Evaluation in Multi-Site Programs[*]

Rajeev Dehejia[**]

Department of Economics and SIPA
Columbia University

and

NBER

This paper discusses the problem of evaluating and predicting the treatment impact of a program that is implemented at multiple sites. Two issues arise: is information from other sites relevant in estimating the impact at a given site? and how can we account for predictive uncertainty in the site effects? Using data from the Greater Avenues for Independence evaluation, I develop a hierarchical model for earnings which allows both for site effects and for smoothing the estimated impact across sites. I show that the degree to which the estimated impact is smoothed across sites does not affect the estimate; i.e. most of the differences across sites are due to differences in the composition of participants. Second I show that predictive uncertainty regarding site effects is important; for example, when the predictive uncertainty regarding site effects is ignored, the treatment impact at the Riverside sites is significant, but when we consider predictive uncertainty, the impact for the Riverside sites is insignificant. Third, I show that the hierarchical model is able to extrapolate site effects with reasonable accuracy when the site for which the prediction is being made does not differ substantially from the sites already observed. For example, the San Diego treatment effects could have been predicted based on observable site characteristics, but the Riverside effects are consistently underestimated.

First version: 1 June 1998
Current version: 28 June 1999

[**] Department of Economics, Room 1022, Columbia University, 420 W. 118th Street, New York, NY 10027. E-mail: dehejia@columbia.edu.

1. Introduction

This paper discusses the problem of evaluating and predicting the treatment impact of a program that is implemented at multiple sites. Many programs operate at, and are evaluated at, multiple sites, e.g., NSW, JTPA, New Chance, and Greater Avenues for Independence (GAIN), which is studied in this paper. Two kinds of differences can emerge across sites: differences in the composition of participants and differences in the treatment (or how it is administered). The former is relatively easy to address if a sufficient number of the participants' characteristics are observed. The latter leads to more difficult questions: are observations on participants at other sites relevant in determining the treatment impact at a given site? and if the program is implemented again, how can we predict the impact at a new site? These questions are important because they influence the interpretation of many evaluations.

One standard method of dealing with site-specific effects is to use a fixed-effects model, or more generally to estimate separate models for each site. To the extent that cross-site differences are administrative or qualitative, this would be a valid approach to estimate the impact in the original implementation. However, it limits us to thinking of subsequent implementations of the program being identical to one of the original sites, because there is no framework to account for predictive uncertainty regarding the value of the fixed effect or site-specific model. Another solution would be to pool the data. This treats units as exchangeable conditional on observable variables, ignoring site effects, and allows us to predict the outcome in a subsequent implementations. Alternatively we can think of pooling as averaging over site effects and predicting the outcome averaged over

subsequent implementations. Neither interpretation allows us to consider how the site effects would vary in subsequent implementations.

This paper explores the middle ground between these methods through the use of Bayesian hierarchical modeling (see Chamberlain and Imbens [1996], Geweke and Keane [1996], and Rossi, McCulloch, and Allenby [1995] for other applications of these methods). Hierarchical modeling is somewhat familiar in the frequentist literature through the related concept of meta-modeling (see Cooper and Hedges [1994]). Meta-modeling involves linking the outcomes of separate studies on the same topic through an over-arching model. It can also be used to model site effects; for example, Card and Krueger (1992) estimate cohort and state-of-birth specific returns to schooling and then use a meta-model to relate these to measures of school quality. The method adopted in this paper is a Bayesian version of meta-modeling.

There are three layers to the model. The first layer consists of separate regression models for each site. The second layer links the coefficients of the site models through a regression-type meta-model. The third layer consists of prior distributions for the unknown parameters. Thus, a hierarchical model combines features of the fixed-effect and pooled models, but also allows for intermediate models. First, compared to standard fixed (or random) effects, it allows for site-specific estimation of all coefficients, not just the constants. Second, participants across sites are not assumed to be exchangeable conditional on individual characteristics, but instead are assumed to be exchangeable within sites conditional on individual characteristics. Third, we use a prior distribution to model the extent to which we believe that site-effects are drawn from a common

distribution; i.e., the extent to which coefficients should be "smoothed" across sites, or observations from one site should influence our estimates in other sites.

This approach is implemented using data from the Greater Avenues for Independence (GAIN) Demonstration, a labor training program implemented in six California counties at 24 sites (see Riccio, *et al.*, [1996]). Much attention in the GAIN program focused on the Riverside county implementation, which was viewed as being highly successful and distinct from other counties (see, for example, Nelson [1997]). From the perspective of GAIN, the interest is in discovering the extent to which a hierarchical model succeeds in capturing these site effects which have been viewed as being primarily qualitative in nature. We focus on three issues. First, to what extent does pooling the data across all sites obscure site effects? Second, if we imagine re-implementing a GAIN-type program, would we be able to predict the site effects based on the observable characteristics of each site, and how important is predictive uncertainty? Third, how well can the model extrapolate to sites that have not been observed? These questions will be addressed below.

Two previous papers on multi-site evaluation issues are Heckman and Smith (1996) and Hotz, Imbens, and Mortimer (1998). Heckman and Smith analyze the sensitivity of experimental estimates to the choice of sites used in the analysis and to different methods of weighting the pooled data. The paper establishes that there is significant cross-site variation in the data from the JTPA evaluation. Hotz, Imbens, and Mortimer analyze the importance of site effects using the key insight that, even if there is heterogeneity in the treatment available at each site, control groups excluded from the program at other sites should still be comparable if there are no site effects. They

conclude that there are, in fact, important site effects in the data they analyze.[1] The

hierarchical model I adopt is one possible approach to deal with site effects.

The paper is organized as follows. Section 2 describes the GAIN program.

Section 3 discusses key features of the GAIN data. Section 4 outlines the hierarchical

model. Section 5 presents the results, and Section 6 concludes.

2. The Data

The GAIN program began operating in California in 1986, with the aim of "increasing

employment and fostering self-sufficiency" among AFDC recipients (see Riccio, *et al.*,

[1994]). In 1988, six counties -- Alameda, Butte, Los Angeles, Riverside, San Diego, and

Tulare -- were chosen for an experimental evaluation of the benefits of GAIN. A subset

of AFDC recipients (single parents with children aged six or older and unemployed heads

of two-parent households) were required to participate in the GAIN experiment (see

Table 1). [2]

Potential participants from the mandatory group were referred to a GAIN

orientation session when they visited an Income Maintenance office (either to sign up for

welfare or to qualify for continued benefits).[3] As a result, the chronology of the data and

subsequent results are in experimental time, rather than calendar time. No sanctions were

used if individuals failed to attend the orientation sessions. However, once individuals

started in the GAIN program, sanctions were used to ensure their ongoing participation.

---

[1] A more recent version of this paper analyzes the GAIN data, arguing that, whereas site effects in control earnings may not be very important, there is evidence of treatment heterogeneity across sites. I use the term site effects to refer to site-specific effects for both the treatment and control groups.
[2] This discussion draws on Dehejia (1999).
[3] In some counties AFDC recipients were allowed to volunteer into the GAIN program, but these units are not included in the public use sample.

At the time of enrollment into the program, a variety of background characteristics were recorded for both treatment and control units including: demographic characteristics; results of a reading and mathematics proficiency test; and data on ten quarters of pre-treatment earnings, AFDC, and food stamp receipts.[4]

Of those who attended the orientation session, a fraction was randomly assigned to the GAIN program,[5] and the others were prohibited from participating in GAIN.[6] Each of the counties randomized a different proportion of its participants into treatment, ranging from a 50-50 split in Alameda to an 85-15 split in San Diego (see Table 1). Because assignment to treatment was random, the distribution of pre-treatment covariates is balanced across the treatment and control groups. In terms of the chronology of data gathering, "experimental" time (which I also refer to as "post-experimental" or "post-treatment" time) begins when individuals attend the GAIN orientation session. The early stages of experimental time thus coincide with the education and training of GAIN participants.[7]

In the GAIN experiment, the treatment is participating in the GAIN program; the control is receiving standard AFDC benefits. The GAIN program works as follows: based on test results and an interview with a case manager, participants are assigned to one of two activities. Those deemed not to be in need of basic education are referred to a

---

[4] Data on AFDC and Food Stamp receipts were taken from each county's welfare records. Data on earnings were taken from the California State Unemployment Insurance Earnings and Benefits Records. Other background characteristics were taken from California's client information ("GAIN-26") form. See Riccio, *et al.*, (1994).

[5] The randomization was (as far as we know) independent of pre-treatment covariates. This is confirmed by the data. A different fraction was randomized into treatment in each county. See Table 1.

[6] Of course, these individuals could participate in non-GAIN employment-creating activities .

[7] More precisely, individuals were registered in the first quarter of experimental time. This means that in some cases the first quarter of experimental time in fact includes information one or two months prior to the commencement of the experiment. So for example, for an individual who attended an orientation session in February 1989, the first quarter of experimental time is from January to March 1989. Of course,

job search activity (which lasts about three weeks); those who do not find work are placed in job training (which includes vocational or on-the-job training and paid or unpaid work experience, lasting about three to four months). Those deemed to be in need of basic education may choose to enter job search immediately, but if they fail to find a job they must register for preparation toward the General Educational Development certificate, Adult Basic Education, or English as a Second Language programs (lasting three to four months).[8] Participants were exempted from the requirement to participate in GAIN activities if they found work on their own.[9]

The counties in the GAIN experiment varied along two important dimensions. First, the composition of program participants varied, because counties chose to focus on particular subsets of their welfare populations and the populations differed. For example, Alameda and Los Angeles counties confined themselves to the subset of long-term welfare recipients (individuals having already received welfare for two years or more). The second difference is that the sub-treatment offered within each county varied due to differences in administrative philosophy. The approach followed by Riverside, which has received much attention, was to focus on job, rather than skills, acquisition. Both are part of the program, but Riverside's emphasis was the former. Instead counties like Alameda

---

some part of the first and second quarters could be spent participating in treatment activities. Pre-treatment data would cover the ten quarters from July 1986 to December 1988.

