Alleviating Ecological Bias in Voter Turnout Models (and other Generalized Linear Models) with Optimal Subsample Design

Adam N. Glynn[*] Jon Wakefield[†] Mark S. Handcock[‡] Thomas S. Richardson[§]

March 4, 2009

Abstract

In this paper, we illustrate that combining ecological data with subsample data in situations in which a generalized linear model (GLM) is appropriate provides two main benefits. First, by including the individual level subsample data, the biases associated with ecological inference in GLMs can be eliminated. Second, available ecological data can be used to design optimal subsampling schemes, so as to maximize information about parameters. We present an application of this methodology to the estimation of regional voter turnout rates across racial groups, showing that small, optimally chosen subsamples can be combined with ecological data to generate precise estimates relative to a simple random subsample.

1 Introduction

Ecological inference (the attempt to make inference about individuals with aggregate data) is well known to be problematic in a number of situations, and may often lead to biased estimates. One solution to this problem is to obtain individual level data, and to make inference based on this data alone. However, it would be inefficient to ignore the ecological data, since such data are usually aggregated from a large number of individuals and therefore provide a great deal of information about some function of the parameters. Furthermore, the ecological data may inform the sampling of individual level data, effectively reducing the cost of data collection. In this paper, we consider these issues within the framework of generalized linear models, specifically addressing the following questions: When does ecological inference lead to bias? How can we correct ecological bias by combining ecological data and subsample data? Finally, if the ecological data are readily available (as is frequently the case), can we design an optimal subsample of individual level

[*] Department of Government and The Institute for Quantitative Social Sciences Harvard University, 1737 Cambridge Street, Cambridge, MA 02138. aglynn@iq.harvard.edu

[†] Department of Statistics and Department of Biostatistics, University of Washington, Seattle, WA 98195.

[‡] Department of Statistics, University of Washington, Seattle, WA 98195.

[§] Department of Statistics, University of Washington, Seattle, WA 98195.

data in order to maximize the information about parameters?

The relevant literature can be divided into three groups: those articles that solely describe ecological inference, those that describe combined inference for aggregate and individual level data, and those that describe sampling design in the combined framework. The literature on ecological inference is quite large as it spans a number of disciplines. Reviews of this literature can be found in Wakefield (2004) and Salway and Wakefield (2004). The literature on combined inference is considerably smaller and can be separated into those techniques that treat the ecological data as population level information and those that do not. When treated as population level information, ecological data are often used in either in a generalized method of moments framework with the ecological data providing extra moment conditions (Imbens and Lancaster, 1994; Hellerstein and Imbens, 1999), or in a likelihood framework where the ecological data provide constraints on the maximization of the likelihood (Handcock et al., 2003; Chaudhuri et al., 2006). When treated as non-population level information, ecological data have been combined with individual level data in a variety of models. Wakefield (2004) uses a likelihood approach for 2×2 tables, while Steel et al. (2004) develops the observed information for this same case, and Haneuse and Wakefield (2008) shows that this approach can be adapted to case-control data. Jackson et al. (2006) develops a model for the combination of ecological and individual level data when logistic regression is appropriate at the individual level. In hierarchical linear models, Raghunathan et al. (2003) considers moment and maximum likelihood estimates of a common within group correlation coefficient, and Steel et al. (2003) considers many variations of aggregate and individual data combinations, developing the properties of moment estimators in this framework. The literature on sampling design in the combined framework has received relatively little attention to date, despite the costs associated with individual level sampling. In the 2×2 table framework, Wakefield (2004) and Haneuse and Wakefield (2008) show that, conditional on the ecological data, rare case observations contain most of the information. For linear models, Glynn et al. (2008) showed that ecological bias can be decomposed into terms based on group intercepts, group slopes, and confounding, and that optimal subsampling design conditional on ecological data can eliminate the bias from these sources while maximizing precision.

In this paper we extend the results of Glynn et al. (2008) to generalized linear models, showing that the problems of ecological bias are similar in this more general framework, and that subsample data can be used to correct this bias in a combined likelihood approach. However, the specification of ecological bias, and the specification of a combined likelihood are complicated by the non-linearity of these models, and therefore we frame much of the discussion in a Poisson regression example with two binary covariates. The non-linearity of these models causes further problems for optimal design, because it is rarely possible to specify the necessary information equations in closed form. Again, we address the question of optimal design in the context of Poisson regression with two binary covariates and in the context of logistic regression with a single binary covariate. The paper is organized in the following manner. In Section 2 we discuss ecological bias in generalized linear models. Section 3 demonstrates the problems of ecological bias in generalized linear models with an illustrative example based on regional voter turnout data in the US. In Section 4 we describe a method for the alleviation of ecological bias through combined data maximum likelihood estimation for the generalized linear model, and we provide an example of Poisson regression with two binary covariates. In Section 5, we describe optimal subsample design conditional on the ecological data; we also provide a number of design examples; including a general look at subsample design for the linear model and some specific design questions for Poisson and logistic regression. Section 6 presents an application of the combined estimation and design results to the US regional voter turnout data. We conclude the paper with a discussion of possible applications and future work in Section 7.

2 Ecological Bias in Generalized Linear Models

Within the framework of generalized linear models, ecological bias is unavoidable in all but a small number of special cases. We first define the data at the individual level, assuming that we could potentially observe the triples $(x_{ij}, y_{ij}, z^c_{ij})$ for individuals $j = 1, ..., n_i$ in groups $i = 1, ..., m$, where $\boldsymbol{y_i} = (y_{i1}, ..., y_{in_i})$ is the vector of responses from group i, $\boldsymbol{x_i} = (x_{i1}, ..., x_{in_i})$ is the vector of univariate exposure/covariates from group i, $\boldsymbol{z^c_i} = (z^c_{i1}, ..., z^c_{in_i})$ is the vector of confounders, and $n = \sum_{i=1}^m n_i$ represents the "full data" sample size. We assume that we observe ecological data that consists of the group averages $(\overline{x}_i, \overline{y}_i)$ for groups $i = 1, ..., m$, and we may observe $\overline{z}^c_i$ for groups $i = 1, ..., m$. Furthermore, we assume that the n_i observed triples in group i represent i.i.d. values produced by some process, and we are interested in the parameters of this process. Within this framework, we consider one of three models:

$$
\begin{aligned}
E[y_{ij}|\boldsymbol{x_i}, \boldsymbol{z^c_i}] &= g(\beta_{0i} + \beta_w x_{ij}) && (1)\\
E[y_{ij}|\boldsymbol{x_i}, \boldsymbol{z^c_i}] &= g(\beta_{0i} + \beta_{wi} x_{ij}) && (2)\\
E[y_{ij}|\boldsymbol{x_i}, \boldsymbol{z^c_i}] &= g(\beta_{0i} + \beta_{wi} x_{ij} + z^c_{ij}), && (3)
\end{aligned}
$$

where $g(\cdot)$ is a non-linear, monotonic link function. In (1), we assume that the linear predictor of each group has a different intercept, but a common within-group slope. For the remainder of this paper, we will refer to β_{0i} and β_w as intercepts and slopes, even though they are not intercepts and slopes in the linear model sense. In (2), we assume that each group may have distinct intercepts and slopes. In (3), we assume that in addition to having distinct intercepts and slopes, z^c_{ij} acts as a confounder so that $E[z^c_{ij}|x_{ij}] \neq E[z^c_{ij}]$. We do not parametrize the final term as it represents the combination of all possible confounding variables and their effects, i.e. $z^c_{ij} = \sum_{k=1}^K \beta_{ki} z_{ijk}$. These three models are nested, in that (1) is a special case of (2), which is a special case of (3).

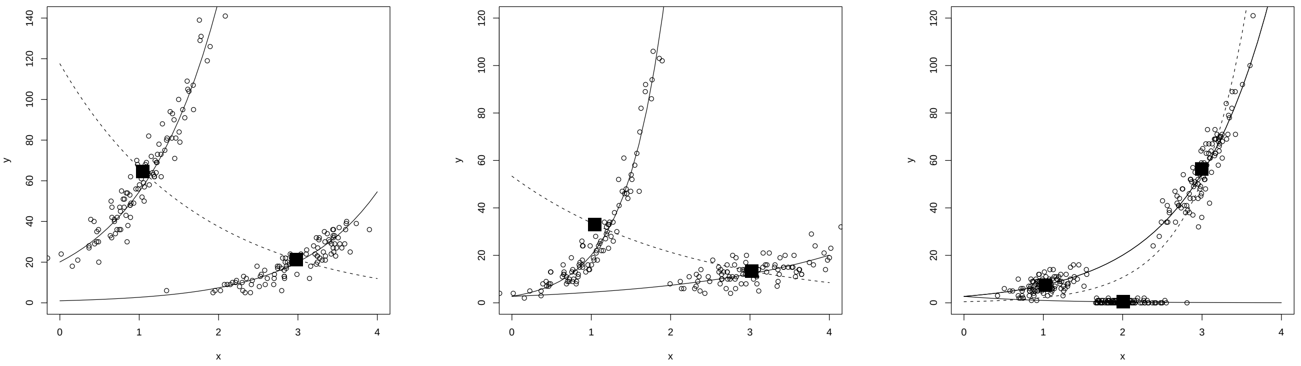
In general, the non-linearity of these models prevents the explicit derivation of their ecological counterparts, and in general, the derived form will not have the same link function. We define naive ecological models as,

$$\begin{aligned} E[\overline{y}_i|\overline{x}_i] &= g(\beta^e_{0i} + \beta^e_w \overline{x}_i) \quad (4)\\ E[\overline{y}_i|\overline{x}_i] &= g(\beta^e_{0i} + \beta^e_{wi} \overline{x}_i) \quad (5)\\ E[\overline{y}_i|\overline{x}_i, \overline{z}_i] &= g(\beta^e_{0i} + \beta^e_{wi} \overline{x}_i + \overline{z}^e_i) \quad (6) \end{aligned}$$

The non-linearity of $g(\cdot)$ causes a number of problems for ecological inference. First, the equations above show that ecological inference will not identify the individual level model parameters, and therefore we have little hope of an unbiased ecological estimate. Second, for the models (2) and (3), it is difficult to select a single parameter of interest. We might assume that some combination of the β_{wi} parameters will be of interest, but the exact combination will depend on the link function and the research question. Finally, the ecological estimate will not in general have a closed form solution, and therefore, we cannot explicitly derive the bias. However, we will see from examples that the sources of linear ecological bias considered in Glynn et al. (2008) (correlated intercepts, correlated slopes, quadratically related slopes, and confounding) will result in bias for GLMs. Using a Poisson regression with $g(\cdot) = \exp(\cdot)$, Figure 1(a) shows an example of (1) with two groups where the intercepts for each group are negatively correlated with the covariate averages for each group. The ecological data are represented by the solid squares, and it is clear, that the slope of the regression curve through these points has a negative slope, as opposed to the positive slopes within the individual groups. Figure 1(b) shows an example of (2) with two groups where the intercepts happen to be equal and the slopes for each group are negatively correlated with the covariate averages for each group. The ecological data are represented by the solid squares, and it is clear that the slope of the regression curve through these points will have a negative slope, as opposed to the positive slopes within each of the two individual groups. As discussed previously, it may be difficult to specify the exact combination of the two slopes that would be of interest to the researcher. However, this negative ecological slope will not correspond to any reasonable combination of these two positive slopes. Figure 1(c) again shows an example of (2), but with three groups in which the slopes for each group are quadratically related but uncorrelated with the covariate averages for each group. The slopes (and intercepts) for the two outer groups are the same and positive, while the slope for the middle group is negative. The ecological Poisson regression, represented by the dashed line, results in a slope which is greater than all three, and hence cannot represent any reasonable combination of the three.

