:

$$\bar{x}_p^w = \frac{\sum_{i=1}^n [\sum_{j=1}^n w_{ij}]\, x_{pi}}{\sum_{i=1}^n [\sum_{j=1}^n w_{ij}]}.$$

The PEs for x_1 are constant for each initial and final value of x_1, and do not vary with any of the other inputs, so the only variation is from posterior simulation in the SPEs. This variation can be summarized graphically in a density estimate or histogram—see the upper left graph in Figure 2—or numerically by APE = 0.99 (the average of the SPEs) and standard error = 0.04 (the standard deviation of the SPEs).

However, the PEs for x_2 in model (13) vary with the value of x_3, a continuous input; this can be presented graphically in a scatterplot of IPEs (8) versus x_3. Since in this case the IPEs do not vary with other input values, the values form a smooth function across the plot (as shown by the thick black line in the upper right graph in Figure 2). We can incorporate posterior variation on this graph by adding similar thin gray lines for the PEs (1) themselves (which represent posterior draws of IPEs in the case of binary inputs). Finally, we can indicate the APE on this graph by adding a red horizontal line at its value.

Figure 2: *Graphical displays of predictive effects for* (x_1, x_2, x_3, x_4) *in model (13). The upper left graph displays variation in the* x_1 *SPEs about the APE represented by the red line. The thick black line in the upper right graph shows how IPEs for* x_2 *vary with the value of* x_3*, while the thin gray lines display posterior variation, and the red line shows the APE. The thick black lines in the lower left graph show how IPEs for* x_3 *vary with the two values of* x_2*, while the inner dark gray bands show 50% posterior intervals, the outer light gray bands show 95% intervals, and the red line shows the APE. The thick black line in the lower right graph shows how IPEs for* x_4 *vary with the value of* x_5*, while the thin gray lines display posterior variation, and the red line shows the APE.*



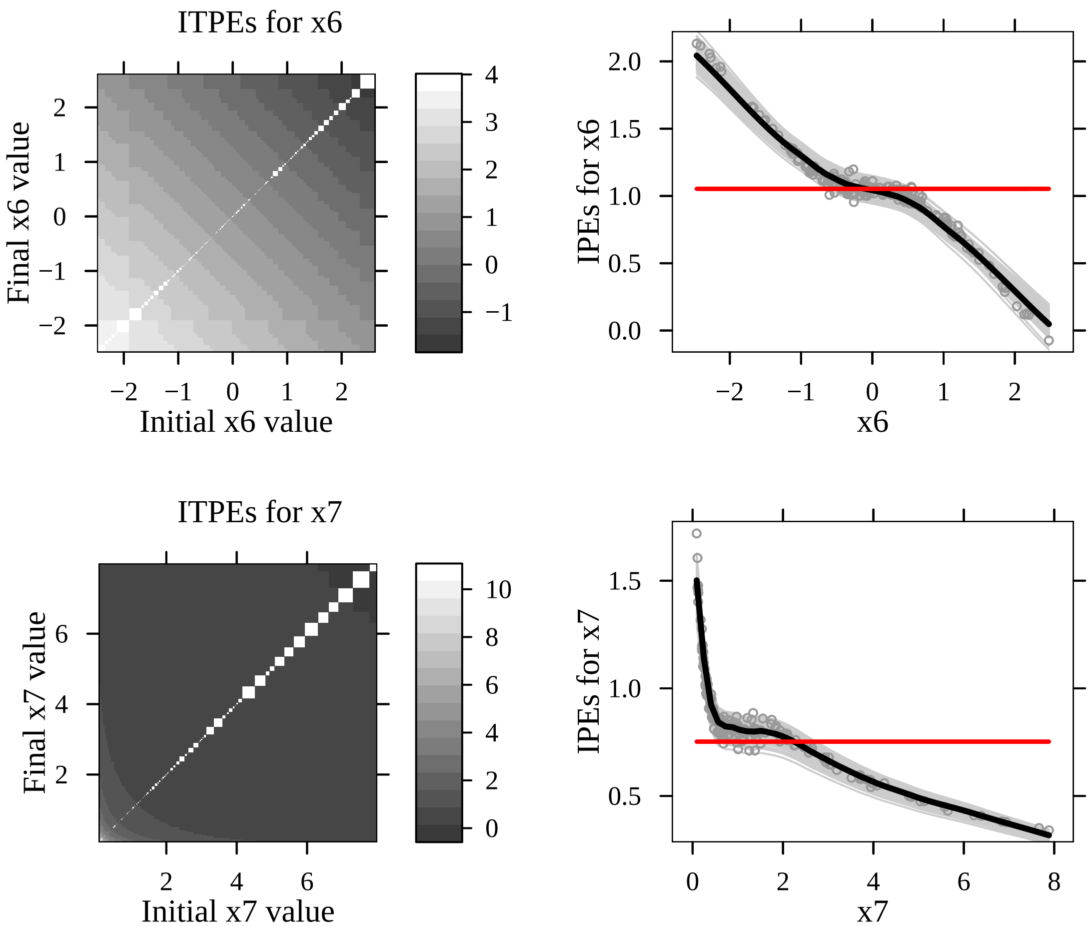
This graph indicates that when x_2 changes from 0 to 1 (representing a change in category from male to female, say), and the other inputs stay constant, we can expect y to increase by 2 units when x_3 is low (-2), but only 1 unit when x_3 is at a central value (0), and to stay about the same when x_3 is high (2). Further, the (posterior) uncertainty surrounding these statements can be seen to be from about ± 0.25 at the center of the range of x_3 to about ± 0.5 at the extremes of the range of x_3. Finally, averaging over the empirical distribution of x_3 leads to an average predictive effect of x_2 on y of about 1 unit.