[8] The public use data do not contain information on each individual's participation in the various components of the program. At the same time, individuals in the control group can participate in non-GAIN activities. Thus, the treatment effect measures the increase in earnings, employment, etc., from the availability of and encouragement (or requirement) to use GAIN-related activities compared to pre-existing employment services.

[9] Note that only about eight-five per cent of the treated units actively participated in any GAIN activities (though by virtue of being in the GAIN sample they did attend an orientation meeting); the balance satisfied the requirements of the GAIN program on their own (in most cases finding employment within the first two or three quarters of experimental time). Thus, as observed earlier, this is important in interpreting the treatment effect as a comparison between earnings, employment, etc., when individuals are required to find a job or to participate in GAIN-related activities and when they are not obliged to find jobs and only pre-existing employment-related services are available.

focused more on skill acquisition. The model will allow for differences in composition by conditioning on pre-treatment covariates and differences in the treatment by allowing for site effects.

3. The GAIN Data

Table 1 presents the six counties that participated in the GAIN experiment, broken down in terms of their 24 administrative sites. The counties vary from one-site counties such as Alameda to multi-site counties such as Los Angeles and San Diego. This paper will analyze the results at the site level because with six counties there is minimal scope for modeling site effects. Table 2 presents the background characteristics of each site in greater detail. We note that the average number of children varies from over four in some sites (site 21) to slightly over two in others (site 6). The proportion of Hispanics in the sample varies from a low of 6 per cent (site 1) to over 50 per cent in other sites (sites 14 and 24).

Table 2 shows that there is significant variation in the treatment impact across sites. The second last column presents the average quarterly post-treatment earnings for the treatment and control groups. The treatment impact ranges from a high of $212 for Site 5 (in Riverside county) to a low of –$132 for Site 17 (in Tulare). In the last column the treatment effect is estimated conditioning on pre-treatment covariates through an OLS regression. The estimates are similar, ranging from –$90 to $292. The sites consistently showing the highest and most significant impacts are those from Riverside county (sites 2 to 5). Their treatment impacts range from $149 to $292, and are significantly different

from zero. The worst-performing county is Tulare, for which some of the impacts are negative and all are statistically insignificant.

4. The Econometric Model

A feature of the data which influences the modeling strategy is the large proportion of zeros in the outcome, earnings. With as many as 75 per cent of the outcomes being zero, the model must explicitly account for the mass point in the earnings distribution. The model has two components. The first is a probit model that determines the probability of positive earnings. The second, conditional on positive earnings, uses a hierarchical model, which is outlined below (see Dehejia [1999], which uses a similar model).

4.1 The Probit Model

Let Y_{it} denote earnings, for $i=1,\ldots,N$, and $t=1,\ldots,\tau$. $Y_{it}^*=0$ if $Y_{it}=0$, and $=1$ if $Y_{it}>0$. The probit model defines a latent variable such that $Y_{it}^*=0$ if $Y_{it}^{**}<0$, $=1$ if $Y_{it}^{**}>0$. The latent variable is assumed to follow a normal likelihood:

$$Y_{it}^{**}\left|\left\{Y_{it}^*=y_{it}^*\right\}_{t=1}^{t-1},\left\{X_{it}=x_{it}\right\}_{t=1}^t,\beta \sim N(x_{it}\beta,1) \tag{1}$$

where $x_{it}=[e_{it}\ T_i \cdot e_{it}]$, T_i is a treatment indicator ($=1$ if treated, $=0$ otherwise), $e_{it}=[Y_{i,t-1}\ c_{it}]$, and c_{it} is a vector of exogenous pre-treatment variables. Rather than using maximum likelihood to obtain a point estimate, we use a Gibbs sampling procedure to obtain the posterior distribution of the parameters (see Appendix A; also Albert and Chib [1993], Gelfland and Smith [1990], and Tanner and Wong [1987]).

4.2 The Hierarchical Model

The hierarchical model (see Gelman, Carlin, Stern, and Rubin [1996]) is a generalization

of the regression model that allows each site to have its own value for the coefficients:

$$Y_{itj} \Big| \{x_{itj}\}_{\forall i,t,j}, \beta_j, \sigma^2 \sim N(\beta_j' x_{itj}, \sigma^2), \qquad (2)$$

where x_{ijt} is defined as in the previous section, except that we index the site as well,

$j=1,\dots,J$. Let $\beta_j' = (\beta_{j1} \cdots \beta_{jM})$, where $m=1,\dots,M$ indexes the regressors. The model

assumes a constant variance across sites. The key feature of the model is that the β šare

linked through a further model:

$$\beta_{jm} \Big| \{z_j\}_{j=1}^J, \gamma_m, \Sigma \sim N(\gamma_m' z_j, \Sigma), \qquad (3)$$

where z_j are a set of site characteristics used to model the site coefficients. The model for

β serves as a prior distribution with respect to the base model for earnings. The model is

completed by defining priors for the parameters:

$$1/\sigma^2 \sim W_1(r, Q^{-1}),$$

$$\Sigma^{-1} \sim W(\rho, K^{-1}),$$

and

$$vec(\gamma) \sim N(d, \Sigma \otimes D).$$

The prior on Σ^{-1} determines the degree of smoothing the model performs. The estimate

of the β's for each site are a precision-weighted average of the OLS estimates within

each site and the β's predicted by the model in (3). The weight, in turn, is influenced by

the prior for Σ^{-1}. The Wishart prior can be interpreted as ρ previous observations with

variance K. When K reflects high variance, this will pull up the estimate of Σ, and lead to

a greater weight being placed on the common prior for β's and a lower weight on the β

estimated within each site. Estimation is again undertaken using a Gibbs sampler

(outlined in Appendix B).

4.3 The Predictive Distribution

Since the object of interest for the policy question is earnings, and only indirectly the

parameters, we generate the predictive distribution, the distribution in the space of

outcomes that captures all of the uncertainty from the model, both intrinsic uncertainty

and parameter uncertainty. This distribution is simulated by repeatedly drawing for

parameter values from their posterior distribution and then drawing from the outcome

distribution conditional on observed data and parameters.

5. The Results

The model outlined in the previous section is implemented on the GAIN data, using age,

education, number and age of children, previous participation in a training program,

reading and writing test scores, ethnicity, and pre-treatment earnings as pre-treatment

individual characteristics. These are interacted with the treatment indicator, so that the

model allows for the site effect for treatment in control earnings to be different. The mean

characteristics of participants (including the mean number of children, mean reading

score, mean level of education, mean age, and mean pre-treatment earnings) are used as

the site characteristics. The Gibbs samplers outlined in the appendices produce estimates

of the posterior distribution of the parameters. These are then used to produce a

predictive distribution of earnings (under treatment and control) for each individual. The

predictive distributions are then averaged over the individuals at a site to produce an estimate of the site impact.

Tables 3 to 5 present the treatment impact for the 24 sites in the GAIN data with a pooled regression model, correlated random effects, and a range of hierarchical models. Note that the same non-hierarchical probit model is used as we vary the model for positive earnings.

5.1 Are site effects important in this data?

The first three sets of estimates in Table 3 allow us to consider to what extent site effects are obscured by pooling all observations. The pooled model (columns (1) and (2)) is a regression of positive earnings on a constant, pre-treatment covariates, and a treatment indicator (which is interacted with the covariates). Instead, the correlated random effects model (columns (3) and (4)) allows for a site-specific constant which depends on the mean characteristics of the participants within each site. The third set of estimates (columns (5) and (6)) is derived from the hierarchical framework, but ignores the meta-model that predicts site effects, focusing on the β_j within each site. The prior variance on the β's, K, is specified to be very low (diagonal elements of 10^{-5}), which means that the model engages in very minimal smoothing across sites. Each β_j is virtually identical to the estimate that would be obtained within each site.

In comparing the pooled estimates with those from the site-specific estimates (Figures 1 to 3), we observe that the results are broadly similar. For some sites (sites 1, 4, 12, 13, 20, 23, and 24) the pooled estimates of earnings are within $20 of the site-specific estimates. Many estimates are within $50 (sites 14 and 18 are exceptions), and almost all

are within $100 for both treatment and control earnings. In terms of the treatment impact, we note that the estimates are closer, with the biggest difference being $63 for site 24 (in San Diego). The estimated treatment impact from the correlated random effects models is even closer to the treatment impacts estimated within each site. The comparison suggests that pooling the data is not especially misleading in estimating the treatment impact for each site.

A more direct means to examine the importance of site effects is to vary the smoothing parameter in the hierarchical model The previous comparisons suggest that pooling the data (hence including out-of-site observations in the regression) does not significantly alter the estimated treatment effect. This is seen directly in Table 3, columns (7) and (8) (and Figure 4), in which the estimates of the hierarchical model with an extremely high degree of smoothing are presented; the elements of K are set to 10^6 along its diagonal. In the table and figure we note that the estimates are extremely similar when a high or low degree of smoothing is used, with all differences less than $20. In principle, by choosing an intermediate value for the smoothing prior we could achieve an intermediate degree of smoothing, but in this case clearly the results would not differ substantially. Thus for these data the question of how much smoothing to use is not a central one. Of course for other data sets the question could more substantial, and the ability to control the degree of smoothing would be more valuable.