It is not surprising that the sources of linear ecological bias can lead to bias in generalized linear models since $g(\cdot)$ is monotonic and can be approximated by a linear function over some ranges of the linear predictor space. However, even in the absence of the sources of bias in Figure 1, ecological bias in GLMs can still occur. We demonstrate this using a simplified model with common parameters across groups.

Suppose we assume a simplified model with common slopes and intercepts across groups. The individual



(a)

(b)

(c)

Figure 1: Sources of ecological bias: (a) Intercept bias, group intercepts correlated with covariate group means. (b) Slope bias, group slopes correlated with covariate group means. (c) Slope bias, group slopes quadratically related to covariate group means. Solid lines are the within group Poisson regression lines. Dashed lines are the ecological Poisson regression lines.

level model and corresponding ecological model are given by:

$$\begin{aligned} E[y_{ij}|\boldsymbol{x_i}, \boldsymbol{z_i^c}] &= g(\beta_0 + \beta_w x_{ij}) && (7) \\ E[\overline{y}_i|\overline{x}_i] &= g(\beta_0^e + \beta_w^e \overline{x}_i), && (8) \end{aligned}$$

where the parameters $\boldsymbol{\beta} = (\beta_0, \beta_w)$ are the same across groups. Loosely following equation (3.5) from Sheppard (2003), the ecological estimating equation based on (8), $U_e(\boldsymbol{\beta})$, will be biased for the unbiased individual level estimating equation, $U(\boldsymbol{\beta})$, and the bias in the ecological parameter estimates $(\widehat{\boldsymbol{\beta}}^{\boldsymbol{e}} - \boldsymbol{\beta})$ can be approximated with a first-order Taylor series expansion of the biased estimating equation about the true value of the parameter. For example, in Poisson regression, $g(\cdot) = \exp(\cdot)$ and the aggregate model (Prentice and Sheppard, 1995),

$$E[\overline{y}_i|\boldsymbol{x_i}] = \frac{1}{n_i} \sum_{j=1}^{n_i} \exp(\beta_0 + \beta_w x_{ij}) \tag{9}$$

does not match the ecological model,

$$E[\overline{y}_i|\overline{x}_i] = \exp(\beta_0^e + \beta_w^e \overline{x}_i) \tag{10}$$

Therefore, an unbiased estimating equation, $U(\boldsymbol{\beta})$, can be written as the sum of the biased ecological estimating equation, $U_e(\boldsymbol{\beta})$, and its bias, $b(\boldsymbol{\beta})$,

$$\begin{aligned}
U(\boldsymbol{\beta}) &= \sum_{i=1}^{m} \left[\begin{array}{c} 1 \\ \overline{x}_i \end{array} \right] \left(\overline{y}_i - \frac{1}{n_i} \sum_{j=1}^{n_i} \exp(\beta_0 + \beta_w x_{ij}) \right) \\
&= \sum_{i=1}^{m} \left[\begin{array}{c} 1 \\ \overline{x}_i \end{array} \right] \left(\overline{y}_i - \exp(\beta_0 + \beta_w \overline{x}_i) + \exp(\beta_0 + \beta_w \overline{x}_i) - \frac{1}{n_i} \sum_{j=1}^{n_i} \exp(\beta_0 + \beta_w x_{ij}) \right) \\
&= \sum_{i=1}^{m} \left[\begin{array}{c} 1 \\ \overline{x}_i \end{array} \right] (\overline{y}_i - \exp(\beta_0 + \beta_w \overline{x}_i)) \\
&+ \sum_{i=1}^{m} \left[\begin{array}{c} 1 \\ \overline{x}_i \end{array} \right] \left(\exp(\beta_0 + \beta_w \overline{x}_i) - \frac{1}{n_i} \sum_{j=1}^{n_i} \exp(\beta_0 + \beta_w x_{ij}) \right) \\
&= U_e(\boldsymbol{\beta}) + b(\boldsymbol{\beta}),
\end{aligned}$$

and the ecological bias in the parameter estimate can be approximated by the first-order Taylor series expansion,

$$\widehat{\boldsymbol{\beta}}^{\boldsymbol{e}} - \boldsymbol{\beta} \approx E[U_e'(\boldsymbol{\beta})]^{-1} b(\boldsymbol{\beta})$$

where,

$$\begin{aligned}
E[U_e'(\beta)] &= - \left[\begin{array}{cc} \sum_{i=1}^{m} \exp(\beta_0 + \beta_w \overline{x}_i) & \sum_{i=1}^{m} \overline{x}_i \exp(\beta_0 + \beta_w \overline{x}_i) \\ \sum_{i=1}^{m} \overline{x}_i \exp(\beta_0 + \beta_w \overline{x}_i) & \sum_{i=1}^{m} \overline{x}_i^2 \exp(\beta_0 + \beta_w \overline{x}_i) \end{array} \right] \\
E[U_e'(\beta)]^{-1} &= - \frac{1}{D} \left[\begin{array}{cc} \sum_{i=1}^{m} \overline{x}_i^2 \exp(\beta_0 + \beta_w \overline{x}_i) & - \sum_{i=1}^{m} \overline{x}_i \exp(\beta_0 + \beta_w \overline{x}_i) \\ - \sum_{i=1}^{m} \overline{x}_i \exp(\beta_0 + \beta_w \overline{x}_i) & \sum_{i=1}^{m} \exp(\beta_0 + \beta_w \overline{x}_i) \end{array} \right]
\end{aligned}$$

and where,

$$D = \sum_{i=1}^{m} \exp(\beta_0 + \beta_w \overline{x}_i) \times \sum_{i=1}^{m} \overline{x}_i^2 \exp(\beta_0 + \beta_w \overline{x}_i) - \left(\sum_{i=1}^{m} \overline{x}_i \exp(\beta_0 + \beta_w \overline{x}_i) \right)^2$$

Since we are primarily interested in β_w, we can write the approximate bias for the ecological estimator as,

$$\begin{aligned}
\beta_w - \widehat{\beta}_w^e &\approx \frac{1}{D} \left[\sum_{i=1}^{m} \overline{x}_i \exp(\beta_0 + \beta_w \overline{x}_i) \right] \times \sum_{i=1}^{m} \left(\exp(\beta_0 + \beta_w \overline{x}_i) - \frac{1}{n_i} \sum_{j=1}^{n_i} \exp(\beta_0 + \beta_w x_{ij}) \right) \\
&- \frac{1}{D} \sum_{i=1}^{m} \exp(\beta_0 + \beta_w \overline{x}_i) \times \left[\sum_{i=1}^{m} \overline{x}_i \left(\exp(\beta_0 + \beta_w \overline{x}_i) - \frac{1}{n_i} \sum_{j=1}^{n_i} \exp(\beta_0 + \beta_w x_{ij}) \right) \right]
\end{aligned} \tag{11}$$

Hence, ecological bias in the simplified model depends on the distribution of the covariate, as will be true in general. If $x_{ij} = x_{ij'}$ for all $j \neq j'$, then $\exp(\beta_0 + \beta_w \overline{x}_i) = \frac{1}{n_i} \sum_{j=1}^{n_i} \exp(\beta_0 + \beta_w x_{ij})$ for all i, and (11) is zero. We can further simplify (11) with the approximation $\frac{1}{n_i} \sum_{j=1}^{n_i} \exp(\beta_0 + \beta_w x_{ij}) \approx \exp(\beta_0 + \beta_w \overline{x}_i) + \frac{1}{2} \beta_w s_{x_i}^2$, where $s^2_{x_i}$ is the sample variance of the covariate within group i. The simplified approximate bias can be written as,

$$\begin{aligned}\beta_w - \widehat{\beta}^e_w \quad &\approx \quad \frac{1}{D}\sum_{i=1}^{m} \overline{x}_i \mathrm{exp}(\beta_0 + \beta_w \overline{x}_i) \times \sum_{i=1}^{m} \mathrm{exp}(\beta_0 + \beta_w \overline{x}_i)\left(1 - \frac{1}{n_i}\sum_{j=1}^{n_i} \mathrm{exp}(1/2\beta_w^2 s^2_{x_i})\right) \\ &- \quad \frac{1}{D}\sum_{i=1}^{m} \mathrm{exp}(\beta_0 + \beta_w \overline{x}_i) \times \sum_{i=1}^{m} \overline{x}_i \mathrm{exp}(\beta_0 + \beta_w \overline{x}_i)\left(1 - \frac{1}{n_i}\sum_{j=1}^{n_i} \mathrm{exp}(1/2\beta_w^2 s^2_{x_i}))\right)\end{aligned} \tag{12}$$

We see that the approximate bias in the slope parameter depends on the variability of the group specific sample variances. In particular, if $s^2_{x_i} = s^2_{x_{i'}}$ for all $i \neq i'$, then the $\left(1 - \frac{1}{n_i}\sum_{j=1}^{n_i} \mathrm{exp}(1/2\beta_w^2 s^2_{x_i})\right)$ multiplier factors out of (12) and the bias for the slope parameter is approximately zero. Furthermore, Richardson et al. (1987) and Wakefield (2003) showed that when x_{ij} is Gaussian, dependence is necessary between s_{x_i} and $\overline{x}_i$ in order for pure specification bias (Greenland, 1992) to be present.

This example with $g(\cdot) = \mathrm{exp}(\cdot)$ illustrates that a non-linear individual level model can result in ecological bias without the effects of intercepts bias, slope bias, or confounding bias that were necessary for ecological bias in the linear model. This result and Figure 1 motivate the use of individual level data for the estimation of individual level parameters. However, while the individual level data provides identification of the parameters, the ecological data can provide precision when combined with the individual data. Furthermore, in many applications, the ecological data will represent an aggregation of all available individual level observations, and therefore any subsample will be dependent on this data. This dependence provides an opportunity for optimal subsample design, because the ecological data are typically available prior to any subsampling. In the following sections, we describe the combination of ecological and subsample data with two goals in mind: estimation of individual regression parameters and optimal subsample design for the estimation of these parameters. There may be some applications where the researcher is only interested in combinations of the individual level parameters, but to simplify the goals of this chapter, we assume that estimates of all parameters are needed.

3 US Voter Turnout Data

Political scientists, policy makers, and campaign strategists want to know who votes (Burden, 2000). However, readily available data sources cannot fully answer this question. As a motivating example, we consider comparing white versus non-white turnout rates in the southern states versus the rest of the US. In the notation of the previous section, a logistic regression is appropriate for this example with i indexing the two regions and j indexing eligible voters within these regions. The dependent variable (y_{ij}) takes the value one

Table 1: The ANES data in counts by region

		Turnout Counts by Region	
Region	Not Vote/Vote	Non-white	White
Non-southern	Not Vote	41	75
	Vote	43	238
Southern	Not Vote	32	27
	Vote	48	90

if eligible voter j in region i voted and zero otherwise, while the independent variable (x_{ij}) takes the value one if eligible voter j in region i is white and zero otherwise. The probability that eligible voter j in region i voted, can be written in logistic regression formulation,

$$E[y_{ij}|x_{ij}] = \frac{\exp(\beta_{0i} + \beta_{wi}x_{ij})}{1 + \exp(\beta_{0i} + \beta_{wi}x_{ij})}, \tag{13}$$

where we allow the slopes to vary by region. Of course, most applications would utilize more complex definitions for the groups, and possibly more covariates, but this simple example will serve to ellucidate the main points of this paper.

