Models, assumptions, and model checking in ecological regressions*

Andrew Gelman,[†] Stephen Ansolabehere,[‡] Phillip N. Price,[§]
David K. Park,[¶] and Lorraine C. Minnite[||]

June 30, 2000

Abstract

Ecological regression is based on assumptions that are untestable from aggregate data. However, these assumptions seem more questionable in some applications than others. There has been some research on implicit models of individual data underlying aggregate ecological regression modeling. We discuss ways in which these implicit models can be checked from aggregate data. We also explore the differences in applications of ecological regressions in two examples: estimating the effect of radon on lung cancer in the United States, and estimating voting patterns for different ethnic groups in New York City.

Keywords: ecological regression, epidemiology, radon, voter turnout

1 Introduction

Ecological regression is the statistical method of running regressions on aggregates (typically, averages within geographic districts) and interpreting these regressions as predictive relations on the level of individual units. For as long as ecological regression methods have been formally used, it has been recognized that they rely on assumptions that are untestable with aggregate data (see Robinson (1950), Goodman (1953, 1959), and, for a recent review, Greenland and Robins (1994). Nonetheless, in many settings ecological regressions will be used in order to make as much sense as possible out of existing aggregate data.

For some applications, ecological regression seems quite natural, whereas for others, the untestable assumptions seem to pose an impossible barrier to interpretation of the results. The fundamental difficulty with the use of ecological regression lies in checking the adequacy of the model's assumptions. For example, Freedman et al. (1991) demonstrate that conventional model assessment statistics such as mean squared error cannot distinguish among the possible models.

*For the Royal Statistical Society conference on the analysis and interpretation of disease clusters and ecological studies, London, 16–17 December 1999. We thank the editor and two referees for helpful comments and the U.S. National Science Foundation for support through grant SBR-9708424 and Young Investigator Award DMS-9796129.

†Department of Statistics, Columbia University, New York, NY
‡Department of Political Science, Massachusetts Institute of Technology, Cambridge, MA
§Indoor Environment Program, Lawrence Berkeley Laboratory, Berkeley, CA
¶Department of Political Science, Columbia University, New York, NY
||Department of Political Science, Barnard College, New York, NY

Despite the broad concern about the assumptions underlying ecological regression, little has been done on developing data-based methods for checking the reasonableness and fit of these models. In this paper we study strategies for model checking based on predictions of the estimates at different levels of aggregation. We explore the assumptions underlying ecological regression in two examples: estimation of the effect of radon on lung cancer in the United States, and estimation of voting patterns for different ethnic groups in New York City. For both examples, we find evidence from the data themselves that the models do not fit; thus, the modeling assumptions can in fact be falsified by data, though never confirmed. We also discuss issues of internal consistency of ecological regression and proposed alternative models.

2 Ecological regression

2.1 Models

2.1.1 Notation and basic model

We work with a modified version of the notation of Ansolabehere and Rivers (1991a,b), which is derived from Goodman (1953, 1959). Consider a population of units i partitioned into districts $j = 1, \ldots, J$, with n_j units in each district j. For each unit there is a predictor variable x_{ij} and an outcome y_{ij}, and interest lies in the linear prediction of y from x within each district:

$$y_{ij} = \alpha_j + \beta_j x_{ij} + \epsilon_{ij},$$

where the ϵ_{ij}'s have zero mean and are independent of the x_{ij}'s and each other. We would like to estimate the α_j's and β_j's; however, the individual measurements x_{ij} and y_{ij} are unavailable; we observe only the means, $\bar{x}_j$ and $\bar{y}_j$, within each district. Assuming the n_j's are reasonably large within each district j, we ignore the average errors, $\bar{\epsilon}_j$, to yield the approximate relation,

$$\bar{y}_j = \alpha_j + \beta_j \bar{x}_j \tag{1}$$

Ecological regression is the procedure of estimating the parameters in (1) using a regression of $\bar{y}_j$ on $\bar{x}_j$, with one data point for each of the J districts:

$$\bar{y}_j = \alpha + \beta \bar{x}_j + \eta_j, \tag{2}$$

where the η_j's have zero mean and are independent of the $\bar{x}_j$'s and each other. The basic idea is to use the aggregate parameters α and β in place of the local regression coefficients α_j and β_j, but such a step requires justification, as discussed in Section 2.2.