5.2 Could site effects be important in other settings? Is there predictive uncertainty?
In the previous results, average earnings at each site are estimated using different individuals at the different sites. Instead, in Table 4, columns (1) and (2), we examine the

average earnings of the same individuals (participants from site 19, Alameda) when assigned to the different sites. The other sites' characteristics and the site coefficients are predicted using the characteristics of the individuals who originally participated in the program at that site. The thought experiment is to determine earnings for Alameda participants if, for example, they had entered the program in the environment of Riverside. As we vary the characteristics of the site we can see that there is variation in both estimated earnings and the treatment impact for these individuals (see Figure 5). The level of treatment (control) earnings varies from $190 ($138) in site 2 (site 2) to $1,157 ($970) in site 15 (site 14). The treatment effect varies from $52 in sites 2 and 13 to $293 in site 6. Even though the pooled model yields an accurate prediction for the treatment impact of site 19 participants when they are assigned to site 19, when we imagine re-assigning the same individuals to other sites, the site effect becomes important.

Another way in which the site effects can be important is their predictive uncertainty. If we were to re-implement the GAIN program how accurately could we predict what would happen at a given site? The second set of estimates in Table 4 (columns (3) and (4)) predict the site-specific parameters based on each site's characteristics. These are then used to predict the outcome for participants at each site. The relevant comparison is to the estimates in Table 3, columns (5) and (6), which ignore uncertainty in the site effects. The immediate observation is that the results are quite similar, always within $100 and typically within $40. At one level this may seem trivial: since the data for a given site are included in the estimation, it may not seem surprising that we are able to predict its outcome with accuracy. The result is not trivial because for each site we draw new site parameters based on the hierarchical model and base

predictions on these. So, for example, when we predict the outcome for site 6, the characteristics of the program participants in site 6 imply a set of site characteristics, which in turn produce a set of site parameters that lead to the average earnings we estimate.

However, we note that the range of uncertainty increases significantly (see Figures 6 and 7). Though the 95 percent posterior probability intervals of the estimates in Table 4, columns (3) and (4) (depicted in Figure 6) rarely fully overlap with the intervals in Table 3, columns (5) and (6), they are substantially wider. The increased uncertainty is meaningful in the sense that the treatment impact for many sites ceases to be significant. In the context of Table 3, in columns (5) and (6), we note that for 13 of the sites the 95 percent probability intervals for treatment and control earnings do not overlap, indicating a significant treatment impact, whereas this is true for only 6 sites in Table 4, columns (3) and (4). For example, for sites 2 to 5 (the Riverside sites) in Table 3, columns (5) and (6), the posterior 95 percent probability intervals do not overlap, but they do in Table 4, columns (7) and (8). Overall the comparison of the two sets of estimates suggests that when the site-specific parameters are re-estimated for each site, we succeed in replicating a profile of outcomes similar to those that are obtained for each site in isolation. However, uncertainty increases, and in some cases significantly.

5.3 Predicting site effects

The next question relating to site effects is whether we would be able to predict the outcomes at a site if we had not observed it in our data. To explore this issue the estimates in Table 5, columns (1) and (2), drop each site successively and use the

remaining sites to predict its outcome. The results are close to those in Table 3, columns (5) and (6), where the site data is included, within $30 for most site, except sites 3 to 5 (the Riverside sites) which are within $75. An important qualification to this result is that, even though we are excluding the site for which we are predicting the outcome, we include other sites from the county. Would it be possible to estimate the profile of treatment effects across sites if we exclude all of these observations from a county when estimating the model? The answer is presented in Table 5, columns (3) and (4) (and Figure 8). For most sites the predictions are worse than when other sites within the county are included, though with some exceptions the model is still able to predict the overall profile of treatment effects. The exceptions are the Riverside sites and two of the Los Angeles sites.

The failure accurately to predict the treatment effects for these sites illustrates the limitation of any model in extrapolating or predicting the treatment impact at a site which is significantly different from the sites observed in the sample. The Los Angeles sites differ substantially in terms of the number of children, which is higher than other sites, and pre-treatment earnings, which are lower than other sites. In contrast, the Riverside sites do not stand out in terms of their pre-treatment site characteristics. The differences with other sites seem to be along other, presumably qualitative, dimensions of the treatment applied. This inability to predict the Riverside treatment effects supports the view that Riverside differed from other counties in the approach it took to administering the treatment. Predictions based on other sites consistently under-estimate the treatment impacts in Riverside.

6. Conclusions

This paper has addressed three questions: (1) To what extent do site-specific effects matter in predicting the outcome for a given site? (2) Does predictive uncertainty regarding site effects influence the interpretation of the treatment effect? and (3) Would we be able to predict the outcome for a site, if its data were never observed. The answer to the first question was that site-specific effects are not especially important in predicting the outcome a given site for the GAIN data. By pooling data together, we only minimally distort the outcome for a particular site. Different degrees of smoothing across sites were explicitly considered and were found to have little impact on the estimated treatment effect, which reinforced this conclusion. The second and third questions are different because they deal with predictive uncertainty for subsequent implementations of the program. When making in-sample predictions, the model is able to predict the profile of site effects with reasonable accuracy. This amounts to saying that even the simple set of site-level characteristics that are used in the hierarchical model are sufficient to identify the distinct profile of site impacts in the GAIN data. However, we also find that the predictive uncertainty is significant: the treatment effect for many sites ceases to be significant when predictive uncertainty is incorporated into the estimate. Finally, when making out-of-sample predictions, the quality of the prediction was found to depend upon observing a sufficient number of sites similar to the site for which predictions are being made. For example, when dropping even some of the Riverside sites, the quality of the predictions for Riverside sites declines. This is not true for the Los Angeles sites when

dropped singly, but is also becomes true when all of the observations from Los Angeles are excluded.

The received wisdom regarding the GAIN program is that qualitative site-specific factors played an important role. Indeed the MDRC reports its findings separately for each county. The results presented here suggest that a simple set of site characteristics are sufficient to distinguish the various site-level effects. However the results also suggest that substantial extrapolation from the existing sites that are observed to new sites can potentially be misleading. For example, the Riverside treatments are consistently under-predicted when using data from the non-Riverside sites.

There are a number of possible extensions to this work. First, the set of site characteristics used were rudimentary, and could in principle be extended. It would be interesting to discover how much additional precision could be obtained. Second, the true economic significance of the different predictions from the range of models can only be assessed if they are linked to a decision problem. Would the added uncertainty when predicting site-level effects be sufficient to alter the policy-maker's decision regarding which program to choose? These are questions for on-going research.

References

Albert, J. and S. Chib (1993). "Bayesian Analysis of Binary and Polychotomous Response Data," *Journal of the American Statistical Association*, 88, 669-679.

Riccio, James, Daniel Friedlander, and Stephen Freedman (1994). GAIN: Benefits, Costs, and Three-Year Impacts of a Welfare-to-Work Program. New York: Manpower Demonstration Research Corporation.

Card, David, and Allan Krueger (1992). "Does School Quality Matter? Returns to Education and the Characteristics of Public Schools in the United States," *Journal of Political Economy,* 100, 1-40.

Chamberlain, Gary, and Guido Imbens (1996). "Hierarchical Bayes Models with Many Instrumental Variables," Harvard Institute of Economic Research, Paper Number 1781.

Cooper, Harris, and Larry Hedges (1994), editors. *The Handbook of Research Synthesis*. New York: Russell Sage Foundation.

Dehejia, Rajeev (1999). "Program Evaluation as a Decision Problem," National Bureau of Economic Research Working Paper No. 6954; http://nberws.nber.org/papers/W6954.pdf

Gelfland, A.E., and A.F.M. Smith (1990), "Sampling-Based Approaches to Calculating Marginal Densities," *Journal of the American Statistical Association*, 85, 398-409.

Gelman, Andrew, John Carlin, Hal Stern, and Donald Rubin (1996). *Bayesian Data Analysis*. London: Chapman and Hall.

Geweke, John, and Michael Keane (1996). "An Empirical Analysis of the Male Income Dynamics in the PSID: 1968-1989," University of Minnesota, unpublished.

Greenberg, David, and Michael Wiseman (1992). "What Did the OBRA Demonstrations Do?" Chapter 1, in Manski and Garfinkel (1992).

Heckman, James, and Jeffrey Smith (1996). "The Sensitivity of Experimental Impact Estimates: Evidence from the National JTPA Study," University of Western Ontario, unpublished.

Hotz, Joseph, Guido Imbens, and Julie Mortimer (1998). "Predicting the Efficacy of Future Training Programs Using Past Experiences". UCLA, unpublished.

Nelson, Doug (1997). "Some 'Best Practices' and 'Most Promising Models' for Welfare Reform," Memorandum, Annie E. Casey Foundation, Baltimore, MD; http://center.hamline.edu/mcknight/caseymemo.htm.

Manski, Charles, and Irwin Garfinkel (1992), editors. *Evaluating Welfare Training Programs*. Cambridge, MA: Harvard University Press.