Ecological and individual level data sources are available for this model. Ecological data can be obtained from the census and the voter rolls, while the American National Election Study (ANES) provides survey data on individuals. For the reasons stated in the previous section, the ecological data are insufficient to estimate the parameters of (13), The ANES data is also problematic because people who did not vote tend to lie to interviewers when asked about their voting history (Anderson and Silver, 1986; Abelson et al., 1992). However, for most of this paper, we treat the ANES data (Table 1) as the gold standard. Figure 2 shows the results of ecological bias for this example based solely on the ANES data. The regional averages for the southern and non-southern states are plotted as S and NS respectively. Clearly these two points cannot be used to estimate all of the parameters in (13), however, we could use these points to get an indication of the "white effect". The dashed curve represents the predicted probabilities from a logistic regression fitted to these two points, and ecological inference from this regression leads us to conclude that there is no predicted difference in the probability of voting for white and non-white eligible voters. The solid curves represent the predicted probabilities for the southern and non-southern states after fitting (13) with the individual level ANES data. This analysis shows a positive and highly significant "white effect", with insignificant differences between the slopes and the intercepts of the two regions. Therefore, if we assume that the ANES data are the gold standard, ecological inference will lead us to incorrectly conclude that there is no difference in voting probabilities between whites and non-whites.

Of course, the ANES response data are not the gold standard, and we see from Table 1 that the turnout proportions are much higher than the true proportions we would get from the ecological data. At the

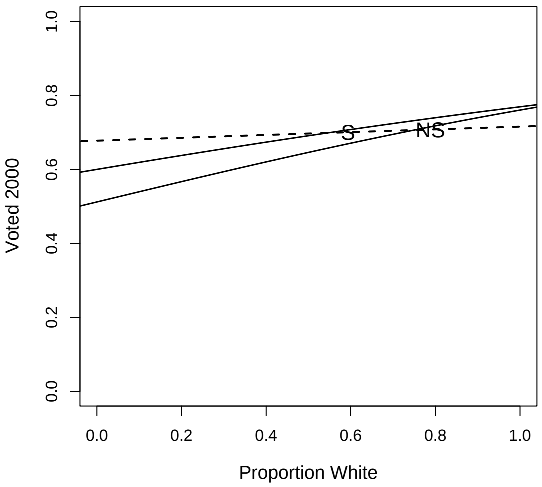


Figure 2: The regional turnout and white/non-white proportions from the ANES for the southern and non-southern states are plotted as S and NS respectively. The dashed curve represents the predicted probabilities of voting based on the ecological logistic regression. The solid curves represent the predicted probabilities of voting based on the within group logistic regressions.

individual level, it is often possible to validate these numbers by cross checking the "not vote/vote" answers of survey respondents against the names on the voting rolls for the corresponding election (Traugott, 1989; Silver et al., 1986). However, this strategy can be quite expensive, because many districts do not have easily accessible records. In the remaining sections of this paper, we illustrate a general methodology that can be utilized to reduce the cost of such studies by using ecological data to design small subsamples of expensive gold standard data such that the combined data will provide precise and accurate estimates.

4 Estimation for Generalized Linear Models with Ecological and Subsample Data

We assume that the individual level model can be written as,

$$f(\boldsymbol{y}, \boldsymbol{X}) = \prod_{i=1}^{m} \prod_{j=1}^{n_i} f_i(y_{ij}|X_{ij}) f_i(X_{ij}), \tag{14}$$

so that individuals are independent of each other, but distributions can be different across groups. In regression, we are only interested in the parameters from the distribution of $\boldsymbol{y}$ given $\boldsymbol{X}$, and the factorization of (14) allows us to extract this conditional distribution without specifying a distribution for the covariates. In this sense, regression is always at least semi-parametric. We assume that the conditional distribution has a generalized linear model form,

$$f(\boldsymbol{y}|\boldsymbol{X}) = \prod_{i=1}^{m}\prod_{j=1}^{n_i} \exp\left\{(y_{ij}\theta_{ij} - b(\theta_{ij}))/a(\phi_i) + c(y_{ij}, \phi_i)\right\} \tag{15}$$

where y_{ij} is the response for individuals $j = 1, ..., n_i$ in groups $i = 1, ..., m$, $\theta_{ij} = X_{ij}\boldsymbol{\beta_i}$ is the linear predictor, $\boldsymbol{\beta_i}$ is the vector of regression parameters for group i with p elements, and X_{ij} is the row of the covariate matrix for individual j in group i. In (14), the (y_{ij}, X_{ij}) pairs are seen as generated by some i.i.d. process.

Suppose that n_i is quite large for all i, and that instead of having observations on all n_i individuals within each group, we only have observations on $k_i << n_i$ individuals for groups $i = 1, ..., m$. Also suppose that we have ecological data; specifically, the ecological response sums $y_{i+} \equiv \sum_{i=1}^{n_i} y_{ij}$ and the vector of ecological covariate sums $X_{i+} \equiv \sum_{i=1}^{n_i} X_{ij}$ for groups $i = 1, ..., m$. In order to estimate the parameters of (15), we need to write down the joint distribution of the subsample and the ecological data. However, to simplify the discussion, we will start with the assumption that we observe the entire covariate matrix within each group i. Since we are only interested in the parameters of (15), this assumption allows us to base estimation and inference on the distribution for subsample and ecological response values $(\boldsymbol{y_i^s}, y_{i+})$ conditional on the full covariate matrix (X_i) within each group. Prentice and Sheppard in Prentice and Sheppard (1995) propose a similar model in which aggregate response data is combined with a sample of covariates. We further simplify the task if we remove the subsampled response values from the ecological response in order to eliminate the dependence. We write the response sum minus the subsampled response sum as $y_{i+}^{(-s)}$, and this sum is independent of the subsampled response vector. Within a single group, we can suppress the i notation, and the distribution of the response subsample, ecological data, and full data on the covariate matrix can be written as,

$$f(\boldsymbol{y^s}, y_+^{(-s)}|\boldsymbol{X}; \boldsymbol{\beta}, \phi) = \prod_{j=1}^{k} \left[\exp\left\{(y_j\theta_j - b(\theta_j))/a(\phi) + c(y_j, \phi)\right\}\right] \times f(y_+^{(-s)}|\boldsymbol{X}; \boldsymbol{\beta}, \phi) \tag{16}$$

where,

$$f(y_+^{(-s)}|\boldsymbol{X}; \boldsymbol{\beta}, \phi) = \int f(y_+, y_{k+2}, ..., y_n|\boldsymbol{X}; \boldsymbol{\beta}, \phi) dy_{k+2} \cdots dy_n$$

Notice that we adopt the notational convention that a semi-colon separates parameters from random variables and that the integral is taken over $y_{k+2}, ..., y_n$. This is because y_{k+1} is determined by y_+ and $y_{k+2}, ..., y_n$.

In (16), $f(y_+^{(-s)}|\boldsymbol{X}; \boldsymbol{\beta}, \phi)$ is a convolution likelihood over the not subsampled response values $(\boldsymbol{y^{(-s)}})$. For some GLMs, this distribution can be written in closed form (e.g. in Poisson regression $(y_+|\boldsymbol{X}; \boldsymbol{\beta}, \phi) \sim Pois\left(\sum_{j=1}^{n} \exp(X_j\boldsymbol{\beta})\right)$. And for any discrete response GLM, this distribution can be written as a sum over all possible values for the individual response data such that the ecological response sum is achieved. Let y_j^* represent a possible value for y_j and $y_+^{(-s)*} = \sum_{j=k+1}^{n} y_j^*$ be the possible ecological response sums. Also let

$\{\boldsymbol{y}^{(-s)*} : y_+^{(-s)*} = y_+\}$ be the set containing the possible values for the not subsampled response values that will sum up to the ecological response sum. Then the combined distribution for a discrete response GLM can be written as,

$$\begin{aligned} f(\boldsymbol{y_i^s}, y_+^{(-s)} | \boldsymbol{X_i}; \boldsymbol{\beta}, \phi) &= \prod_{i=1}^{k} \left[\exp\left\{(y_j\theta_j - b(\theta_j))/a(\phi) + c(y_j, \phi)\right\}\right] \\ &\times \sum_{\{\boldsymbol{y}^{(-s)*} : y_+^{(-s)*} = y_+\}} \prod_{i=1}^{k} \left[\exp\left\{(y_j^*\theta_j - b(\theta_j))/a(\phi) + c(y_j^*, \phi)\right\}\right] \end{aligned} \tag{17}$$

Unfortunately, we will usually not observe the entire covariate matrix. Instead we observe only the vector of ecological covariate sums and the covariate matrix for the subsample. Therefore, we are faced with two choices on how to proceed. First, we can assume that the design matrix is drawn from a distribution whose parameters we must estimate. Second, we can interpret the design matrix as a finite population whose parameters we must estimate. Both approaches can be useful in different circumstances, and both present difficulties.

If we assume that the design matrix is drawn from a distribution, then we must model the distribution of $\boldsymbol{X}$ and treat the parameters of this distribution as nuisance parameters. For example, suppose that the design matrix has one continuous covariate $X_j = \left[\begin{array}{cc} 1 & x_j \end{array}\right]$, we could model this covariate as Gaussian with parameters (μ_x, σ_x^2). Furthermore, in the random draw interpretation, the ecological covariate sums do not provide population level information about (μ_x, σ_x^2). Instead, X_+ is just an aggregation of draws, and the researcher will need to develop the joint distribution of X_+ and $\boldsymbol{X}^s$. Therefore, in this interpretation (μ_x, σ_x^2) are the nuisance parameters that we must estimate along with the parameter of interest $(\boldsymbol{\beta})$.

We now consider a multivariate example that was treated in Richardson et al. (1987), Jackson et al. (2006), and Wakefield and Salway (2001). Suppose the design matrix has two binary variables so that $X_j = \left[\begin{array}{ccc} 1 & x_j & z_j \end{array}\right]$. In this case there are many possible parameterizations of the joint distribution of x and z, with three parameters being required. For example, we could take as the nuisance parameters the odds ratio and the marginal probabilities for one of the two binary variables.

$$x_j | z_j \quad \sim_{ind} \quad Bern\left(\text{expit}(\gamma_0 + \gamma_1 z_j)\right) \tag{18}$$

$$z_j \quad \sim_{iid} \quad Bern(\gamma_2) \tag{19}$$

Again, the ecological covariate sums do not provide population level information about $\boldsymbol{\gamma}$, and the researcher

Table 2: Example with two binary covariates. The internal cells of the table represent the counts of the four unique row types within $\boldsymbol{X}$. These would typically not be included for ecological data. The margins represent the sums over these totals, which would be included in the ecological data.

n_{00}	n_{01}	$n - x_+$
n_{10}	n_{11}	x_+
$n - z_+$	z_+	n

will need to develop the joint distribution of X_+ and $\boldsymbol{X}^s$. This can be done with the convolution likelihood of Wakefield (2004) and the hybrid likelihood of Haneuse and Wakefield (2008).