The PEs for x_3 in model (13) take on a different form to those for x_2, since they vary with x_2, a binary input. The lower left graph in Figure 2 displays the IPEs (5) as thick black lines at the centers of boxes representing posterior variation in the PEs (1)— the dark gray boxes extend to the 25th and 75th percentiles, and the light gray boxes extend to the 2.5th and 97.5th percentiles. Again, we indicate the APE on this graph with a red horizontal line. We interpret this graph similar to the graph for x_2, although the PEs in this case refer to expected changes in y per unit change in x_3, holding all other inputs constant.

The predictive effect graph for x_4 is similar to that for x_2, since the PEs for x_4 vary with x_5, a continuous input—see the lower right graph in Figure 2. This shows expected changes in y per unit change in x_4, holding all other inputs constant. (The predictive effect graph for x_5 is very similar and so is not shown.)

The PEs for x_6 in model (13) vary with the initial and final values of x_6. This can be seen by averaging over the posterior simulations and displaying the resulting TPEs (4) in the "levelplot" in the upper left of Figure 3. This graph displays expected changes in y per unit change in x_6 starting from an initial $x_6 = x_{6i}$

Figure 3: *Graphical displays of predictive effects for* (x_6, x_7) *in model (13). The levelplots on the left display TPEs, while the scatterplots on the right display the corresponding IPEs which average the TPEs in the levelplots across the vertical axis. The thick black line shows how the IPEs vary with the input values, the thin gray lines display posterior variation, and the red line shows the APE.*



value (on the horizontal axis) and ending at a final $x_6 = x_{6j}$ value (on the vertical axis). For example, we can expect y to increase by around 2.5 units per unit increase in x_6 when x_6 increases from -2 to -1, but only 0.5 units when x_6 increases from 0 to 1. The graph has constant bands running from lower right to upper left because along these paths $x_{6i} + x_{6j}$ = a constant and the TPEs for this example are $\hat{\beta}_9 + \hat{\beta}_{10}(x_{6i} + x_{6j})$. Thus, one way to summarize this graph is to look along 45° lines from lower left to upper right since these all have identical profiles of TPE values. These 45° paths represent $x_{6j} = x_{6i} + k$, where k is a constant, and so the main diagonal represents the limit as $k \to 0$ of the predictive effect of x_6 on y, i.e. the derivative of (13) with respect to x_6, $(\hat{\beta}_9 + 2\hat{\beta}_{10}x_6)$.

We take an alternative approach in this paper which is to collapse the levelplot across one of the axes (the vertical axis, say), by averaging across the estimated distribution of the x_{6j} values using weights representing how likely each transition is. This is exactly what the IPE (5) values are, and these can be displayed graphically in a similar manner to that for the predictive effects of x_4 in the lower right graph of Figure 2.

Such a graph is displayed is the upper right of Figure 3, where the IPEs are plotted as light gray circles. Since they do not fall along a line as they did for the predictive effects of x_4 in Figure 2, but rather scatter across a range of values between 0 and 2, we find it helpful to smooth these points using the "loess" (Cleveland and Devlin, 1988) function in R. This thick black line in the scatterplot highlights any strong trends in how the IPEs vary with x_6. We can construct thin gray lines similarly to show posterior variation, and, as before, the horizontal red line shows the APE. This approach can easily be generalized for more complex models.

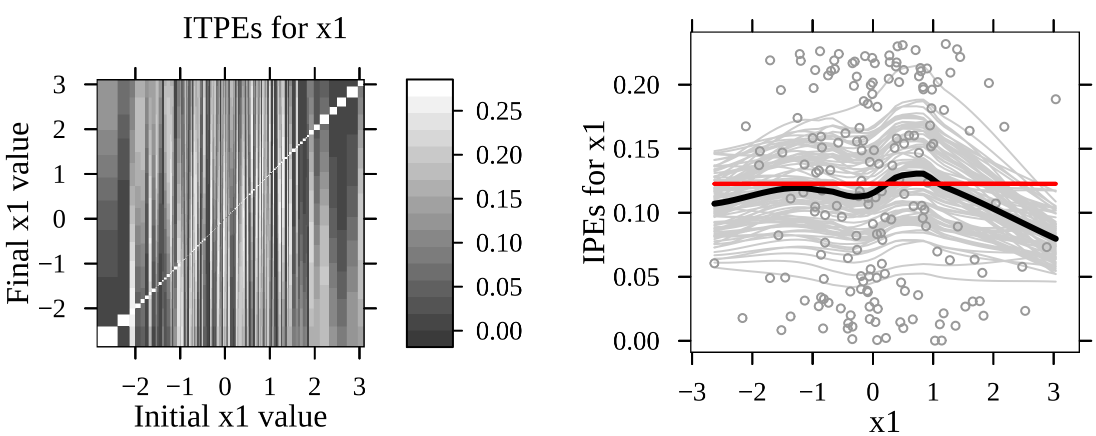
Finally, the PEs for x_7 in model (13) vary with the initial and final values of x_7. Thus we take the same approach as for x_6 and construct a levelplot of the TPEs (lower left graph of Figure 3) and a lineplot of the APE and IPEs including posterior variation (lower right graph of Figure 3). The appearance of the graph makes sense because the logarithmic transformation of x_7 in the model means that the predictive effect on y of a change in x_7 is much larger for low values of x_7 than for high values.