Rossi, Peter, Robert McCulloch, and Greg Allenby (1995). "Hierarchical Modeling of Consumer Heterogeneity: An Application to Target Marketing," in C. Gatsonis, J. Hodges, R. Kass, and N. Singpurwalla (eds.), *Case Studies in Bayesian Statistics*, Volume II, *Lecture Notes in Statistics*, 105. New York: Springer-Verlag.

Tanner, M., and W. Wong (1987). "The Calculation of Posterior Distributions by Data Augmentation," *Journal of the American Statistical Association*, 82, 528-550.

Appendix A: Probit Model

Given diffuse priors for β and an arbitrary starting value $\beta^{(0)}$ the Gibbs sampling scheme is:

(1) Conditional on $\beta^{(j)}$, draw values for y_{it}^{**}: for $\{it : y_{it}^*=0\}$, from the negative portion of a normal distribution with mean $x_{it}\beta^{(j)}$ and variance 1, and for $\{it : y_{it}^*=1\}$, from the positive portion of the same distribution. Denote the filled-in dependent variable $y_{it,Z}^{(j+1)}$, so that $y_{i,Z}^{(j+1)} = (y_{i1,Z}^{(j+1)}\ldots y_{i\tau,Z}^{(j+1)})'$.

(2) Conditional on $Y_Z^{(j+1)} = (Y_{1Z}^{*(j+1)},\ldots,Y_{IZ}^{*(j+1)})''$, draw for $\beta^{(j+1)}$ from

$$N\left(\hat{\beta}^{(j+1)},(X'X)^{-1}\right),$$

where $\hat{\beta}^{(j+1)} = (X'X)^{-1}X'Y_Z^{(j+1)}$, with $X_i = (x_{i1}'\ldots x_{i\tau}')'$, and $X=(X_1',\ldots,X_I')'$.

From an arbitrary starting value, this is iterated 2000 times, producing $\left(Y_Z^{(j)},\beta^{(j)}\right)$. The first 500 iterations are discarded, leaving 1500 draws from posterior distribution of the parameters, which will be indexed $j=1,\ldots,1500$.

Appendix B: The Hierarchical Model

(1) $\beta_j^{(l)} \sim N(\overline{\beta}_j,V_\beta^{(l-1)})$, where $\overline{\beta}_j = (X_j'X_j\sigma_{(l-1)}^{-2} +\Sigma_{(l-1)}^{-1})^{-1}(X_j'X_j\sigma_{(l-1)}^{-2}\hat{\beta} +\Sigma_{(l-1)}^{-1}\beta^p)$, $\hat{\beta} = (X_j'X_j)^{-1}X_j'y_j$, $\beta^p = \gamma'_{(l-1)}z_j$, and $V_\beta = [X_j'X_j\sigma_{(l-1)}^{-2} +\Sigma_{(l-1)}^{-1}]^{-1}$,

(2) $1/\sigma_{(l-1)}^2 \sim \chi_{(n+r)}^2/(Q^{-1}+s^2)$, where $[s_{itj}] = y_{itj} - \beta_j^{(l)}x_{itj}$ and $s^2=s's$,

(3) $\Sigma_{(l)}^{-1} \sim W(J-M+\rho,(S+K^{-1})^{-1})$, where $S = \sum_{t=1}^J e'_t e_t$ and $e_t = \beta_t - \gamma_t z_t$ (the Mx1 vector of residuals for each site observation),

(4) $\gamma^{(l)} \sim N(\tilde{\gamma},\Sigma_{(l)}\otimes(Z'Z+D^{-1})^{-1})$, where $\gamma = (\gamma_1\cdots\gamma_M)'$, $\tilde{\gamma} = vec(\overline{\gamma})$, $\overline{\gamma} = ((Z'Z)+D^{-1})^{-1}(Z'Z\hat{\gamma}+D^{-1}d)$, and $\hat{\gamma}_j = (z_j'z_j)^{-1}z_j'\beta_j$.

This procedure produces a sequence of draws from the parameters, the first 500 of which we discard, leaving us with draws from the posterior distribution of the parameters.

Table 1: The Sample

	Alameda	Butte	Los Angeles	Riverside	San Diego	Tulare
GAIN:						
Treated Group	685	1717	3730	5808	8711	2693
Control Group	682	458	2124	1706	1810	1146
Total	1367	2175	5854	7514	10521	3839
Number of Sites	1	1	5	4	8	5

Notes: The GAIN sample sizes are from the public use file of the GAIN data. The AFDC total represents the number of AFDC cases (both single-parent and two-parent households) in the six evaluation counties in December 1990 (see Riccio, *et al.* (1994), Table 1.1).

Table 2: Site Characteristics and Outcomes from the GAIN Experiment

	Site (T=treatment C=control)	Number of children	Reading test score	Grade	Previous training	Hispanic	Black	Lagged earnings, 1 quarter before treatment	Lagged earnings, 2 quarters before treatment	Average post-treatment earnings (quarterly)	Quarterly treatment impact conditional on covariates
Butte (n= 2165)	Site 1										
	T	2.49	232	10.99	0.22	0.06	0.03	457	616	657	
	C	2.54	227	10.83	0.24	0.07	0.03	432	551	478	145
		(2165)	(4.56)	(0.14)	(0.02)	(0.01)	(0.01)	(54)	(67)	(57)	(50)
River-side (n= 7480)	Site2										
	T	2.69	231	10.80	0.23	0.24	0.18	388	498	642	
	C	2.69	228	10.68	0.24	0.24	0.17	380	500	501	149
		(0.08)	(2.67)	(0.10)	(0.02)	(0.02)	(0.02)	(44)	(51)	(42)	(39)
	Site3										
	T	2.81	232	10.66	0.28	0.18	0.10	329	442	642	
	C	2.91	227	10.65	0.25	0.20	0.07	359	491	445	216
		(0.10)	(4.28)	(0.14)	(0.02)	(0.02)	(0.02)	(53)	(63)	(55)	(48)
	Site4										
	T	2.87	231	9.59	0.14	0.56	0.09	478	592	693	
	C	2.77	227	9.57	0.14	0.58	0.07	548	718	532	190
		(0.13)	(3.55)	(0.21)	(0.02)	(0.03)	(0.02)	(74)	(83)	(66)	(56)
	Site5										
	T	2.65	231	10.96	0.22	0.18	0.20	332	471	560	
	C	2.44	226	11.02	0.22	0.19	0.15	381	489	348	292
		(0.14)	(4.89)	(0.17)	(0.04)	(0.03)	(0.03)	(93)	(101)	(82)	(72)
San Diego (n= 10485)	Site6										
	T	2.19	230	11.75	0.10	0.13	0.12	487	627	811	
	C	2.27	229	11.98	0.09	0.17	0.10	449	684	617	213
		(0.15)	(1.73)	(0.24)	(0.03)	(0.03)	(0.03)	(110)	(131)	(117)	(109)
	Site7										
	T	2.80	231	10.56	0.12	0.30	0.53	333	409	562	
	C	2.74	229	10.40	0.11	0.32	0.44	447	513	603	21
		(0.13)	(2.02)	(0.20)	(0.02)	(0.03)	(0.03)	(59)	(70)	(70)	(64)
	Site8										
	T	2.50	231	11.43	0.04	0.17	0.09	514	647	730	
	C	2.53	227	11.29	0.06	0.14	0.06	666	779	690	49
		(0.13)	(3.84)	(0.18)	(0.02)	(0.03)	(0.02)	(97)	(110)	(94)	(90)
	Site9										
	T	2.34	232	11.56	0.03	0.20	0.27	536	644	703	
	C	2.42	227	11.45	0.04	0.14	0.27	470	524	531	116
		(0.15)	(4.63)	(0.21)	(0.02)	(0.04)	(0.04)	(116)	(124)	(112)	(104)
	Site10										
	T	2.34	231	11.34	0.08	0.17	0.33	456	533	689	
	C	2.40	228	11.63	0.11	0.17	0.34	382	493	496	195
		(0.10)	(2.81)	(0.15)	(0.02)	(0.02)	(0.03)	(70)	(81)	(72)	(65)
	Site11										
	T	2.57	231	10.39	0.11	0.55	0.12	493	557	696	
	C	2.77	229	10.31	0.10	0.51	0.15	531	729	715	39
		(0.10)	(2.59)	(0.18)	(0.02)	(0.03)	(0.02)	(70)	(74)	(71)	(62)
	Site12										
	T	2.34	231	11.35	0.13	0.09	0.08	498	608	744	
	C	2.39	228	11.15	0.16	0.12	0.06	558	721	676	104
		(0.09)	(3.35)	(0.12)	(0.02)	(0.02)	(0.02)	(66)	(77)	(70)	(65)

Table 2: Site Characteristics and Outcomes from the GAIN Experiment (continued)