If we utilize the second interpretation of the design matrix as a finite population, then effectively, we must estimate the rows that were not included in the subsample. The order of these rows will not matter, since $y_+^{(-s)}$ does not reference a specific ordering of the rows in $\boldsymbol{X}^{(-s)}$. However, we will be unable to estimate the composition of the rows in $\boldsymbol{X}^{(-s)}$, unless, $\boldsymbol{X}^s$ contains enough information about the unique rows in $\boldsymbol{X}^{(-s)}$. Specifically, the parameters that need to be estimated are the counts of unique row types in the matrix $\boldsymbol{X} = \left\{\boldsymbol{X}^s \cup \boldsymbol{X}^{(-s)}\right\}$, where $\boldsymbol{X}^s$ is known. These counts can be seen as nuisance parameters that we need to estimate in order to estimate $\boldsymbol{\beta}$ and ϕ.

For example, if the design matrix has two binary variables $X_j = \left[\begin{array}{ccc} 1 & x_j & z_j \end{array}\right]$, then there are only four unique types of row in the design matrix: n_{00} is the number of observations in the full data with $x_j = 0$ and $z_j = 0$, n_{10} is the number of observations in the full data with $x_j = 1$ and $z_j = 0$, n_{01} is the number of observations in the full data with $x_j = 0$ and $z_j = 1$, and n_{11} is the number of observations in the full data with $x_j = 1$ and $z_j = 1$. We can estimate these counts using the corresponding subsample counts: k_{00}, k_{10}, k_{01}, and k_{11}.

As we can see from this example, in order for this method to work, there has to be a limit on the number of unique rows. Specifically, if one of the covariates is continuous, then all the rows of $\boldsymbol{X}$ are potentially unique, and $\boldsymbol{X}^s$ does not provide enough information about $\boldsymbol{X}^{(-s)}$. However, we receive extra information about $\boldsymbol{X}$ from the ecological covariate sums. In the finite population interpretation, these sums (X_+) provide population level information about $\boldsymbol{X}$, and if we restrict ourselves to discrete covariates (or discretize any continuous covariates), then the ecological covariate sums provide population level information about the counts of the unique row types. In our example with two binary covariates, the four unique types of rows in the design matrix can be represented by the 2×2 table shown in Table 2 in which the ecological covariate sums $X_+ = \left[\begin{array}{ccc} n & x_+ & z_+ \end{array}\right]$ specify the margins. Since we do not observe the internal cell counts $(n_{00}, n_{01}, n_{10}, n_{11})$, we only have partial information about the design matrix as specified by the margins of the table. However, $n_{00} + n_{01} + n_{10} + n_{11} = n$, $n_{10} + n_{11} = x_+$, and $n_{01} + n_{11} = z_+$, so the dimensionality of the distribution of the design matrix has been reduced further, and we can write the unknown internal cell counts in terms of a single cell. For example, writing the unknowns in terms of n_{11} gives,

$$\begin{aligned}
n_{01} &= z_+ - n_{11} \\
n_{10} &= x_+ - n_{11} \\
n_{00} &= n_i - x_+ - (z_+ - n_{11}).
\end{aligned}$$

and the nuisance parameter has been reduced to n_{11} which has support over the integers between $\max(0, x_+ + z_+ - n)$ and $\min(x_+, z_+)$. Therefore, in this example, the ecological sums have reduced the dimensionality of the parameter space from four unknown counts to one unknown count, and the subsample design matrix ($\boldsymbol{X_i}^s$) is only needed to estimate this final count. Notice the difference in interpretation between y_+ and X_+; y_j is seen as generated by an independent process, and therefore y_+ has a distribution, whereas we only care about the finite population distribution of $\boldsymbol{X}$ and therefore X_+ is population level information for this finite population. In the remainder of this paper, we utilize both the random draw interpretation and the finite population interpretation depending on the task at hand. In general, we utilize the random draw interpretation for estimation and the finite population interpretation for design.

If we denote the vector of unknown counts for unique row types in $\boldsymbol{X}$ as $\boldsymbol{n_X}$, then we can incorporate the ecological covariate information into the likelihood with an indicator function,

$$f(X_+; \boldsymbol{n_X}) = \begin{cases} 1 & \text{if } \boldsymbol{n_X} \Rightarrow X_+ \\ 0 & \text{otherwise} \end{cases} \tag{20}$$

Utilizing (16) and the finite population interpretation of $\boldsymbol{X}$, inference can be achieved with the joint distribution of the subsample and ecological data, treating the unknown counts for unique row types as parameters. The joint distribution is given by

$$\begin{aligned}
f(\boldsymbol{y^s}, y_+^{(-s)}, \boldsymbol{X^s}, X_+; \boldsymbol{n_X}, \boldsymbol{\beta}, \phi) &= \prod_{j=1}^{k} [\exp\{(y_j\theta_j - b(\theta_j))/a(\phi) + c(y_j, \phi)\}] \\
&\times \; f(y_+^{(-s)} | X_+, \boldsymbol{n_{X^s}}; \boldsymbol{n_X}, \boldsymbol{\beta}, \phi) \\
&\times \; f(\boldsymbol{n_{X^s}} | X_+; \boldsymbol{n_X}) \times f(X_+; \boldsymbol{n_X})
\end{aligned} \tag{21}$$

where, $\theta_j = X_j\boldsymbol{\beta}$. Utilizing (16) and the random draw interpretation of $\boldsymbol{X}$, inference can be achieved by integrating over a complete data likelihood that has been factored into the known and the unknown parts. If we write $\boldsymbol{n_X^{(-s)}}$ as the unknown counts for unique row types in $\boldsymbol{X^{(-s)}}$, then the complete data likelihood is given by

$$\begin{aligned}
f(\boldsymbol{y^s}, \boldsymbol{X^s}, \boldsymbol{n_X^{(-s)}}, y_+, X_+; \boldsymbol{\beta}, \boldsymbol{\gamma}) &= f(\boldsymbol{y^s}, \boldsymbol{X^s}, \boldsymbol{n_X^{(-s)}}, y_+^{(-s)}, X_+^{(-s)}; \boldsymbol{\beta}, \boldsymbol{\gamma}) \\
&= f(\boldsymbol{y^s} | \boldsymbol{X^s}; \boldsymbol{\beta}) \\
&\times \; f(y_+^{(-s)} | \boldsymbol{n_X^{(-s)}}; \boldsymbol{\beta}) \\
&\times \; f(\boldsymbol{n_{X^s}}; \boldsymbol{\gamma}) \\
&\times \; f(\boldsymbol{n_X^{(-s)}} | X_+^{(-s)}; \boldsymbol{\gamma}) \\
&\times \; f(X_+^{(-s)}; \boldsymbol{\gamma}) \qquad (22)
\end{aligned}$$

where $\boldsymbol{n_X^{(-s)}}$ must be averaged over. In principle, (21) or (22) can be used as a basis for estimation and inference. For estimation in the finite population interpretation, we need only maximize over the values of $(\boldsymbol{n_X, \beta}, \phi)$ in order to obtain the MLE. For inference, we can use a profile likelihood where the nuisance parameters $\boldsymbol{n_X}$ have been profiled out. Since the likelihood for any given $\boldsymbol{n_X}$ is well behaved, the profile likelihood will just be the maximum of these individual likelihoods over the range of $(\boldsymbol{\beta}, \phi)$. In the random draw interpretation, we utilize the EM algorithm, taking expectations over the unknown $\boldsymbol{n_X^{(-s)}}$ values. For inference in the random draw interpretation, we use the standard EM approaches to inference, but note that an MCMC approach is straightforward with the $\boldsymbol{n_X^{(-s)}}$ values being auxiliary variables.

4.1 Example: Poisson Regression with Two Binary Covariates

For a Poisson regression with two binary covariates and an additive model (we continue to supress the i notation),

$$\begin{aligned}
\lambda_+^{(-s)} &= \sum_{j=k+1}^{n} \exp(X_j \boldsymbol{\beta}) \\
&= (n_{00} - k_{00}) e^{\beta_0} + (n_{01} - k_{01}) e^{\beta_0 + \beta_c} \\
&+ \; (n_{10} - k_{10}) e^{\beta_0 + \beta_w} + (n_{11} - k_{11}) e^{\beta_0 + \beta_c + \beta_w},
\end{aligned}$$

and $\phi = 1$. Therefore, in the finite population interpretation, (21) simplifies to,

$$\begin{aligned}
f(\boldsymbol{y^s} | \boldsymbol{X^s}; \boldsymbol{\beta}, \phi) &= \prod_{j=1}^{k} Pois(\exp(X_j \boldsymbol{\beta})) \\
f(\boldsymbol{y_+^{(-s)}} | \boldsymbol{n_{X^s}}; \boldsymbol{n_{X^{(-s)}}, \beta}, \phi) &= Pois\left(\lambda_+^{(-s)}\right) \\
f(\boldsymbol{n_{X^s}} | X_+; \boldsymbol{n_X}) &= \frac{\binom{n_{11}}{k_{11}} \binom{n - n_{11}}{k - k_{11}}}{\binom{n}{k}} \qquad (23)
\end{aligned}$$

where n_{11} is an unknown parameter, and $n_{01} = z_+ - n_{11}$, $n_{10} = x_+ - n_{11}$, and $n_{00} = n - x_+ - (z_+ - n_{11})$.

In the random draw interpretation for this example, $\boldsymbol{n}_{\boldsymbol{X}}^{\boldsymbol{(-s)}}|X_+^{(-s)}$ has the non-central hypergeometric distribution of McCullah and Nelder (1989). The joint distribution of $X_+^{(-s)}$ in this example can be parameterized by factoring $f(x_+^{(-s)}, z_+^{(-s)}; \boldsymbol{\gamma})$ into $f(x_+^{(-s)}|z_+^{(-s)}; \boldsymbol{\gamma}) \times f(z_+^{(-s)}; \boldsymbol{\gamma})$, where $f(x_+^{(-s)}|z_+^{(-s)}; \boldsymbol{\gamma})$ has the binomial convolution distribution of Wakefield (2004) and $f(z_+^{(-s)}; \boldsymbol{\gamma})$ follows a binomial distribution. The complete data likelihood simplifies to

$$\begin{aligned}
f(\boldsymbol{y^s}, \boldsymbol{X^s}, \boldsymbol{n}_{\boldsymbol{X}}^{\boldsymbol{(-s)}}, y_+, X_+; \boldsymbol{\beta}, \boldsymbol{\gamma}) &= \prod_{j=1}^{k} \{Pois\,(y_j|\lambda_j)\} \\
&\times\ Pois\left(y_+^{(-s)}|\lambda_+^{(-s)}\right) \\
&\times\ Multinom(k_{00}, k_{10}, \gamma_{01}, k_{11}, k_{00}; \gamma_{10}, \gamma_{01}, \gamma_{11}) \\
&\times\ NCHyperGeo\left(n_{11}^{(-s)}|n-k, X_+^{(-s)}, \frac{\gamma_{11}\gamma_{00}}{\gamma_{10}\gamma_{01}}\right) \\
&\times\ BinConvolution(x_+^{(-s)}|z_+^{(-s)}, n-k; \boldsymbol{\gamma}) \\
&\times\ Bin(z_+^{(-s)}|n-k; (\gamma_{01} + \gamma_{11}, \gamma_{00} + \gamma_{10})
\end{aligned}$$

where the distribution of the not subsampled row counts follows the non-central hypergeometric distribution and the distribution of the ecological counts follows a binomial convolution.