4 Logistic regression

We next generalize the graphs in Figures 2 and 3 for a logistic regression example in which simple algebraic expressions for the predictive effects are not available. We simulated $n = 180$ data-points using the same linear predictor as (13) and response Bernoulli with probability $1/(1+\exp(-f(x)))$. We then obtained L =100 posterior simulation draws for the β-parameters (under standard noninformative prior distributions).

The PEs for x_1 in this example vary not only with the initial and final values of x_1, but also with the values of the other inputs. A levelplot of the TPEs is therefore more complex than in the previous example—see left graph of Figure 4. There is little systematic variation in this levelplot, but we can again collapse

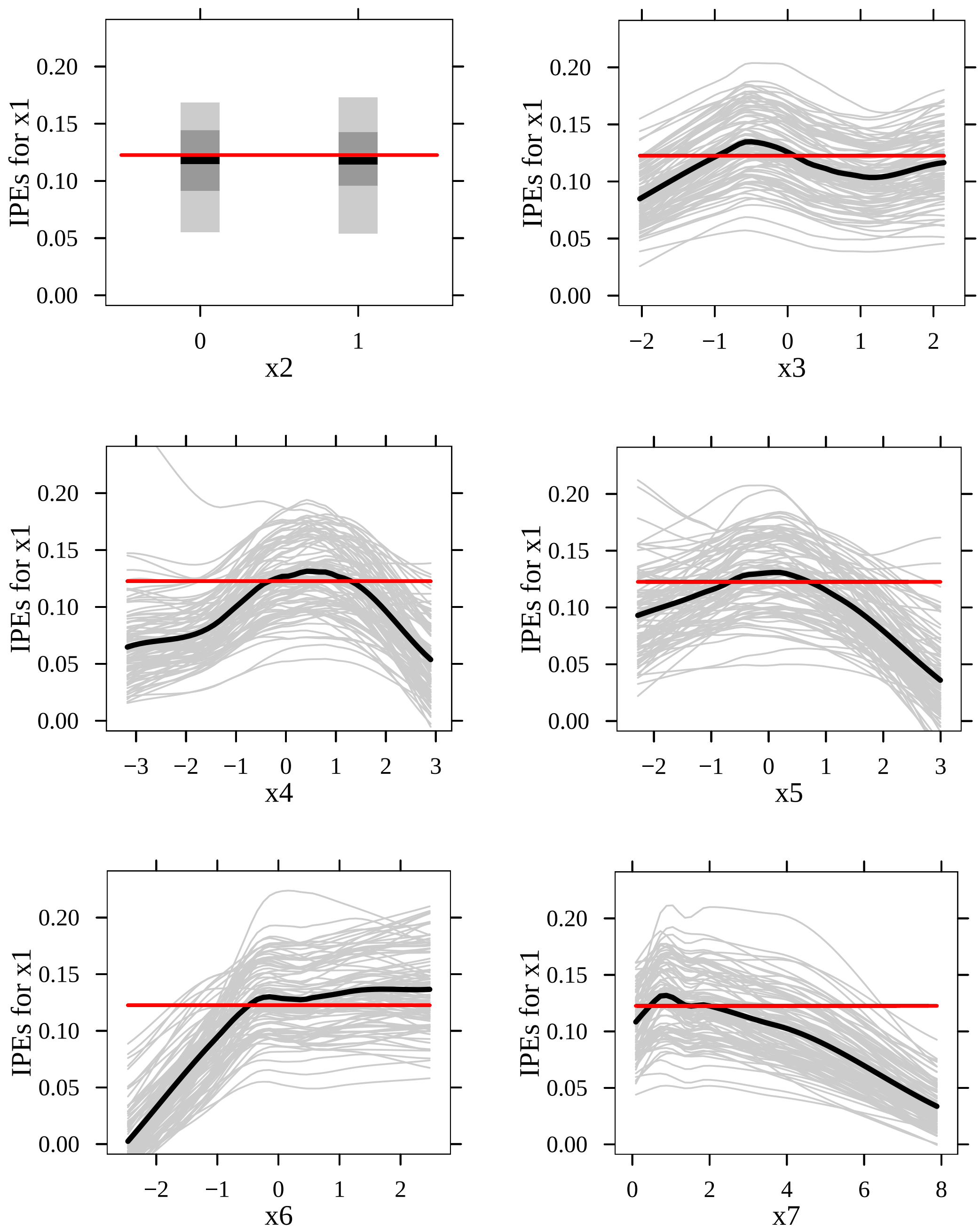
Figure 4: *Graphical displays of predictive effects for x_1 in the logistic example. The levelplot on the left displays TPEs, while the scatterplot on the right shows the corresponding IPEs which average the TPEs in the levelplot across the vertical axis. The thick black smooth in this scatterplot shows how the IPEs vary with x_6, while the thin gray smooths display posterior variation, and the red line shows the APE.*