County	Site (T=treatment C=control)	Number of children	Reading tests core	Grade	Previous training	Hispanic	Black	Lagged earnings, 1 period before treatment	Lagged earnings, 2 periods before treatment	Average post-treatment earnings (quarterly)	Quarterly treatment impact conditional on covariates
San Diego (cont'd)	Site13										
	T	3.92	231	6.45	0.12	0	0.00	298	321	414	
	C	4.44	228	7.03	0.12	0	0	234	205	455	-90
		(0.25)	(2.61)	(0.44)	(0.03)	(0)	(0.01)	(55)	(60)	(84)	(77)
Tulare (n= 3835)	Site14										
	T	2.87	230	9.63	0.06	0.62	0.01	476	603	610	
	C	3.02	227	9.50	0.08	0.63	0.01	527	513	641	-15
		(0.19)	(2.7)	(0.32)	(0.03)	(0.05)	(0.01)	(103)	(104)	(96)	(87)
	Site15										
	T	3.03	232	9.45	0.25	0.43	0	573	665	612	
	C	3.33	226	9.21	0.22	0.49	0.01	366	435	515	25
		(0.20)	(5.47)	(0.33)	(0.04)	(0.05)	(0.01)	(103)	(112)	(87)	(77)
	Site16										
	T	3.04	231	9.47	0.25	0.34	0.01	439	521	530	
	C	3.05	228	9.38	0.20	0.29	0.01	440	530	444	80
		(0.14)	(3.23)	(0.22)	(0.03)	(0.03)	(0.01)	(67)	(74)	(61)	(53)
	Site17										
	T	2.98	232	10.08	0.26	0.40	0.11	403	504	558	
	C	3.05	226	10.36	0.19	0.38	0.09	597	716	691	-18
		(0.15)	(6.29)	(0.20)	(0.03)	(0.04)	(0.02)	(74)	(92)	(75)	(68)
	Site18										
	T	3.04	231	9.91	0.12	0.36	0.02	544	598	578	
	C	3.06	228	9.77	0.11	0.38	0.03	475	518	590	-2
		(0.14)	(3.85)	(0.26)	(0.02)	(0.04)	(0.01)	(100)	(104)	(79)	(62)
Alameda (n= 1360)	Site19										
	T	2.38	231	10.78	0.23	0.09	0.63	139	118	377	
	C	2.39	228	10.80	0.25	0.06	0.65	145	141	301	84
		(0.09)	(3.01)	(0.16)	(0.02)	(0.01)	(0.03)	(33)	(30)	(46)	(41)
Los Angeles (n= 5850)	Site20										
	T	3.54	232	9.36	0.20	0.40	0.11	136	145	381	
	C	3.44	223	9.55	0.14	0.30	0.13	160	157	301	108
		(0.16)	(8.63)	(0.25)	(0.03)	(0.03)	(0.02)	(34)	(36)	(52)	(49)
	Site21										
	T	4.20	230	7.84	0.11	0.28	0.09	178	193	340	
	C	4.45	228	7.38	0.12	0.24	0.07	199	178	253	80
		(0.18)	(2.32)	(0.27)	(0.02)	(0.03)	(0.02)	(31)	(34)	(40)	(37)
	Site22										
	T	3.21	230	9.54	0.05	0.22	0.47	175	179	301	
	C	3.25	227	9.17	0.05	0.18	0.41	173	162	311	-6
		(0.12)	(3.68)	(0.20)	(0.01)	(0.02)	(0.03)	(33)	(32)	(43)	(39)
	Site23										
	T	3.72	231	9.69	0.17	0.15	0.64	165	139	309	
	C	4.07	229	9.49	0.15	0.16	0.56	146	121	299	-7
		(0.11)	(1.5)	(0.17)	(0.02)	(0.02)	(0.02)	(30)	(26)	(36)	(32)
	Site24										
	T	3.73	231	7.61	0.16	0.78	0.06	62	67	210	
	C	3.84	228	7.39	0.17	0.73	0.06	137	152	269	-50
		(0.17)	(3.13)	(0.29)	(0.03)	(0.03)	(0.02)	(33)	(37)	(46)	(44)

Table 3: Estimated Site Earnings (Quarterly)

Coun-ty		Pooled		Correlated Random Effects		Hierarchical Model, No smoothing		Hierarchical Model, Maximum smoothing	
		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
		Treatment	Control	Treatment	Control	Treatment	Control	Treatment	Control
Butte	Site 1	668	541	659	541	665	532	663	532
		[641,696]	[512,571]	[631,686]	[513,572]	[631,700]	[495,571]	[621,707]	[493,578]
River-side	Site2	623	495	609	492	578	501	579	496
		[599,645]	[472,518]	[586,630]	[469,515]	[549,605]	[472,528]	[549,612]	[465,528]
	Site3	587	463	570	457	624	502	641	518
		[562,615]	[437,489]	[545,598]	[432,484]	[591,661]	[464,540]	[598,688]	[472,566]
	Site 4	622	519	619	526	633	494	662	499
		[587,657]	[487,551]	[584,654]	[493,557]	[596,678]	[456,534]	[616,713]	[449,549]
	Site 5	629	492	658	523	587	465	639	502
		[581,678]	[449,533]	[608,709]	[477,568]	[533,648]	[413,522]	[573,716]	[434,571]
San Diego	Site 6	724	575	761	614	778	621	830	701
		[674,774]	[536,624]	[708,813]	[573,668]	[711,851]	[546,696]	[752,913]	[603,812]
	Site 7	555	444	524	428	500	432	510	450
		[523,586]	[414,478]	[495,553]	[399,461]	[465,539]	[390,474]	[469,555]	[403,499]
	Site 8	714	576	724	594	669	542	644	524
		[677,756]	[538,614]	[687,768]	[554,633]	[624,714]	[496,594]	[590,703]	[468,583]
	Site 9	692	554	681	554	648	494	670	511
		[640,748]	[508,603]	[628,735]	[508,603]	[585,710]	[432,563]	[594,739]	[434,600]
	Site 10	641	507	672	540	608	437	627	452
		[610,673]	[478,535]	[639,705]	[509,570]	[571,644]	[401,472]	[583,676]	[404,500]
	Site 11	645	528	673	560	675	619	679	645
		[614,676]	[502,553]	[640,704]	[532,586]	[637,716]	[574,664]	[639,725]	[594,696]
	Site 12	704	564	735	598	722	572	725	576
		[671,735]	[534,594]	[700,767]	[566,631]	[679,766]	[530,621]	[673,780]	[527,624]
	Site 13	404	319	339	274	413	342	384	350
		[367,442]	[286,353]	[308,371]	[246,304]	[372,464]	[290,394]	[336,437]	[290,413]
Tu-lare	Site 14	646	549	631	546	551	513	556	541
		[585,710]	[498,599]	[572,692]	[496,597]	[495,609]	[458,575]	[487,630]	[468,623]
	Site 15	642	527	653	545	584	493	551	469
		[589,697]	[477,579]	[599,708]	[494,601]	[528,644]	[436,562]	[483,625]	[400,546]
	Site 16	602	497	596	500	571	486	557	472
		[564,639]	[463,534]	[559,633]	[466,537]	[525,617]	[440,535]	[506,609]	[419,522]
	Site 17	652	537	652	545	599	533	577	520
		[610,699]	[497,576]	[610,698]	[505,586]	[551,650]	[487,586]	[519,632]	[461,580]
	Site 18	644	524	686	569	584	504	586	500
		[599,695]	[485,568]	[638,741]	[526,617]	[537,634]	[455,560]	[523,656]	[444,564]
Ala-meda	Site 19	457	328	412	302	387	267	354	244
		[427,491]	[301,356]	[384,442]	[277,327]	[355,425]	[241,298]	[316,394]	[216,271]
LosAn-geles	Site 20	410	322	355	287	419	295	398	264
		[373,447]	[290,354]	[324,386]	[258,315]	[376,461]	[256,335]	[351,451]	[227,304]
	Site 21	386	317	287	243	327	287	326	277
		[352,421]	[284,348]	[263,313]	[218,267]	[294,360]	[253,325]	[288,368]	[236,324]
	Site 22	395	308	341	273	372	313	370	314
		[369,424]	[282,334]	[318,366]	[250,296]	[344,405]	[281,349]	[330,408]	[278,354]
	Site 23	380	282	337	257	364	293	360	295
		[356,403]	[260,306]	[317,357]	[237,279]	[335,393]	[265,326]	[332,392]	[263,329]
	Site 24	321	254	274	223	326	322	361	349
		[288,357]	[224,286]	[246,305]	[196,251]	[286,367]	[271,373]	[310,420]	[291,410]

Table 4: Estimated Site Earnings (Quarterly)

County		Alameda Participants in Other Sites		Hierarchical Model, Predicting Site Effect	
		(1) Treatment	(2) Control	(3) Treatment	(4) Control
Butte	Site 1	310 [282,343]	225 [200,253]	659 [607,713]	532 [488,586]
River-side	Site2	190 [170,209]	138 [121,156]	595 [540,650]	495 [443,546]
	Site3	310 [277,344]	228 [196,261]	567 [510,627]	481 [426,537]
	Site 4	443 [400,486]	351 [309,394]	579 [521,634]	512 [461,566]
	Site 5	898 [805,1000]	645 [564,734]	607 [538,682]	507 [443,577]
San Diego	Site 6	898 [794,1000]	605 [524,691]	708 [624,813]	540 [467,625]
	Site 7	440 [405,478]	326 [293,363]	540 [499,585]	455 [418,496]
	Site 8	601 [540,659]	422 [371,474]	689 [627,762]	561 [495,628]
	Site 9	1007 [893,1127]	689 [601,772]	680 [603,758]	529 [463,606]
	Site 10	359 [325,395]	247 [219,279]	620 [565,674]	475 [429,526]
	Site 11	303 [277,327]	223 [200,249]	611 [569,658]	514 [478,554]
	Site 12	354 [319,385]	245 [217,277]	680 [622,744]	536 [485,596]
	Site 13	877 [766,1009]	825 [658,993]	392 [321,466]	308 [253,375]
Tu-lare	Site 14	1188 [1076,1316]	970 [838,1110]	604 [531,684]	546 [471,625]
	Site 15	1157 [1049,1273]	943 [821,1068]	622 [561,698]	549 [486,612]
	Site 16	592 [532,653]	473 [410,537]	595 [537,661]	518 [460,578]
	Site 17	730 [652,803]	583 [507,667]	646 [580,715]	555 [495,626]
	Site 18	739 [677,805]	552 [489,616]	633 [578,698]	537 [483,594]
Ala-meda	Site 19	458 [393,532]	315 [257,375]	457 [389,530]	315 [261,377]
LosAn-geles	Site 20	698 [622,783]	536 [468,612]	394 [347,442]	319 [272,370]
	Site 21	673 [566,784]	574 [468,692]	345 [285,416]	311 [254,380]
	Site 22	390 [349,440]	301 [261,346]	371 [330,415]	306 [268,350]
	Site 23	335 [285,391]	252 [208,301]	394 [338,454]	300 [248,360]
	Site 24	672 [561,804]	573 [447,715]	319 [256,388]	311 [233,411]