5 Optimal Subsample Design for Generalized Linear Models Conditional on Ecological Data

For optimal design conditional on the ecological data, we must carefully consider the information available to the researcher at the time of the subsample. We know that the ecological data are available, but most design schemes will also require some information on individual covariate values. For example, if white/non-white is a covariate, and we would like to sample based on this covariate, we need to be able to choose to sample whites or non-whites. There are a number of scenarios where this may be feasible. First, it may be possible to obtain a sampling frame based on full or partial data for all the covariates. With a full sampling frame, design is based on the distribution of the subsampled response values conditional on the ecological response and individual covariate values because the sampling frame provides the full design matrix. In the second, more realistic scenario without a full sampling frame, it may be possible to perform a two stage sampling design where individuals are sampled at random, and we choose whether or not to sample response values from these individuals based on their covariates. For example, the ANES data provides a probability sample of eligible voters. If we want to obtain the true "not vote/vote" status of these eligible voters, one could subsample among the ANES respondents and crosscheck against the voting rolls for the names of selected respondents. A two stage sample will slightly complicate the analysis, but any real application will likely be further complicated (e.g. with demographic covariates, the inclusion of data from the Public Use Microdata Survey (PUMS), Ruggles et al. (2004) will often be desirable). That being said, results for optimal subsample design based on a sampling frame for the full design matrix and the results from the previous section provide the building blocks for a more complicated analysis. In this section, we present results for design in the sampling frame context.

If a sampling frame is available for the covariate data, then design questions will be based on the distribution of subsample response data conditional on the ecological response data and the individual covariate data. We assume that we can sample within each group, therefore independence between groups is preserved, and we can consider the distribution for a single group, again supressing the i notation.

$$\begin{aligned} f(\boldsymbol{y^s}|y_+,\boldsymbol{X}) &= \frac{f(\boldsymbol{y^s},y_+|\boldsymbol{X})}{f(y_+|\boldsymbol{X})} \\ &= \frac{f(\boldsymbol{y^s},y_+^{(-s)}|\boldsymbol{X})}{f(y_+|\boldsymbol{X})} \\ &= \frac{f(\boldsymbol{y^s}|\boldsymbol{X}) \times f(y_+^{(-s)}|\boldsymbol{X})}{f(y_+|\boldsymbol{X})} \\ &= \frac{f(\boldsymbol{y^s}|\boldsymbol{X^s}) \times f(y_+^{(-s)}|\boldsymbol{X^{(-s)}})}{f(y_+|\boldsymbol{X})} \end{aligned} \tag{24}$$

Notice that in this design scenario, the design matrix is known, and we don't need to estimate the counts of unique row types. Also notice that redundancy in the subsampled responses and ecological responses allows us to re-write the joint distribution in terms of the subsampled responses and the sum of the "not subsampled" responses, which in turn allows the numerator of (24) to factor in the same manner as (21). The slight complication here is that we need to divide by the distribution of the ecological sum.

5.1 Example: The Linear Model

Subsample design is relatively easy in the linear model because sums of normals in this framework only depend on sums of the covariates, $f(y_+|X_+;\boldsymbol{X},\boldsymbol{\beta},\phi) = f(y_+|X_+;\boldsymbol{\beta},\phi)$, and therefore,

$$\begin{aligned} y_+|X_+ &\sim N(X_+\boldsymbol{\beta}, n\sigma^2) \\ y_+^{(-s)}|X_+^{(-s)} &\sim N(X_+^{(-s)}\boldsymbol{\beta}, (n-k)\sigma^2) \\ \boldsymbol{y^s}|\boldsymbol{X^s} &\sim_{ind} N_k\left(\boldsymbol{X^s}\boldsymbol{\beta}, \sigma^2 I_k\right) \end{aligned}$$

If we have a sampling frame for the entire covariate matrix, (24) simplifies to

$$\begin{aligned} f(\boldsymbol{y^s}|y_+,\boldsymbol{X};\boldsymbol{\beta},\phi) &= \frac{f(\boldsymbol{y^s}|\boldsymbol{X^s};\boldsymbol{\beta},\sigma^2)\times f(y_+^{(-s)}|\boldsymbol{X^{(-s)}};\boldsymbol{\beta},\sigma^2)}{f(y_+|\boldsymbol{X};\boldsymbol{\beta},\sigma^2)} \\ &= \frac{f(\boldsymbol{y^s}|\boldsymbol{X^s};\boldsymbol{\beta},\sigma^2)\times f(y_+^{(-s)}|X_+^{(-s)};\boldsymbol{\beta},\sigma^2)}{f(y_+|X_+;\boldsymbol{\beta},\sigma^2)}, \end{aligned}$$

where the second step occurs because of the linear model. If we denote covariate vectors of averages as $\overline{X} \equiv \frac{1}{n}X_+$ and the centered covariate subsample matrix as $\boldsymbol{X^s} - \overline{\boldsymbol{X}}$, then the Fisher information about $\boldsymbol{\beta}$ can be written as

$$\begin{aligned} I(\boldsymbol{\beta}) &= \frac{1}{\sigma^2}\left(\boldsymbol{X^s}^T\boldsymbol{X^s} + \frac{1}{n-k}(X_+^{(-s)})^T X_+^{(-s)} - \frac{1}{n}(X_+)^T X_+\right) \\ &= \frac{1}{\sigma^2}\left(\boldsymbol{X^s}^T\boldsymbol{X^s} + \frac{1}{n-k}(X_+ - X_+^s)^T(X_+ - X_+^s) - \frac{1}{n}(X_+)^T X_+\right) \\ &= \frac{1}{\sigma^2}\left((\boldsymbol{X^s} - \overline{\boldsymbol{X}})^T(\boldsymbol{X^s} - \overline{\boldsymbol{X}}) + \frac{1}{n-k}\left[\frac{k}{n}(X_+)^T X_+ - 2(X_+^s)^T X_+ + \frac{n}{k}(X_+^s)^T X_+^s\right]\right) \\ &= \frac{1}{\sigma^2}\left((\boldsymbol{X^s} - \overline{\boldsymbol{X}})^T(\boldsymbol{X^s} - \overline{\boldsymbol{X}}) + \frac{nk}{n-k}(\overline{X}^s - \overline{X})^T(\overline{X}^s - \overline{X})\right) \qquad (25) \end{aligned}$$

where (25) is the matrix version of the design result from Glynn et al. (2008). In order to maximize information about $\boldsymbol{\beta}$, we want to subsample widely varying covariate values whose averages are far away from the ecological averages.

5.2 Example: Poisson Regression

In our Poisson regression example,

$$\begin{aligned} \boldsymbol{y^s}|\boldsymbol{X^s} &\sim \prod_{j=1}^{k} Pois(\exp(X_j\boldsymbol{\beta})) \\ y_+^{(-s)}|\boldsymbol{X^{(-s)}} &\sim Pois\left(\sum_{j=k+1}^{n}\exp(X_j\boldsymbol{\beta})\right) \\ y_+|\boldsymbol{X} &\sim Pois\left(\sum_{j=1}^{n}\exp(X_j\boldsymbol{\beta})\right) \qquad (26) \end{aligned}$$

If the conditional likelihood (24) can be derived, then optimal sampling design can be based on the expected information from the conditional likelihood. If we also have a sampling frame for the entire design matrix, (24) simplifies to

$$
\begin{aligned}
f(\boldsymbol{y^s}|y_+, \boldsymbol{X}) &= \frac{f(\boldsymbol{y^s}|\boldsymbol{X^s}) \times f(y_+^{(-s)}|\boldsymbol{X^{(-s)}})}{f(y_+|\boldsymbol{X})} \\
&= \frac{f(\boldsymbol{y^s}|\boldsymbol{X^s}) \times f(y_+^{(-s)}|X_+^{(-s)})}{f(y_+|X_+)}
\end{aligned}
$$

If we subsample within each group, the observed information in the group i can be written as

$$
\begin{aligned}
\frac{\partial^2 \ell}{\partial \boldsymbol{\beta}} &= -\sum_{j=1}^{k} X_j^T X_j e^{X_j\boldsymbol{\beta}} \\
&+ \left(\frac{y_+^{(-s)}}{\lambda_+^{(-s)}} - 1\right) \sum_{j=k+1}^{n} X_j^T X_j e^{X_j\boldsymbol{\beta}} - \frac{y_+^{(-s)}}{(\lambda_+^{(-s)})^2} \sum_{j=k+1}^{n} X_j^T e^{X_j\boldsymbol{\beta}} \sum_{j=k+1}^{n} X_j e^{X_j\boldsymbol{\beta}} \\
&- \left(\frac{y_+}{\lambda_+} - 1\right) \sum_{j=1}^{n} X_j^T X_j e^{X_j\boldsymbol{\beta}} - \frac{y_+}{\lambda_+^2} \sum_{j=1}^{n} X_j^T e^{X_j\boldsymbol{\beta}} \sum_{j=1}^{n} X_j e^{X_j\boldsymbol{\beta}} \\
&= \frac{y_+^{(-s)}}{\lambda_+^{(-s)}} \left(\sum_{j=k+1}^{n} X_j^T X_j e^{X_j\boldsymbol{\beta}} - \frac{1}{(\lambda_+^{(-s)})} \sum_{j=k+1}^{n} X_j^T e^{X_j\boldsymbol{\beta}} \sum_{j=k+1}^{n} X_j e^{X_j\boldsymbol{\beta}} \right) \\
&- \frac{y_+}{\lambda_+} \left(\sum_{j=1}^{n} X_j^T X_j e^{X_j\boldsymbol{\beta}} - \frac{1}{\lambda_+} \sum_{j=1}^{n} X_j^T e^{X_j\boldsymbol{\beta}} \sum_{j=1}^{n} X_j e^{X_j\boldsymbol{\beta}} \right) \qquad (27)
\end{aligned}
$$

Notice that we can factor e^{β_0} out of the top and bottom of all of these terms, and hence (27) no longer depends on β_0. This is not surprising, because the ecological sums contain all the information about the overall average. Furthermore, the two terms of (27) can be written as weighted sums of squares if we write $\overline{X}^w = \frac{1}{\lambda_+} \sum_{j=1}^n X_j e^{X_j\boldsymbol{\beta}}$ and $\boldsymbol{X_j^w} = \left(\lambda_j^{1/2}(X_j - \overline{X}^w)\right)_j$ for rows $j = 1, ..., n$ and $\overline{X}^{w(-s)} = \frac{1}{\lambda_+} \sum_{j=1}^k X_j e^{X_j\boldsymbol{\beta}}$ and $\boldsymbol{X_j^{w(-s)}} = \left(\lambda_j^{1/2}(X_j - \overline{X}^{w(-s)})\right)_j$ for rows $j = 1, ..., k$.