it across its vertical axis to produce the scatterplot of IPEs on the right of Figure 4. The IPEs, plotted as light gray circles, scatter across a range of values between 0 and 0.25, so we smooth these points as before to highlight any strong trends—this is the thick black line in the scatterplot. Again, the thin gray lines show posterior variation, and the horizontal red line shows the APE. There is some suggestion that IPEs are slightly higher in the middle of the range of x_1 than at the extremes, but this variation is dominated by the posterior variation.

However, since in this example the IPEs for x_1 also vary with the values of the other inputs, it makes sense to produce similar graphs with the other inputs along the horizontal axis—see Figure 5. These graphs suggest that the predictive effect of x_1 is a little lower than average for low values of x_4 and x_6 and also for high values of x_4, x_5, and x_7. On revisiting the data, this appears to be a consequence of most of the values

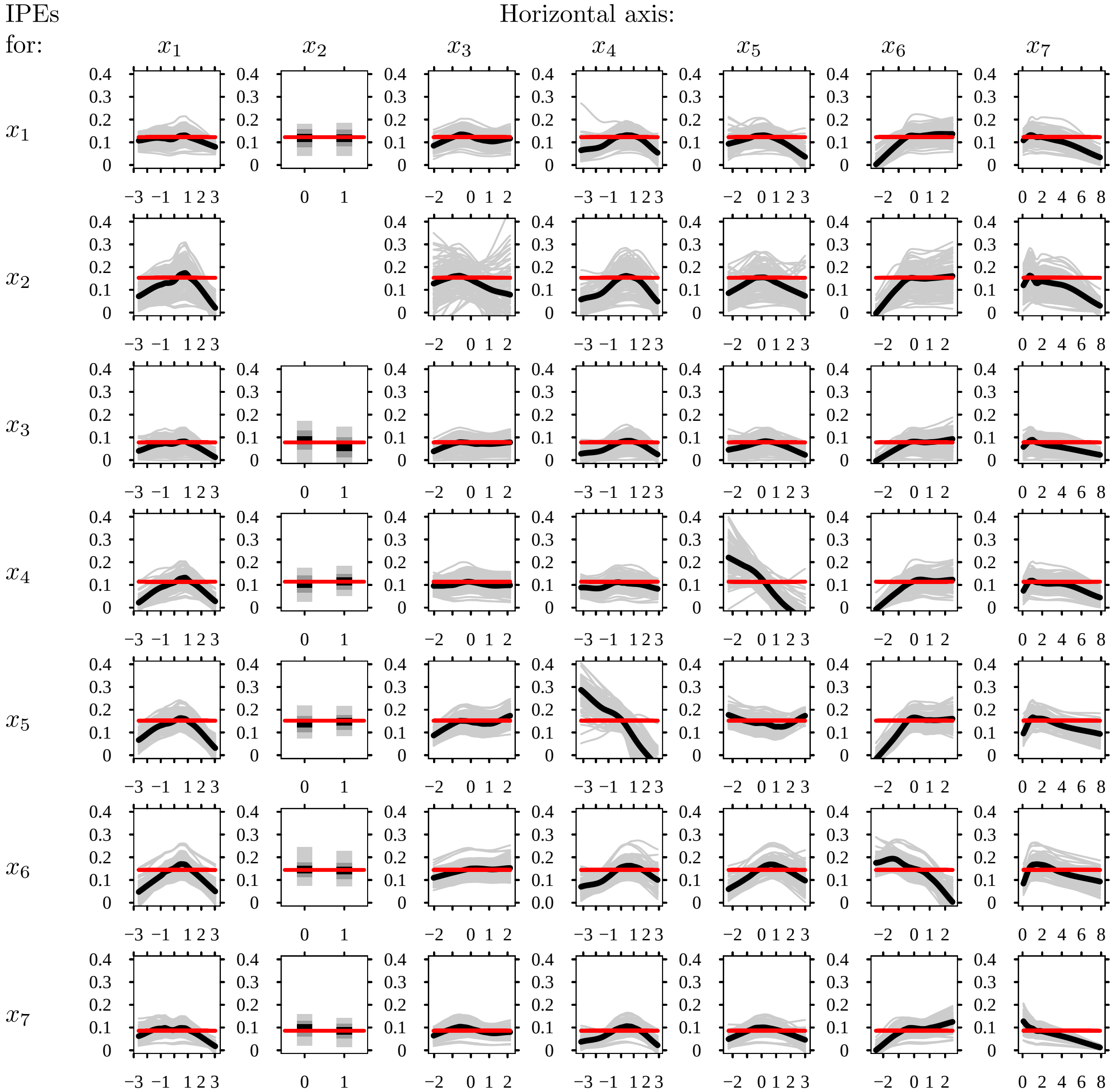
Figure 5: *Graphical displays of predictive effects for x_1 in the logistic example. The thick black lines show how IPEs vary with values of the inputs, while the gray lines and bands display posterior variation, and the red lines show APEs.*



of y for low values of x_4 and x_6 being 0, and most of the values of y for high values of x_4, x_5, and x_7 being 1. With little response variation at these input values, changing x_1 has little impact on y.