Table 5: Estimated Site Earnings (Quarterly)

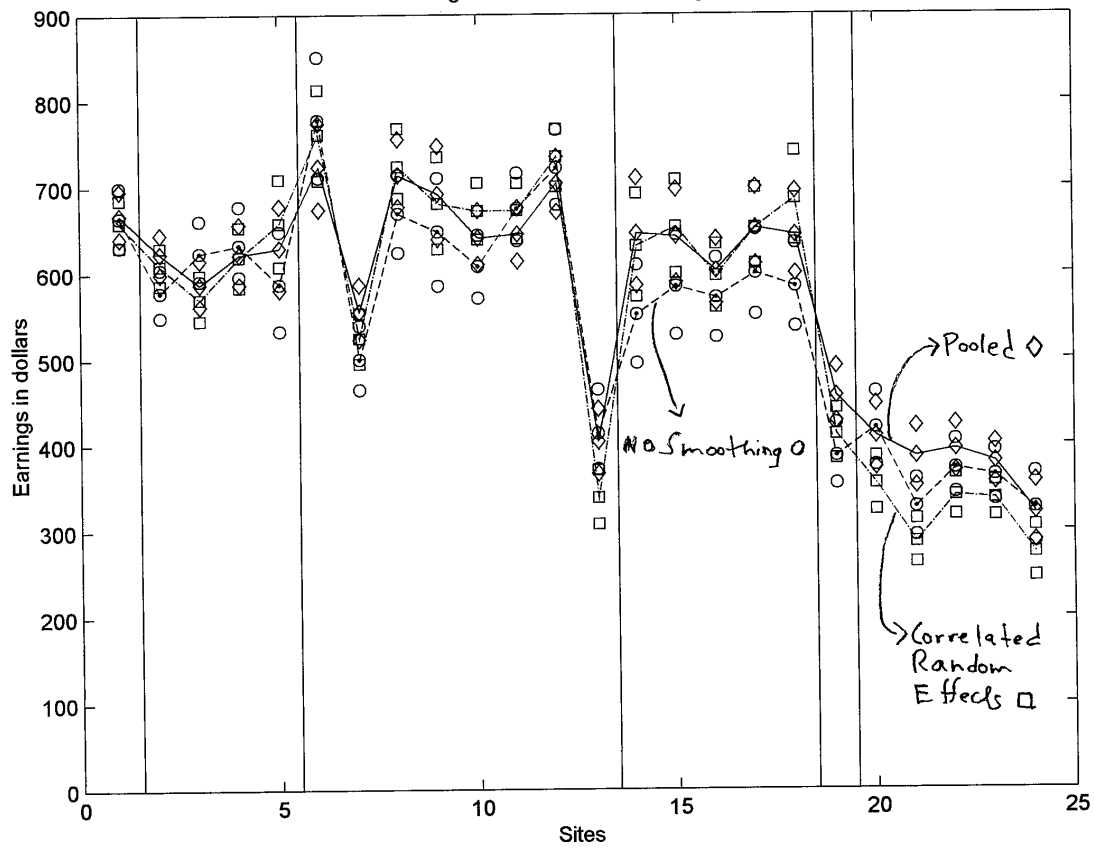| County		Predicting site effect, dropping that site		Dropping all sites for that county	
		(1)	(2)	(3)	(4)
		Treatment	Control	Treatment	Control
Butte	Site 1	670 [613,729]	539 [481,605]	670 [613,729]	539 [481,605]
River-side	Site2	593 [537,652]	495 [438,554]	556 [491,630]	487 [416,572]
	Site3	546 [491,603]	467 [413,523]	532 [460,610]	479 [403,569]
	Site 4	564 [504,631]	512 [450,583]	514 [459,567]	492 [433,555]
	Site 5	614 [546,695]	527 [463,604]	588 [507,679]	525 [431,630]
San Diego	Site 6	701 [623,791]	540 [464,634]	648 [556,747]	498 [410,584]
	Site 7	539 [497,577]	456 [416,501]	555 [514,600]	455 [416,500]
	Site 8	701 [628,774]	573 [508,642]	677 [607,748]	548 [481,619]
	Site 9	685 [608,777]	543 [472,636]	638 [562,724]	495 [428,566]
	Site 10	619 [567,670]	484 [434,534]	587 [526,651]	450 [391,503]
	Site 11	605 [557,656]	504 [454,551]	624 [578,675]	501 [457,543]
	Site 12	701 [646,764]	550 [502,603]	658 [592,727]	516 [451,574]
	Site 13	361 [294,446]	289 [225,368]	420 [351,499]	331 [272,401]
Tu-lare	Site 14	618 [548,699]	554 [486,632]	666 [581,753]	599 [510,694]
	Site 15	628 [559,700]	569 [498,645]	686 [599,770]	595 [519,678]
	Site 16	614 [540,692]	533 [471,599]	656 [587,733]	565 [492,644]
	Site 17	539 [497,577]	456 [416,501]	710 [629,791]	606 [517,698]
	Site 18	701 [628,774]	573 [508,642]	684 [620,747]	567 [508,632]
Ala-meda	Site 19	685 [608,777]	543 [472,636]	685 [608,777]	543 [472,636]
LosAn-geles	Site 20	399 [350,452]	335 [291,390]	430 [329,556]	382 [276,509]
	Site 21	362 [287,454]	331 [253,449]	420 [269,618]	480 [279,755]
	Site 22	387 [336,442]	307 [262,357]	388 [312,474]	337 [253,434]
	Site 23	445 [361,543]	309 [238,389]	471 [343,626]	379 [261,541]
	Site 24	306 [234,396]	259 [180,359]	336 [240,452]	313 [203,441]