$$
\frac{\partial^2 \ell}{\partial \boldsymbol{\beta}} = \frac{y_+^{(-s)}}{\lambda_+^{(-s)}} \left(\boldsymbol{X^{w(-s)}}^T \boldsymbol{X^{w(-s)}} \right) - \frac{y_+}{\lambda_+} \left(\boldsymbol{X^w}^T \boldsymbol{X^w} \right) \qquad (28)
$$

Therefore, the observed information can be seen as the combination of the weighted sums of squares for the "not subsampled" covariates and the weighted sums of squares for the full data. From the design standpoint, we will not know $y_+^{(-s)}$ prior to subsampling, so we need to derive the expected information conditional on the ecological data. Conveniently, $y_+^{(-s)}|y_+$ follows a binomial distribution and hence $E[y_+^{(-s)}|y_+] = \frac{y_+ \lambda_+^{(-s)}}{\lambda_+}$. Therefore, the Fisher information can be written as,

$$
I(\boldsymbol{\beta}) = \frac{y_+}{\lambda_+} \left\{ \left(\boldsymbol{X^w}^T \boldsymbol{X^w} \right) - \left(\boldsymbol{X^{w(-s)}}^T \boldsymbol{X^{w(-s)}} \right) \right\} \qquad (29)
$$

The expected information in the subsample conditional on the ecological data provides an objective function to answer design questions. In general, we want to maximize this information, and we achieve this by maximizing the first term and minimizing the second term. The second term can be minimized by subsampling covariate values in order to minimize the weighted sum of squares for the “not subsampled” individuals within group i. The first term cannot be maximized within group i, because all of the elements are fixed by the ecological data. However, in some cases it will be appropriate to subsample only from a subset of groups, and this term can tell us about the potential information within each group. With more than one covariate, $I(\boldsymbol{\beta})$ is a matrix, and hence we must define what is meant by maximization. Maximizing the determinant is one possibility, but we will assume that there is one parameter of interest, and the coefficients on all other covariates can be thought of as nuisance parameters.

One plausible example with complete information about the design matrix involves a Poisson regression with a single binary covariate. When we have a single binary covariate, the ecological covariate data will fully describe the distribution of the covariates. Therefore, $\sum_{j=1}^{n} \lambda_j$ will be known despite the non-linear link function, and the joint distribution of the ecological and subsample data has an analytical form. This form will allow not only estimation, but also derivation of the Fisher Information that will inform optimal subsample design conditional on the ecological data. However, design will depend on which parameters are group specific. Therefore, we divide the discussion into two cases: group specific risk ratio and common risk ratio.

If we utilize a group specific risk ratio (β_{wi}), then we can again supress the i notation by considering a single group, and $\sum_{j=1}^{n} \lambda_j = n\overline{x}e^{\beta_0+\beta_w} + n(1-\overline{x})e^{\beta_0}$, where $\overline{x}$ is the ecological covariate average within each group. Furthermore, $\sum_{j=k+1}^{n} \lambda_j = n\overline{x}^{(-s)}e^{\beta_0+\beta_w} + n(1-\overline{x}^{(-s)})e^{\beta_0}$, where $\overline{x}^{(-s)}$ is the covariate average for the individuals who aren’t in the subsample. We can calculate this value from the ecological and subsample covariate values, therefore, every element of the ecological distribution is known.

The expected information conditional on the ecological data (29) simplifies for binary covariates. Recall that e^{β_0} cancels from the information equation, and therefore, β_w is the only parameter of interest for group i, and its expected information is given by

$$I(\beta_w) = -\frac{(n-k)}{\lambda_+/n}\frac{\overline{x}^{(-s)}(1-\overline{x}^{(-s)})e^{\beta_w}}{(\overline{x}^{(-s)}e^{\beta_w} + (1-\overline{x}^{(-s)}))} + \frac{y_+}{\lambda_+/n}\frac{\overline{x}(1-\overline{x})e^{\beta_w}}{(\overline{x}e^{\beta_w} + (1-\overline{x}))} \tag{30}$$

This equation allows us to address a number of optimal design questions, but as in most non-linear design scenarios, we must specify β_w in order to proceed. In general, we can specify an optimal design for each possible value of β_w, and the researcher can choose the appropriate design for a given application. However, in this example, the optimal design only changes at one point in the parameter space.

In Figure 3 we show the plots of expected information within a single group (i.e. i notation supressed) as a function of the “not subsampled” covariate average for a group with $n = 1000$, $k = 50$, and an ecological covariate average of 0.3. The vertical lines represent the values we can choose for $\overline{x}^{(-s)}$. This range depends

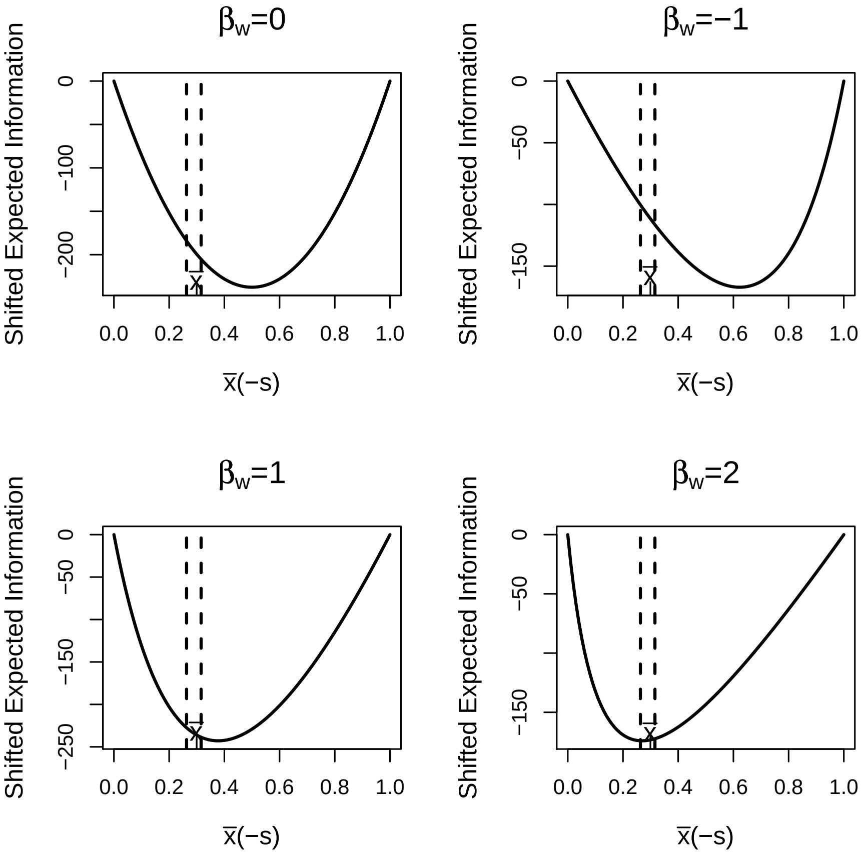


Figure 3: Vertically shifted expected information for different values of β_w as a function of the "not subsampled" covariate average. The vertical dashed bars represent the possible values for the "not subsampled" covariate average, which depends on the ecological covariate average (.3 in this example), the full data size (n=1000), and the subsample size (k=50): When $\beta_w = 0$, the optimal design is to choose $\overline{x}^{(-s)}$ to be as close to zero as possible. Since subsampling large covariates will make $\overline{x}^{(-s)}$ smaller, we should subsample covariates equal to one. When $\beta_w = -1$, the optimal design is to choose $\overline{x}^{(-s)}$ to be as close to zero as possible. Therefore, we should subsample covariates equal to one. When $\beta_w = 1$, the optimal design is to choose $\overline{x}^{(-s)}$ to be as close to zero as possible. Therefore, we should subsample covariates equal to one. When $\beta_w = 2$, the optimal design is to choose $\overline{x}^{(-s)}$ to be as close to one as possible. Therefore, we should subsample covariates equal to zero.

on the ecological covariate average, the full data size and the subsample size. In Figure 3 when $\beta_w = 0$, the expected information curve is symmetric as a function of the "not subsampled" covariate average. Within the vertical dashed lines, we maximize the expected information by choosing $\overline{x}^{(-s)}$ to be as small as possible. Since large subsampled covariates will make $\overline{x}^{(-s)}$ small, we should sample covariates equal to one. This result shows that rare covariates will provide the most information, when the covariate has a weak effect on the response count. In Figure 3 when $\beta_w = -1$, the expected information curve is left skewed as a function of the "not subsampled" covariate average. Within the vertical lines, we still maximize the expected information by choosing $\overline{x}^{(-s)}$ to be as small as possible, therefore, we should sample covariates equal to one. In fact, the skewness of the expected information function increases the benefit that we get from using the optimal design. Hence when there is a negative effect and covariate "ones" are rare, then low count responses are also rare. Therefore, we especially want to sample the rare covariate "ones" because these will bring us the rare lower count responses on average. In Figure 3 when $\beta_w = 1$, the expected information curve is right skewed as a function of the "not subsampled" covariate average. However, the skewness of this function is not enough to overcome the effect of the ecological covariate average. Within the vertical lines, we still maximize the expected information by choosing $\overline{x}^{(-s)}$ to be as small as possible, therefore, we should sample covariates equal to one. Notice, the right skewness of the expected information function decreases the benefit that we get from using the optimal design. Hence when there is a small positive effect and covariate "ones" are rare, the rare covariate bonus, outweighs the expected rare response bonus. Therefore, we should sample the rare covariate values, even though these will lead to non-rare responses on average. In Figure 3 when $\beta_w = 2$, the expected information curve is right skewed to a greater extent as a function of the "not subsampled" covariate average. Within the vertical lines, we now maximize the expected information by choosing $\overline{x}^{(-s)}$ to be as large as possible, therefore, we should sample covariates equal to zero. Hence the larger right skewness of the expected information changes the optimal design, and the bonus we get from rare responses outweighs the bonus we get from the rare covariates. Therefore, we should sample non-rare covariates in order to get rare responses on average. Figure 3 shows that for a specific data set we should specify a decision rule based on $\overline{x}$ and our beliefs about β_w. Furthermore, for the case of a single binary covariate, there are only two possible optimal designs, and the β_w values that specify one design or the other partition the parameter space into two parts.