We can construct similar graphs for each of the inputs—Figure 6 displays all possible graphs for this example. We designed these graphs following the "small multiples" idea of Tufte (2001). If the inputs have comparable scales—as they do in this constructed example—each of the graphs can be given a common vertical scale to allow easy comparisons of predictive effects across inputs.

Figure 6: *Graphical displays of all predictive effects for the logistic example. Each row contains multiple graphs of the IPEs for an input, with each graph having a different input on the horizontal axis. Each column contains the same input along the horizontal axis. The thick black lines show how IPEs vary with values of the inputs, while the gray lines and bands display posterior variation, and the red lines show APEs.*



5 Application revisited

Figure 7: *Graphical displays of predictive effects for first six individual inputs in the multilevel criminal justice application. Each row contains multiple graphs of the IPEs for an input, with each graph having a different county-level input on the horizontal axis. Each column contains the same county-level input along the horizontal axis. The thick black lines show how IPEs vary with values of the county-level inputs, while the gray lines and bands display posterior variation, and the red lines show APES.*

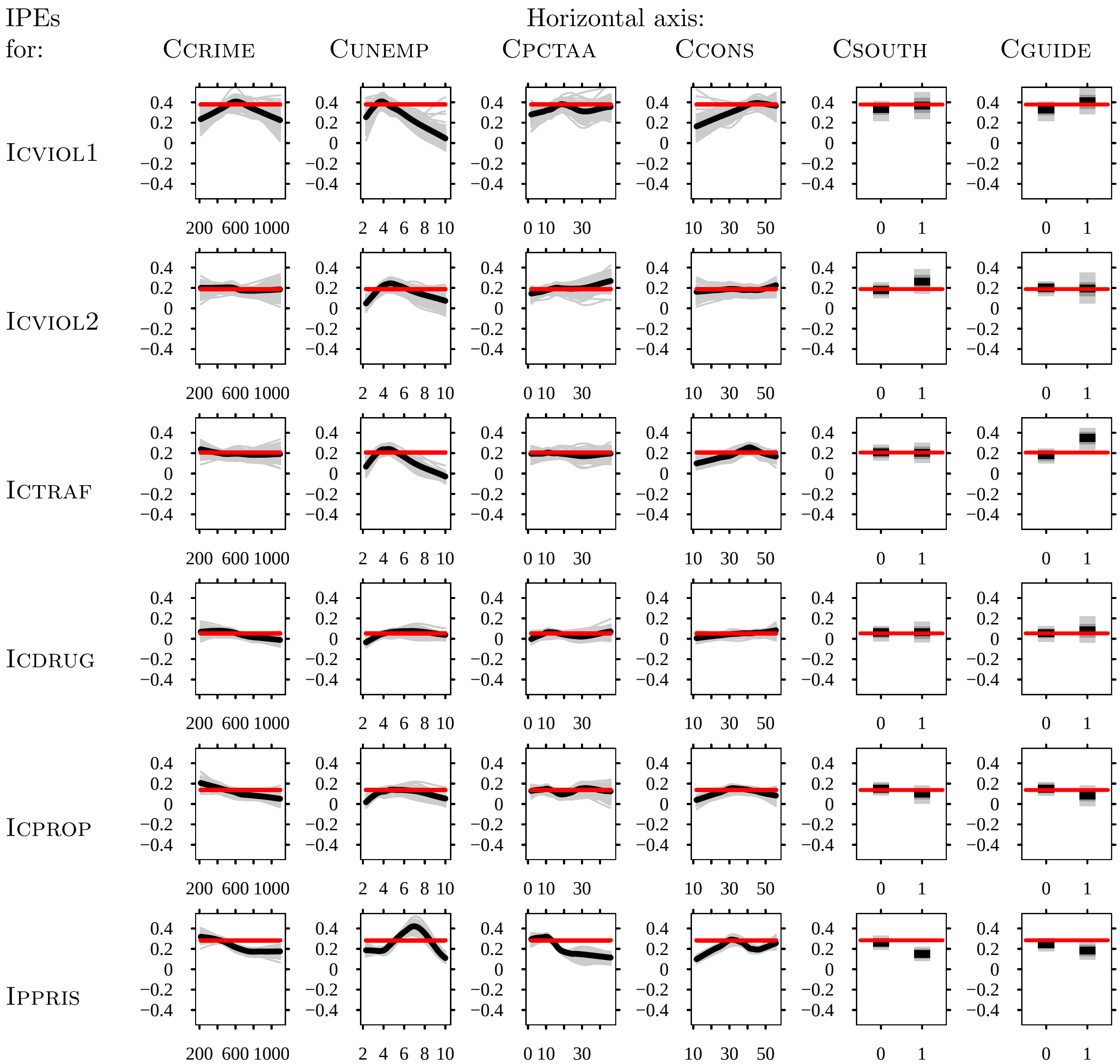


Figure 8: *Graphical displays of predictive effects for last six individual inputs in the multilevel criminal justice application. Each row contains multiple graphs of the IPEs for an input, with each graph having a different county-level input on the horizontal axis. Each column contains the same county-level input along the horizontal axis. The thick black lines show how IPEs vary with values of the county-level inputs, while the gray lines and bands display posterior variation, and the red lines show APEs.*