Figure 1: Treatment Earnings

Figure 2: Control Earnings

Figure 3: Treatment Effect

Figure 3: Treatment Effect

Annotations on figure: "Pooled ◊", "No Smoothing", "Correlated Random Effects □"
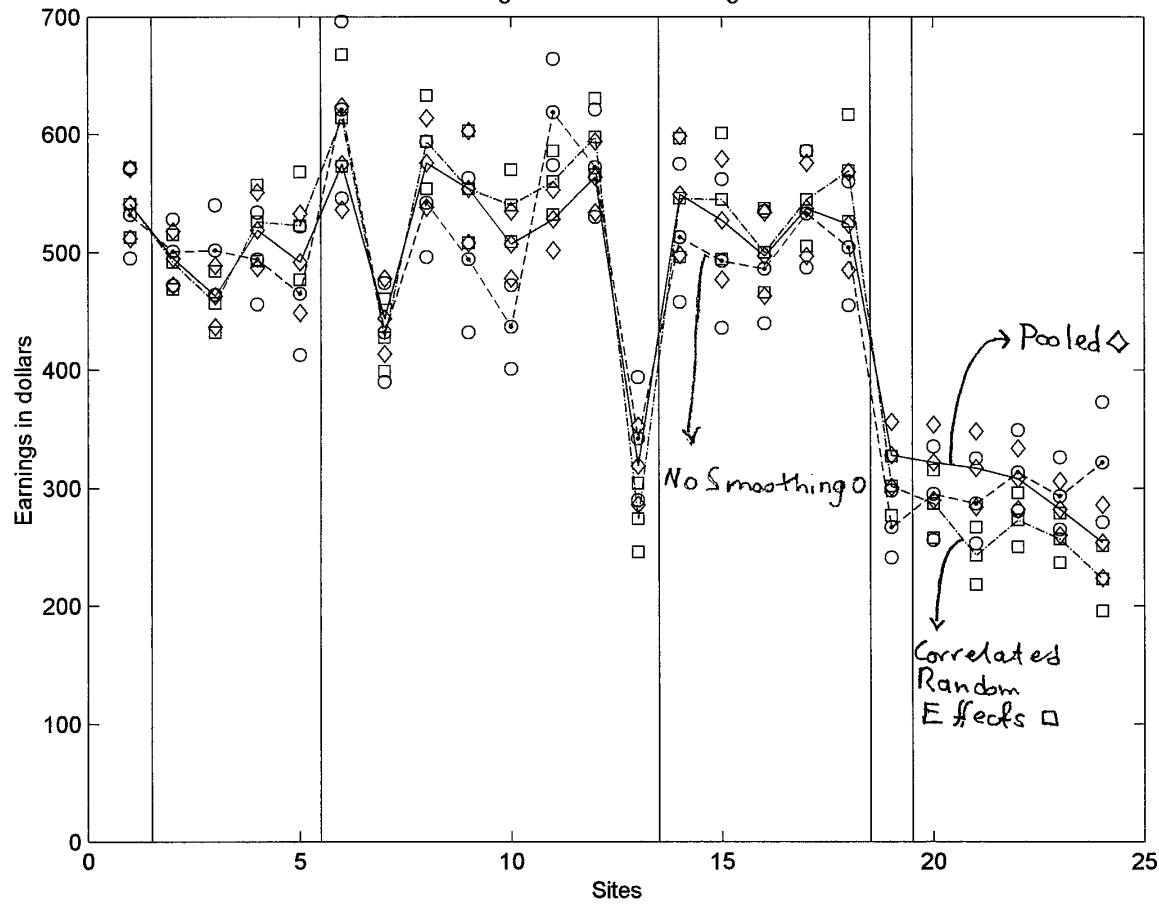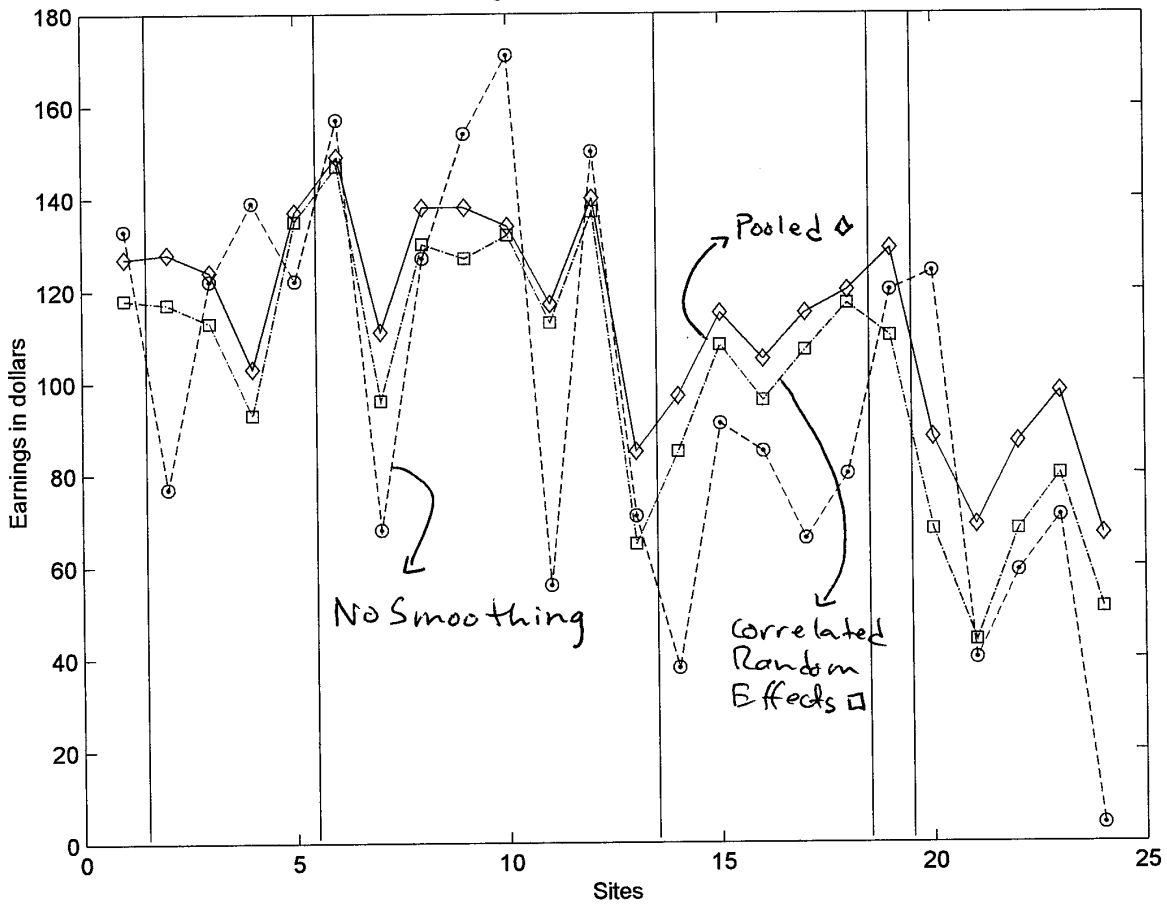
X-axis: Sites

Y-axis: Earnings in dollars

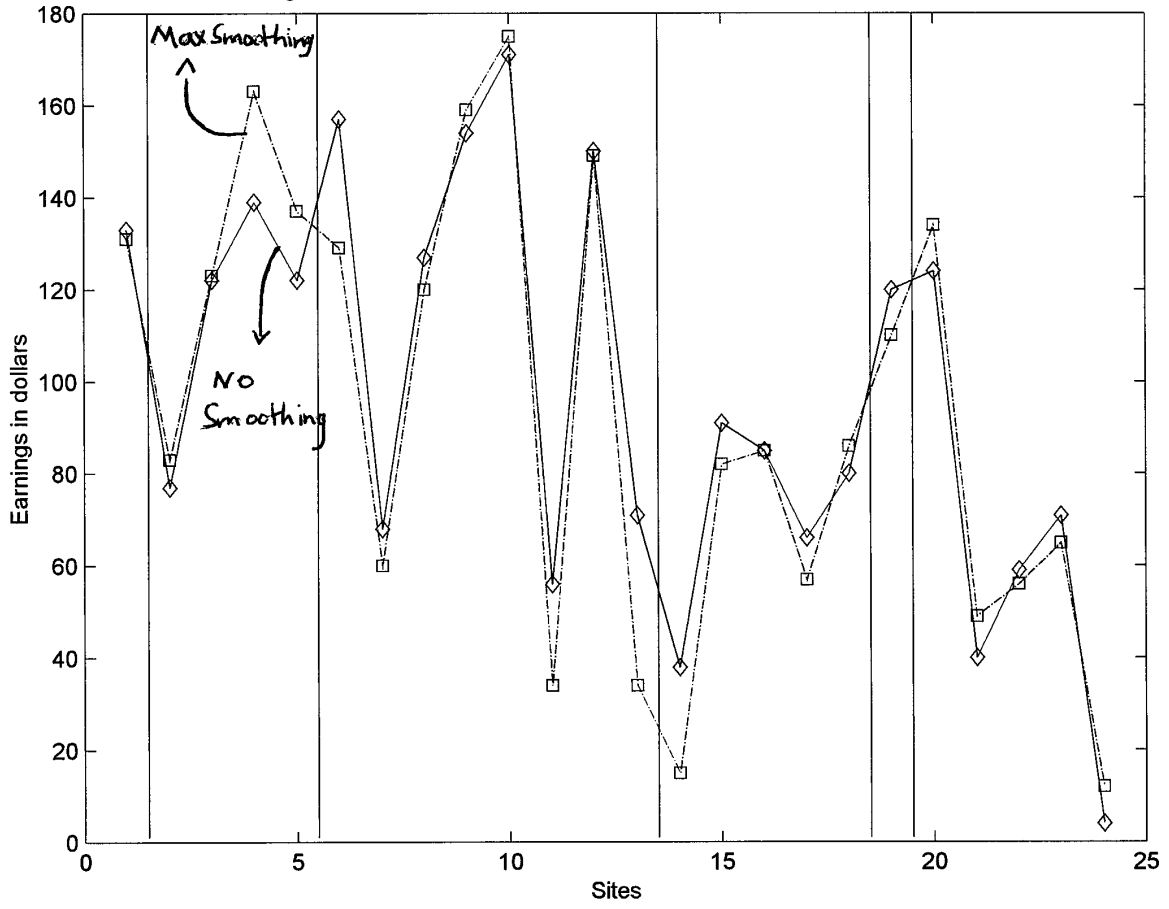Figure 4: Treatment Effect, Maximum Smoothing vs. No Smoothing