If we additively include a binary confounder to the model and treat the coefficient on this variable (β_c) as a nuisance parameter, then the design problem is to maximize the information about β_w while accounting for our uncertainty about β_c. Figure 4 shows a contour plot for maximized expected information for β_w treating β_c as a nuisance parameter, with $k = 100$, $n = 1000$, $n_{11} = 200$, $\overline{x} = .3$, $\overline{z} = .6$ and β_w and β_c close to zero. The heights on the contour plot represent the maximum expected information $I(\beta_w)$ for different values of the subsampled binary covariate average ($\overline{x}^s$) and the subsampled binary confounder average ($\overline{z}^s$) where the maximization is carried out over the possible values of k_{11}. However, the maximum (represented

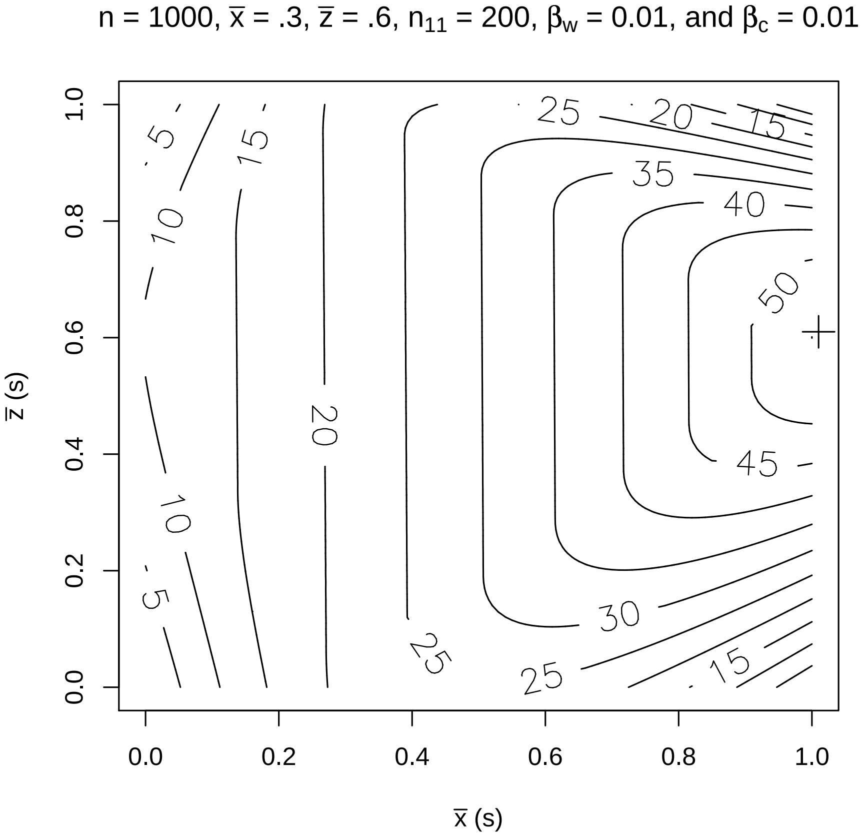


Figure 4: Maximized expected information for β_w treating β_c as a nuisance parameter, with $k = 100$, $n = 1000$, $n_{11} = 200$, $\overline{x} = .3$, $\overline{z} = .6$ and β_w and β_c close to zero. The heights on the contour plot represent the maximum expected information $I(\beta_w)$ for different values of the subsampled binary covariate average ($\overline{x}^s$) and the subsampled binary confounder average ($\overline{z}^s$) where the maximization is carried out over the possible values of k_{11}. However, the maximum (represented by the "+") is on the boundary, and hence k_{11} is determined by $\overline{x}^s$ and $\overline{z}^s$. This shows that to maximize $I(\beta_w)$ when β_w and β_c are close to zero, we should sample rare values of the binary covariate, and sample the binary confounder in proportions equal to the ecological proportions.

by the "+") is on the boundary, and hence k_{11} is determined by $\overline{x}^s$ and $\overline{z}^s$. This maximizing value of $\overline{x}^s$ is one, showing that we replicate the design rule for the one variable case. The maximizing value of $\overline{z}^s$ is 0.6 showing that we should sample the binary confounder in proportions equal to the ecological proportions. This result is analogous to the linear model result from Glynn et al. (2008). We can produce similar plots for different values of β_w and β_c, and as in the one variable case, we have found that this maximizing value of $\overline{x}^s$ is fairly robust against different parameter values. However, the maximizing value of $\overline{z}^s$ is not robust to parameter values away from zero.

If we are willing to assume a common risk ratio across groups, then we can choose which group to sample in addition to the design within each group. We also need to re-introduce the i notation. Re-writing (30) with the i notation and a common risk ratio,

$$I(\beta_w) = \sum_{i=1}^{m} \left\{ -\frac{(n_i - k_i)}{\lambda_{i+}/n_i} \frac{\overline{x}_i^{(-s)}(1-\overline{x}_i^{(-s)})e^{\beta_w}}{(\overline{x}_i^{(-s)}e^{\beta_w} + (1-\overline{x}_i^{(-s)}))} + \frac{y_{i+}}{\lambda_{i+}/n_i} \frac{\overline{x}_i(1-\overline{x}_i)e^{\beta_w}}{(\overline{x}_i e^{\beta_w} + (1-\overline{x}_i))} \right\} \tag{31}$$

we can choose a group to sample. By examination of (31), we see that large y_{i+} will maximize the second term as will large $\overline{x}_i(1-\overline{x}_i)$ when β_w is close to zero. If we were only concerned with this term, we would often choose a group with large y_{i+} and $\overline{x}_i$ close to $\frac{1}{2}$. However, we also need to minimize the first term, and therefore we need a group with a small n_i, and $\overline{x}_i^{(-s)}$ not close to $\frac{1}{2}$. Given the within group design rule of the previous paragraph, we will subsample all zeros or ones so as to move $\overline{x}_i^{(-s)}$ far away from $\frac{1}{2}$, but unfortunately, we can move this variable farther away from $\frac{1}{2}$ when $\overline{x}_i$ isn't close to $\frac{1}{2}$. Hence, choosing the group with the most information depends on a number of factors.

5.3 Example: Logistic Regression

If we use a group specific slopes for logistic regression, we can again suppress the i notation by considering a single group. In this case, subsample design is described by the following distributions.

$$\begin{aligned} f(\boldsymbol{y^s}|\boldsymbol{X^s}) &= \prod_{j=1}^{k} \frac{\exp(X_j\boldsymbol{\beta}y_j)}{1+\exp(X_j\boldsymbol{\beta})} \\ f(y_+^{(-s)}|\boldsymbol{X^{(-s)}}) &= \sum_{\{\boldsymbol{y^{(-s)*}}:y_+^{(-s)*}=y_+^{(-s)}\}} \left(\prod_{j=1}^{k} \frac{\exp(X_j\boldsymbol{\beta}y_j^*)}{1+\exp(X_j\boldsymbol{\beta})} \right) \\ f(y_+|\boldsymbol{X}) &= \sum_{\{\boldsymbol{y^*}:y_+^*=y_+\}} \left(\prod_{j=1}^{n} \frac{\exp(X_j\boldsymbol{\beta}y_j^*)}{1+\exp(X_j\boldsymbol{\beta})} \right) \end{aligned} \tag{32}$$

where recall that y_j^* represents a possible value for y_j, $y_+^* = \sum_{j=1}^n y_j^*$ is the possible ecological response sum, $\{\boldsymbol{y}^* : y_+^* = y_+\}$ is the set containing the $\binom{n}{y_+}$ ways of allocating the y_+ ones into the n spots, and $y_+^{(-s)*}$ and $\{\boldsymbol{y}^{(-s)*} : y_+^{(-s)*} = y_+^{(-s)}\}$ are the "not subsampled" equivalents.

If we write $S_+ = \sum_{\{\boldsymbol{y}^*:y_+^*=y_+\}} e^{\sum_{j=1}^n (X_j\boldsymbol{\beta})y_j^*}$ and $S_+^{(-s)} = \sum_{\{\boldsymbol{y}^{(-s)*}:y_+^{(-s)*}=y_+^{(-s)}\}} e^{\sum_{i=1}^n (X_j\boldsymbol{\beta})y_j^*}$, and $\overline{X}^w = \sum_{\{\boldsymbol{y}^*:y_+^*=y_+\}} \left\{ e^{\sum_{j=1}^n (X_j\boldsymbol{\beta})y_j^*} \sum_{j=1}^n y_j^* X_j \right\}$ and $\overline{X}^{w(-s)} = \sum_{\{\boldsymbol{y}^{(-s)*}:y_+^{(-s)*}=y_+^{(-s)}\}} \left\{ e^{\sum_{j=1}^n (X_j\boldsymbol{\beta})y_j^{(-s)*}} \sum_{j=1}^n y_j^{(-s)*} X_j \right\}$ then the observed information can be written as,

$$\begin{aligned} \frac{\partial^2 \ell}{\partial \boldsymbol{\beta_i^2}} &= \frac{1}{S_+^{(-s)}} \left[\sum_{\{\boldsymbol{y}^{(-s)*}:y_+^{(-s)*}=y_+^{(-s)}\}} \left\{ e^{\sum_{j=k+1}^n (X_j\boldsymbol{\beta})y_j^*} \sum_{j=k+1}^{n} X_j^T y_j^* \sum_{j=k+1}^{n} y_j^* X_j \right\} - \frac{1}{S_{i+}^{(-s)}} \overline{X}^{w(-s)T} \overline{X}^{w(-s)} \right] \\ &- \frac{1}{S_+} \left[\sum_{\{\boldsymbol{y}^*:y_+^*=y_+\}} \left\{ e^{\sum_{j=1}^n (X_j\boldsymbol{\beta})y_j^*} \sum_{j=1}^{n} X_j^T y_j^* \sum_{j=1}^{n} y_j^* X_j \right\} - \frac{1}{S_{i+}} \overline{X}^{wT} \overline{X}^{w} \right] \end{aligned} \tag{33}$$

Therefore, the observed information takes on a weighted sum of squares formulation in the same manner as the Poisson model, Unfortunately, the expected information is not available in closed form, but we can produce plots similar to those in the Poisson example. Furthermore, for the case of a single binary covariate, there are again only two possible optimal designs, and the β_w values that specify one design or the other partition the parameter space into two parts. In the application section, we will calculate exact cut-off points for the US voter turnout example.

6 Application of Estimation and Design Methodology to the US Voter Turnout Data

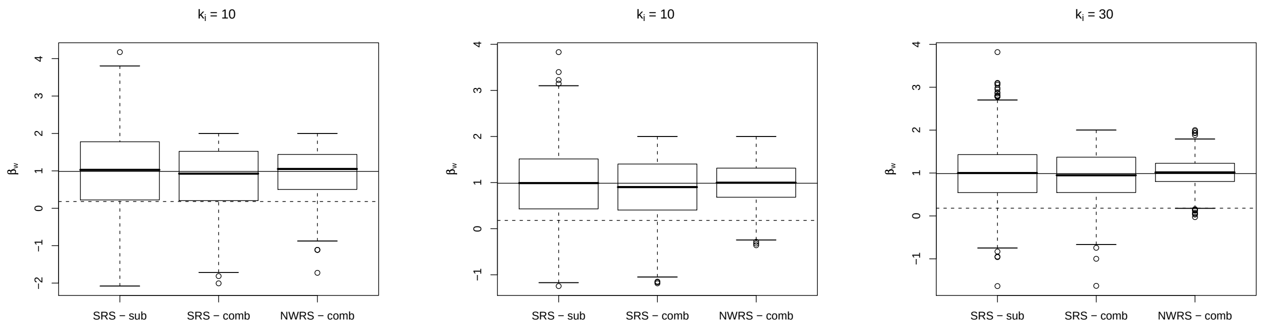
In this section we demonstrate the efficacy of the estimation and design methodology with an illustrative application using the ANES turnout data. We examine a model with a single binary covariate and three different within region sample sizes: $k_i = 10$, $k_i = 20$, and $k_i = 30$. In order to simplify the discussion, we assume that the within group subsample sizes are the same across the southern and the non-southern states, and we assume a common within group slope. We show two main results. First, when the subsample is a simple random sample from the full data, the combined approach improves the precision for estimates of the slope parameters slightly. Second, we show that optimal design in the combined approach will provide substantially increased precision about the slope parameters.