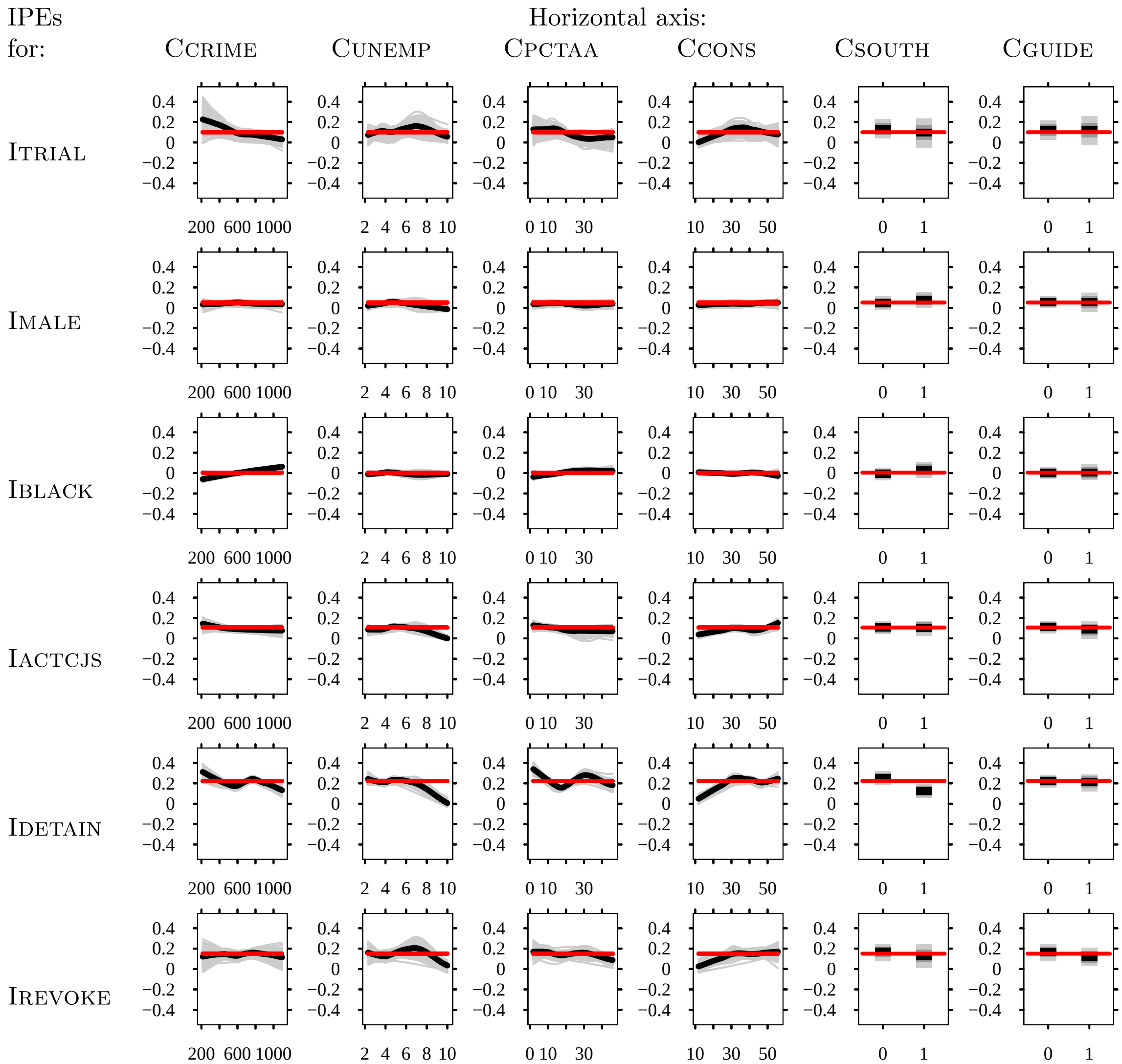


Figure 9: *Graphical displays of predictive effects for county-level inputs in the multilevel criminal justice application. Each row contains multiple graphs of the IPEs for an input, with each graph having a different individual input on the horizontal axis. Each column contains the same individual input along the horizontal axis. The thick black lines show how IPEs vary with values of the individual inputs, while the gray bands display posterior variation, and the red lines show APEs.*