Figure 5: Alameda Participants in Other Sites
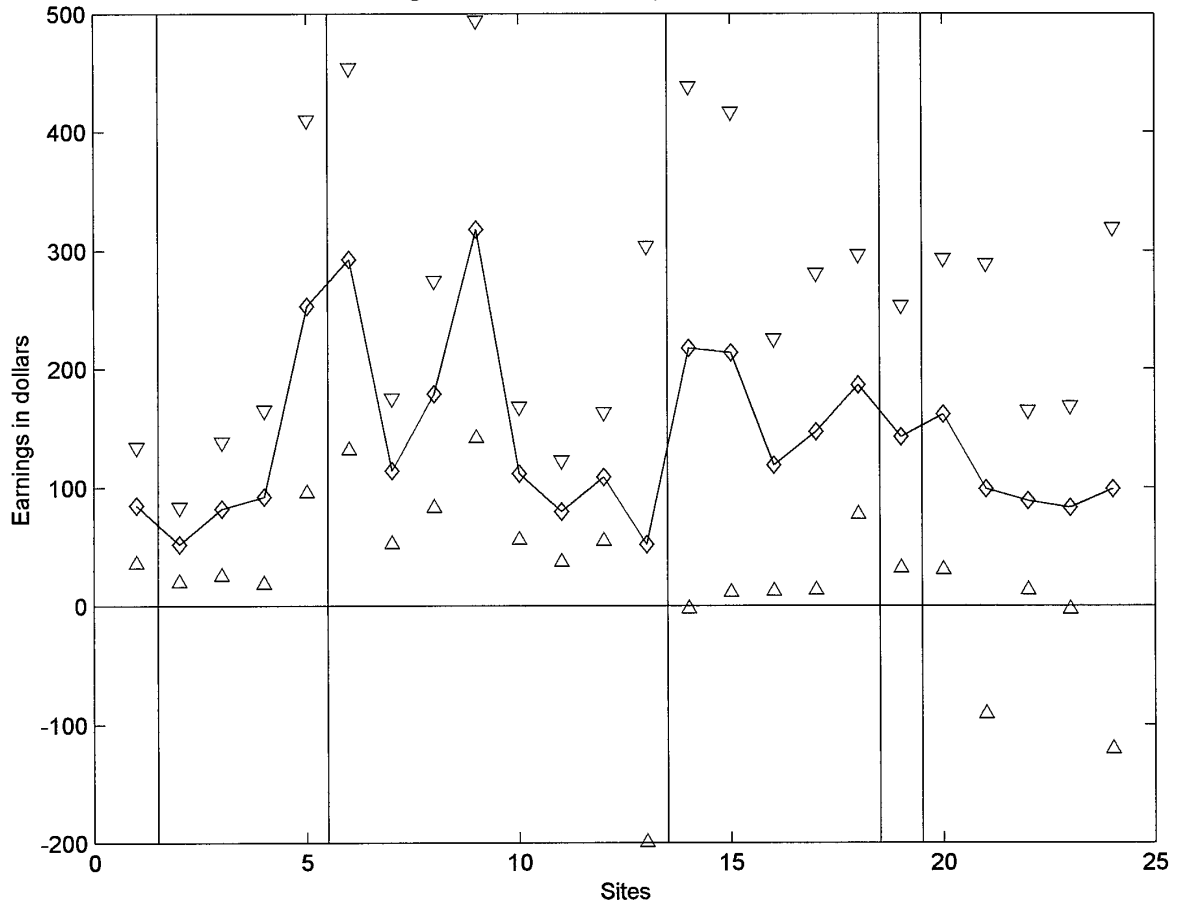
Figure 6: Treatment and Control Earnings, Given Site Effects
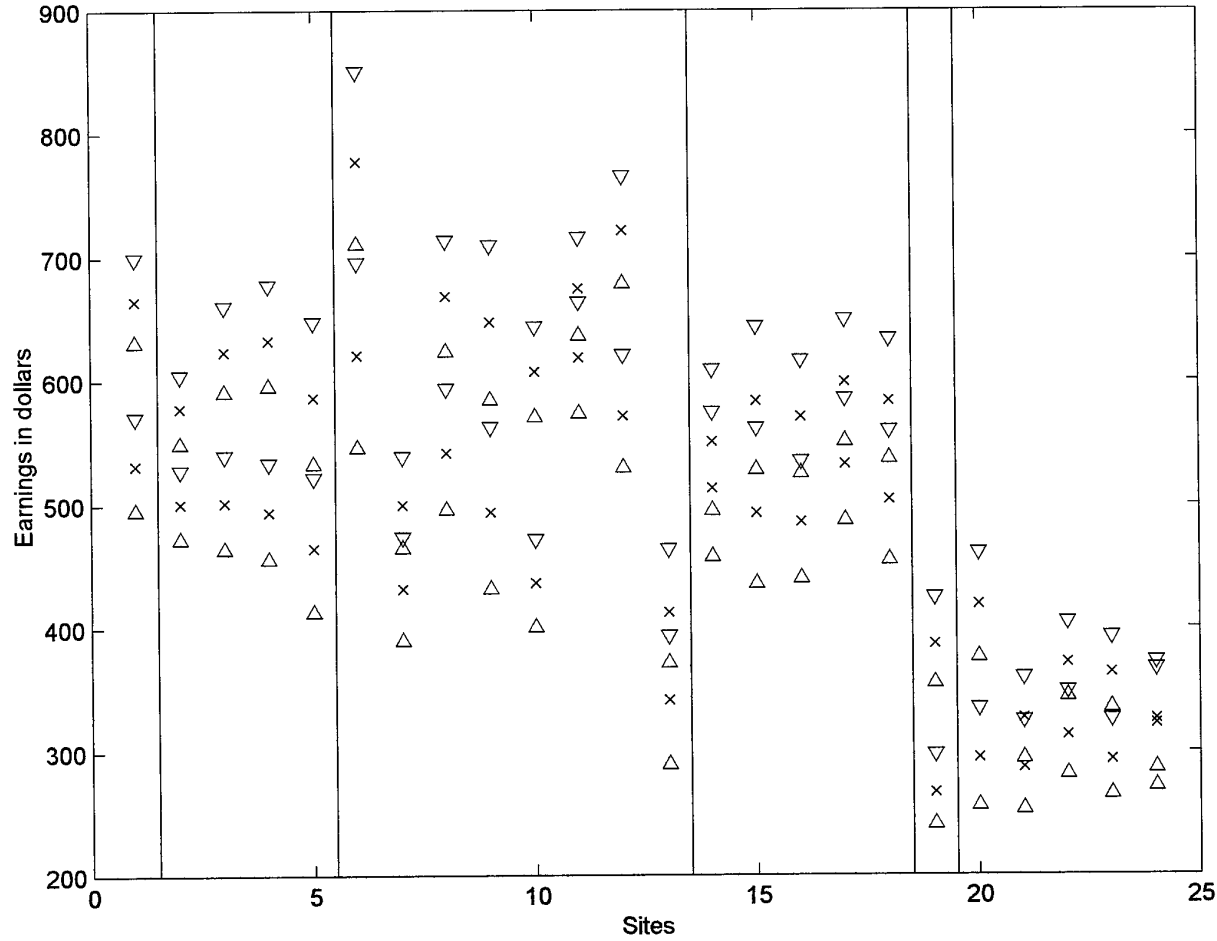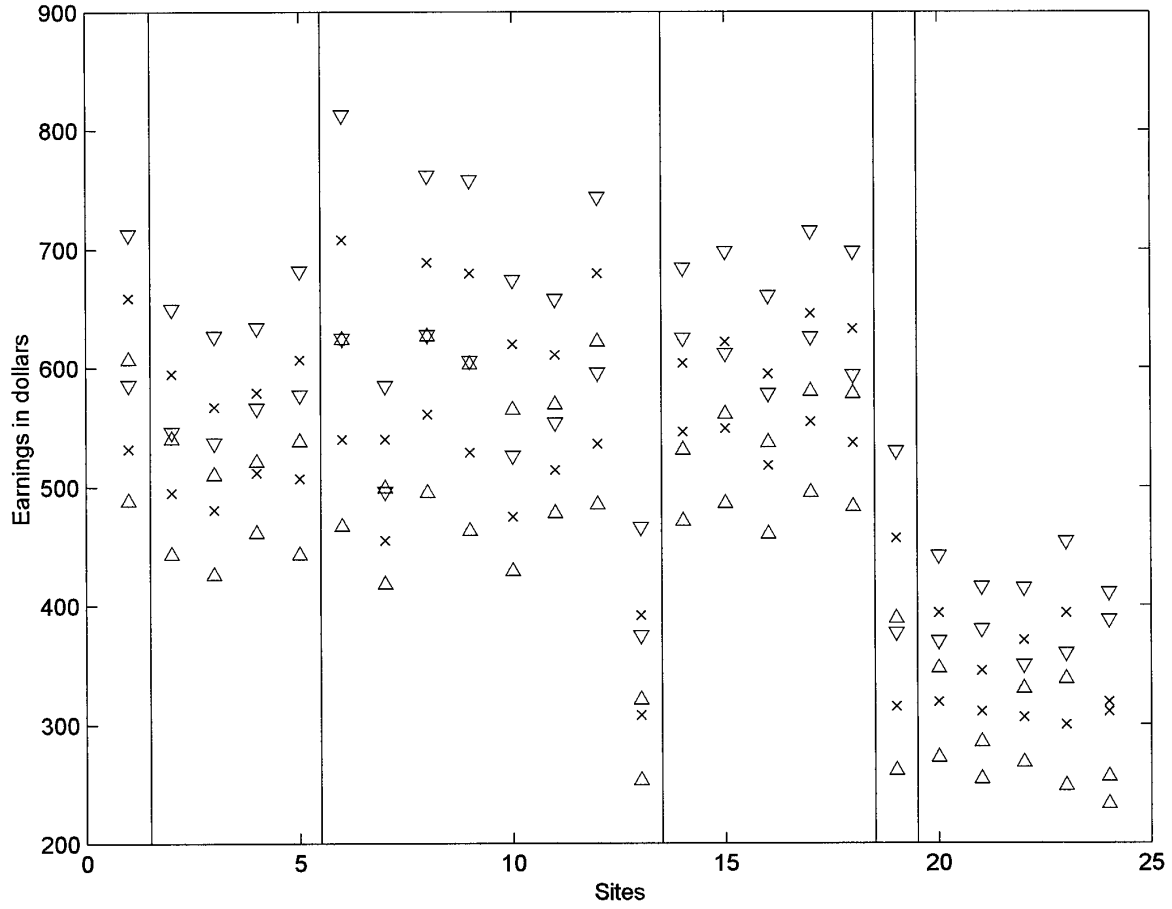
Figure 7: Treatment and Control Earnings, Predicting Site Effects

Figure 8: Predicting Site Effects

Figure 8: Predicting Site Effects