If we believe the true model is a logistic regression with regional intercepts and a common slope for the white/non-white indicator variable, then the true model can be written as the following:

$$E[y_{ij}|x_{ij}] = \frac{\exp(\beta_{0i} + \beta_w x_{ij})}{1 + \exp(\beta_{0i} + \beta_w x_{ij})}, \tag{34}$$

where x_{ij} is a white indicator. We do not know the true values of the parameters, but we can calculate the full data MLEs ($\hat{\beta}_w^{full} = 0.986$, $\hat{\beta}_{0NS}^{full} = 0.136$, and $\hat{\beta}_{0S}^{full} = 0.310$).

In this section, we investigate the benefit to be gained from subsampling design conditional on the ecological data. Additionally, the binary covariate in our example allows a simple solution to the problem of optimal design. In the following, we show that for a logistic regression with regional intercepts and a common slope for x_{ij}, optimal design in the combined approach will provide substantially increased precision about the slope parameter. We showed in Section 5 that we can maximize our information about the common within region slope by carefully subsampling based on covariate values. In the context of our application, the covariate is binary, and the percentage of white individuals is greater than 50% in all regions. We calculate that as long as $\beta_w < 2.12$ for $k_i = 10$, $\beta_w < 2.25$ for $k_i = 20$, and $\beta_w < 2.42$ for $k_i = 30$, the optimal design is to subsample only non-whites. Since β_w is unlikely to be large, we use the design that samples



(a) (b) (c)

Figure 5: Subsampling distributions for $\hat{\beta}_w$ in the single covariate model (34). We employ three estimation approaches: subsample data based on simple random subsamples (SRS-sub), combined data based on simple random subsamples (SRS-comb), combined data based on non-white random subsamples (NWRS-comb). The solid horizontal line represents the full data MLE, and the dashed line represents the ecological estimate: (a) Ten subsampled individuals in each group ($k_i = 10$), (b) Twenty subsampled individuals in each group ($k_i = 20$), (c) Thirty subsampled individuals in each group ($k_i = 30$)

only non-whites.

In order to compare the combined estimator under optimal design to the combined and subsample estimators under simple random sampling, we generated 1000 simple random subsamples (SRS) and 1000 non-white random samples (NWRS, i.e. random subsamples of non-whites). To simplify, we sampled equal numbers from within each group, and the process was repeated for three different within-group sample sizes: $k_i = 10$, 20, 30. We then created three sampling distributions: $\hat{\beta}_w^{sub}$ under SRS (SRS-sub), $\hat{\beta}_w^{comb}$ under SRS (SRS-comb), $\hat{\beta}_w^{comb}$ under NWRS (NWRS-comb). Because of the relatively small sample sizes for the full data set, we conducted these random samples with replacement in order to approximate the situation where the full data set is large. The results from this section become stronger if sampling is performed without replacement.

In Figure 5, we see that the NWRS-comb estimator performs better than the SRS-sub and the SRS-com estimators for all subsample sizes. The solid line represents the true value of the parameter and the dashed line represents the ecological estimate while the boxplots summarize the sampling distributions for the three estimators. For all three subsample sizes, the sampling distribution for NWRS-comb is tighter around the true value of the parameter. Additionally, when $k_i = 10$ and $k_i = 20$ the SRS-sub and the SRS-comb estimators sometimes generate problematic estimates that are not plausible. These have been removed from the plot in order to maintain a reasonable scale. The NWRS-comb estimator avoids this problem altogether.

In Table 3 we report effective sample sizes based on the variance ratios from the subsampling distributions.

Table 3: Effective Sample Sizes for $\widehat{\beta}_w$ based on variance ratios from the subsampling distributions. The problematic estimates for the SRS-sub and SRS-comb estimators have been removed.

Data & Subsample Scheme	Group Size $k_i = 10$	 $k_i = 20$	 $k_i = 30$
Subsample only (SRS-sub)	•	•	•
Eco plus Subsample (SRS-comb)	11	27	40
Eco plus Non-white Subsample (NWRS-comb)	28	65	127

For all subsample sizes, the SRS-comb estimator effectively buys a few more observations when compared to the SRS-sub estimator, while the CRS-comb estimator nearly triples the effective subsample size. Additionally, the problematic estimates for the SRS-sub and SRS-comb estimators have been removed for this analysis. Therefore, the effective sample sizes reported for the NWRS-comb estimator are conservative. Furthermore, the predicted probabilities (not the odds ratios) are usually the parameters of interest. The use of the NWRS-comb estimator provides an even greater amount of information about the intercepts, and since the predicted probabilities are a combination of the slopes and the intercepts, the NWRS-comb estimator will be far more informative for the predicted probabilities.

7 Discussion

In this paper we introduced the problem of ecological inference in GLMs, developed a maximum likelihood approach for estimation and inference, and proposed methodology for optimal subsampling design conditional on the ecological data. While, this methodology was only applied to an illustrative example and hence the full benefits (and possible drawbacks) were not fully developed, there appears to be great promise for optimal design conditional on ecological data. In addition to the voter turnout example that was introduced in this paper, there are a number of potential applications for this methodology in disciplines where ecological inference is common. This is particularly true in environmental epidemiology where aggregate data are often available on disease counts, covariates, and some measure of exposure. With rare diseases, Poisson regression is often appropriate, and hence an estimate of the joint distribution of exposure and covariates is sufficient to remove ecological bias. Additionally, individual level data are now available for a number of covariates through the Public Use Microdata Survey (Ruggles et al., 2004). Hence, we need only obtain information on the distribution of exposure conditional on covariates. Unfortunately, individual level data on exposure can only be obtained at great cost. The techniques developed in this paper are ideally suited to reducing this cost.

References

Abelson, R. P., E. F. Loftus, and A. G. Greenwald (1992). Attempts to improve the accuracy of self-reports of voting. In J. M. Tanur (Ed.), *Questions About Questions: Inquiries into the Cognitive Bases of Surveys.* New York: Russell Sage.

Anderson, B. and B. Silver (1986). Measurement and mismeasurement of the validity of the self-reported vote. *American Journal of Political Science 80*, 771–785.

Burden, B. C. (2000). Voter turnout and the national election studies. *Political Analysis 8*, 389–398.

Chaudhuri, S., M. Handcock, and M. Rendall (2006). Generalised linear models incorporating population level information: An empirical likelihood based approach. *Submitted for Publication.*

Freedman, D., S. Klein, M. Ostland, and M. Roberts (1998). Review of a solution to the ecological inference problem. *Journal of the American Statistical Association 93*, 1518–1522.

Freedman, D., S. Klein, J. Sacks, C. Smyth, and C. Everett (1991). Ecological regression and voting rights (with discussion). *Evaluation Review 15*, 673–816.

Glynn, A., J. Wakefield, M. Handcock, and T. Richardson (2008). Alleviating linear ecological bias and optimal design with subsample data. *Journal of the Royal Statistical Society: Series A 171*, 179–202.

Goodman, L. (1953). Ecological regression and the behavior of individuals. *American Sociological Review 18*, 663–664.

Greenland, S. (1992). Divergent biases in ecologic and individual-level studies. *Statistics in Medicine 11*, 1209–1223.

Greenland, S. and H. Morgenstern (1989). Ecological bias, confounding, and effect modification. *International Journal of Epidemiology 18 (1)*, 269–274.

Handcock, M., M. Rendall, and J. Cheadle (2003). Improved regression estimation of a multivariate relationship with population data on the bivariate relationship. *Sociological Mehthodology 35*, 303–346.

Haneuse, S. J. and J. C. Wakefield (2008). The combination of ecological and case–control data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 70*(1), 73–93.

Hellerstein, J. and G. Imbens (1999). Imposing moment restrictions from auxillary data by weighting. *The Review of Economics and Statistics 81 (1)*, 1–14.

Imbens, G. and T. Lancaster (1994). Combining micro and macro data in microeconomic models. *Review of Economic Studies 61*, 655–380.

Jackson, C., N. Best, and S. Richardson (2006). Improving ecological inference using individual-level data. *Statistics in Medicine 25*, 2136–2159.

King, G. (1997). *A Solution to the Ecological Inference Problem.* Princeton: Princeton.

King, G., O. Rosen, and M. Tanner (2004). Information in ecological inference: An introduction. In G. King, O. Rosen, and M. Tanner (Eds.), *Ecological Inference: New Methodological Strategies.* Cambridge: Cambridge University Press.

McCullah, P. and J. Nelder (1989). *Generalized Linear Models, 2nd Edition.* London: Chapman & Hall.

Piantadosi, S., D. Byar, and S. Green (1988). The ecological fallacy. *American Journal of Epidemiology 127*, 893–904.

Prentice, R. and L. Sheppard (1995). Aggregate data studies of disease risk factors. *Biometrika 82*, 113–125.

Raghunathan, T., P. Diehr, and A. Cheadle (2003). Combining aggregate and individual level data to estimate an individual level correlation model. *Journal of Educational and Behavioral Statistics 28*, 1–19.

Richardson, S. (1992). Statistical methods for geographical correlation studies. In P. Elliott, J. Cuzick, D. English, and R.Stern (Eds.), *Analysis of Survey Data*, pp. 181–204. New York: Oxford University Press.

Richardson, S., I. Stucker, and D. Hemon (1987). Comparison of relative risks obtained in ecological and individual studies: some methodological considerations. *International Journal of Epidemiology 16*, 111–120.

Robinson, W. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review 15*, 351–357.

Ruggles, S., M. Sobek, T. Alexander, C. Fitch, R. Goeken, P. Hall, M. King, and C. Ronnander (2004). Integrated public use microdata series: Version 3.0 [machine-readable database].

Salway, R. and J. Wakefield (2004). A common framework for ecological inference in epidemiology, political science, and sociology. In G. King, O. Rosen, and M. Tanner (Eds.), *Ecological Inference: New Methodological Strategies*. Cambridge: Cambridge University Press.

Sheppard, L. (2003). Insights on bias and information in group-level studies. *Biostatistics 4 (2)*, 265–278.

Silver, B., B. Anderson, and P. Abramson (1986). Who overreports voting? *American Political Science Review 80*, 613–624.

Steel, D., E. Beh, and R. Chambers (2004). The information in aggregate data. In G. King, O. Rosen, and M. Tanner (Eds.), *Ecological Inference: New Methodological Strategies*. Cambridge: Cambridge University Press.

Steel, D., M. Tranmer, and D. Holt (2003). Analysis combining survey and geographically aggregated data. In R. Chambers and C. Skinner (Eds.), *Analysis of Survey Data*. New York: Wiley.

Traugott, S. (1989). Validating self-reported vote: 1964-1988. *American National Election Study Technical Report*.

Wakefield, J. (2003). Sensitivity analysis for ecological regression. *Biometrics 59*, 9.

Wakefield, J. (2004). Ecological inference for 2x2 tables (with discussion). *Journal of the Royal Statistical Society - A 167*, 385–445.

Wakefield, J. and R. Salway (2001). A statistical framework for ecological and aggregate studies. *Journal of the Royal Statistical Society - Series A 164*, 119–137.