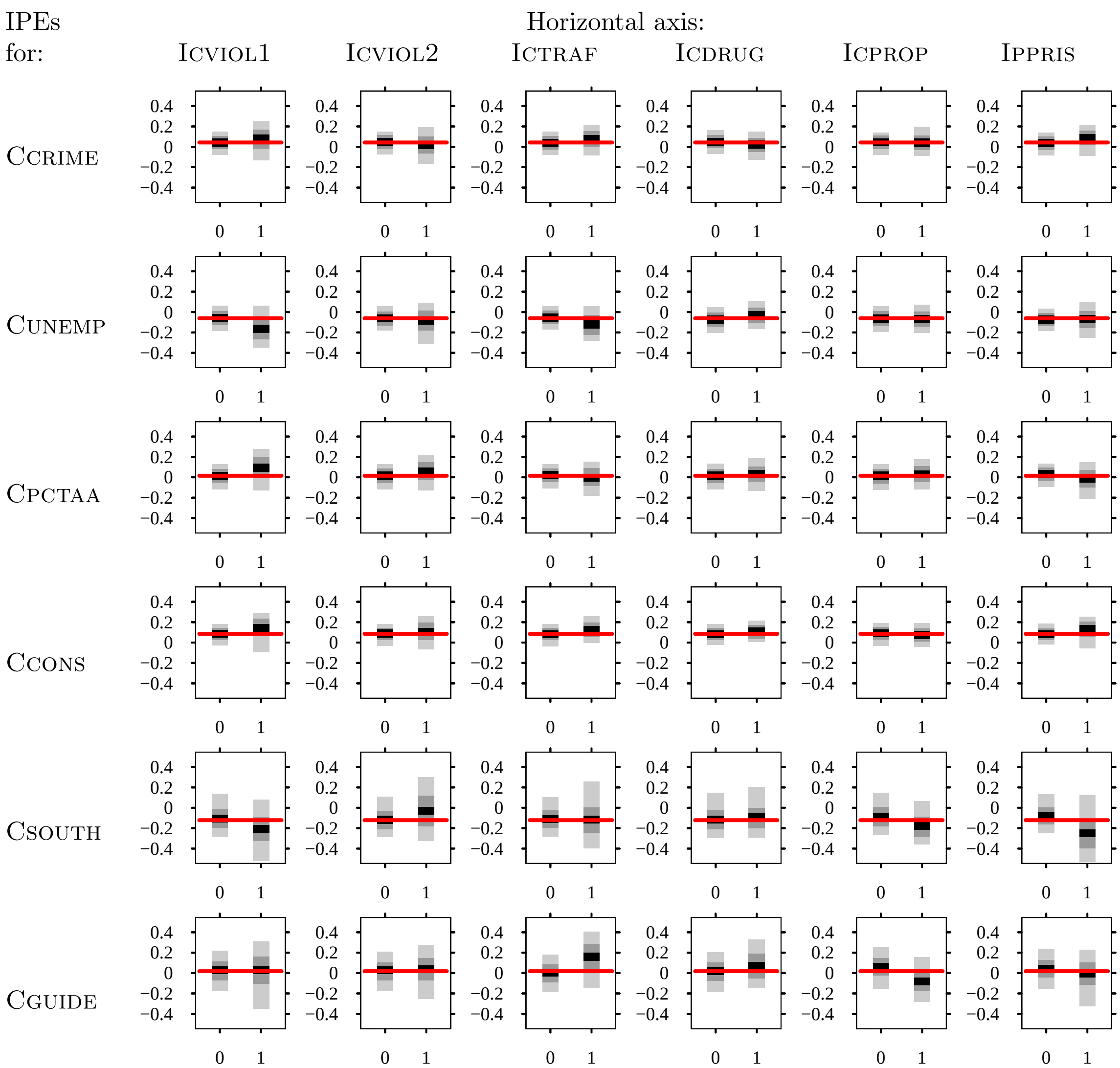


Figure 10: *Graphical displays of predictive effects for county-level inputs in the multilevel criminal justice application. Each row contains multiple graphs of the IPEs for an input, with each graph having a different individual input on the horizontal axis. Each column contains the same individual input along the horizontal axis. The thick black lines show how IPEs vary with values of the individual inputs, while the gray bands display posterior variation, and the red lines show APEs.*

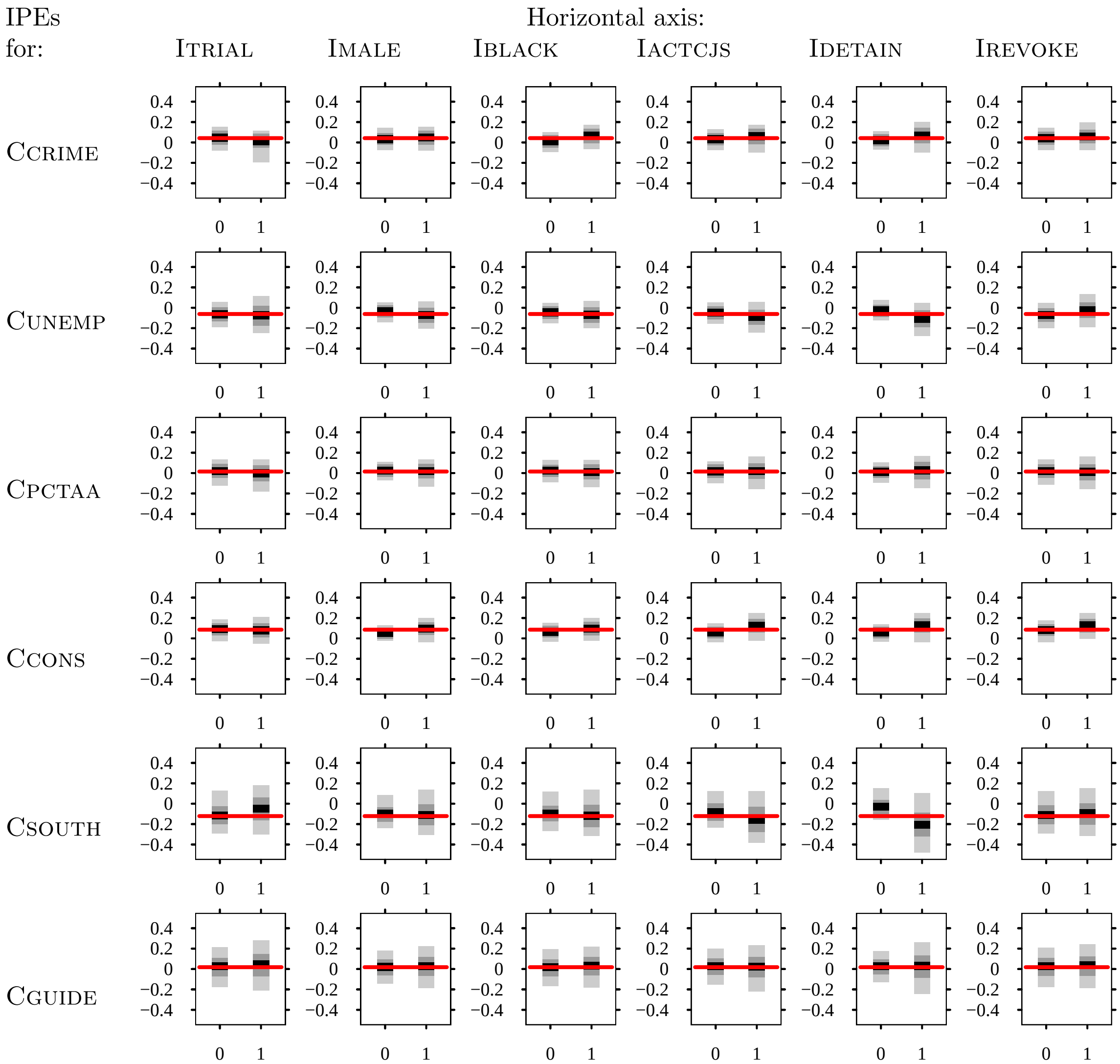
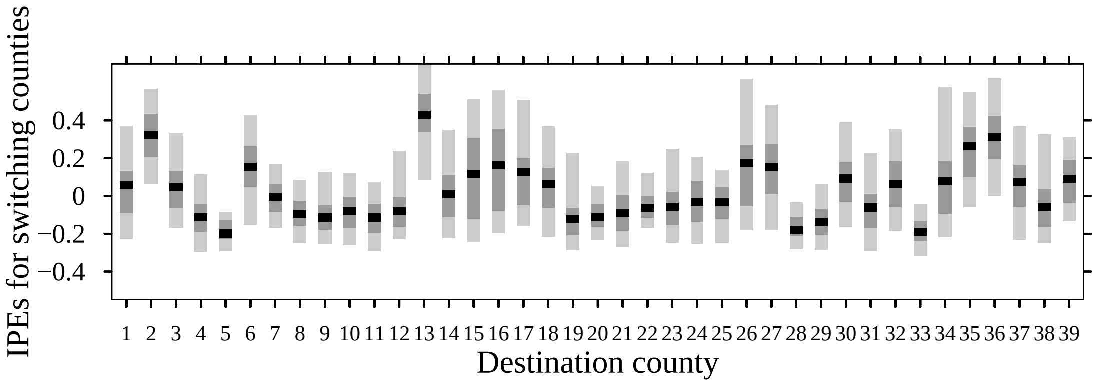


Figure 11: *Graphical display of predictive effects for switching counties. The thick black lines show how the CPEs vary by destination county, while the gray bands display posterior variation.*



References

Cleveland, W. S. and Devlin, S. J. (1988). Locally-weighted regression: An approach to regression analysis by local fitting. *Journal of the American Statistical Association* **83**, 596–610.

Gelman, A., and Pardoe, I. (2004). Average predictive effects for models with nonlinearity, interactions, and variance components. Technical report, Department of Statistics, Columbia University.

Graubard, B. I., and Korn, E. L. (1999). Predictive margins with survey data. *Biometrics* **55**, 652–659.

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.

McCullagh, P., and Nelder, J. A. (1989). *Generalized Linear Models*, second edition. New York: Chapman & Hall.

Pardoe, I., and Weidner, R. W. (2004). Sentencing convicted felons in the United States: a Bayesian analysis using multilevel covariates. *Journal of Statistical Planning and Inference*, to appear.

Tufte, E. (2001). *The Visual Display of Quantitative Information*, second edition. Cheshire, Conn.: Graphics Press.