2.1.2 Reexpression for binary data

Ecological regression is often used for binary predictors x_{ij} and binary outcomes y_{ij}. In this case, $\bar{x}_j$ and $\bar{y}_j$ are interpreted as observed proportions within districts, and the identity (1) is often reparameterized as

$$\bar{y}_j = \beta_{j1}\bar{x}_j + \beta_{j2}(1 - \bar{x}_j) \tag{3}$$

and fitted with the ecological regression model

$$\bar{y}_j = \beta_1\bar{x}_j + \beta_2(1 - \bar{x}_j) + \eta_j, \tag{4}$$

In the notation of (1), $\beta_{j1} = \alpha_j + \beta_j$ and $\beta_{j2} = \alpha_j$. These new parameters β_{j1} and β_{j2} are the conditional probabilities of $y = 1$ given $x = 1$ or $x = 0$. The most important feature of this new model is not the linear transformation of the parameters but rather the restriction, caused by the binary nature of x and y, that the β_{j1}'s and β_{j2}'s lie between 0 and 1. In fact, this information alone can be used to bound the values of the aggregate population quantities. This technique is known as the method of bounds (Shively, 1991).

2.1.3 Quantities of interest

Ecological regression is commonly used for two purposes, corresponding to two different estimands or quantities of interest based on equations (1) and (3). The first purpose, which we illustrate in Section 3 with the radon example, is to understand the individual-level relation between the outcome, y, and the predictor, x, in which case the parameters β_j in (1) are of direct interest, and they are typically estimated all at once by estimating β in (2). In this setting, the ecological regression is hard to interpret if the β_j's vary much from district to district, because this corresponds to an interaction between district and the predictive relation of interest.

The second potential goal of ecological regression, which we illustrate in Section 4 with the voting example, is to estimate the average value of y for different categories of x. Voting rights litigation, in particular, relies heavily on ecological regression, as experts use the technique to estimate the voting patterns of individuals of different ethnic groups. In the notation of (3), the variance among the β_{j1}'s and β_{j2}'s corresponds to local variation in voting patterns within each ethnic group, and the aggregate quantities of interest are

$$\tilde{\beta}_1 = \frac{\sum_{j=1}^{J} n_j\bar{x}_j\beta_{j1}}{\sum_{j=1}^{J} n_j\bar{x}_j} \quad \text{and} \quad \tilde{\beta}_2 = \frac{\sum_{j=1}^{J} n_j(1-\bar{x}_j)\beta_{j2}}{\sum_{j=1}^{J} n_j(1-\bar{x}_j)}. \tag{5}$$

Unlike the coefficient estimates described in the previous paragraph, the aggregate estimands in (5) can well be of interest even if the coefficients vary greatly with j and even if the models underlying ecological regression are not even approximately true (although then some other set of assumptions would be required in order to obtain estimates). In any case, $\tilde{\beta}_1$ and $\tilde{\beta}_2$ are directly interpretable as the population averages of y for the $x = 1$ and $x = 0$ subpopulations.

Under the ecological regression assumptions, inference for the parameters α and β (or, equivalently, β_1 and β_2) is straightforward. For inference about $\tilde{\beta}_1, \tilde{\beta}_2$, however, one can sometimes do quite a bit better in the binary case by constraining the individual β_{j1}'s and β_{j2}'s to be bounded between 0 and 1 (Duncan and Davis (1953), Goodman (1959), Ansolabehere and Rivers (1991a), King (1997)).

2.1.4 Adding more predictor variables

The model can be most obviously expanded by replacing the data vector x with a data matrix X of k variables, in which case $k+1$ ecological regression coefficients can be estimated. Another way to add covariates, in the context of the partitioning model (3), is to allow each of β_{j1}, β_{j2} to depend on additional covariates X. The advantage of this formulation is that it keeps the interpretation of β_{j1}, β_{j2} as $\mathrm{E}(y|x = 1)$ and $\mathrm{E}(y|x = 0)$, which would not work if additional X's were simply added as predictors in (3). Hanushek, Jackson, and Kain (1974) demonstrate how to estimate ecological regressions with multiple predictors and how to use the regression results to estimate aggregate quantities such as the literacy rates of blacks, native whites, and foreign-born whites.

Another variant, which we shall discuss in Section 4, involves categorical data, in which each unit in the population must fall in one of M categories, with a set of binary variables $x_{ijm}, m = 1, \ldots, M$, indicating which category is appropriate for unit i in district j. It is then natural to write the ecological regression model as,

$$\bar{y}_j = \beta_{j1}\bar{x}_{j1} + \cdots + \beta_{jM}\bar{x}_{jM}, \quad \text{with } \bar{x}_{j1} + \cdots + \bar{x}_{jM} = 1, \tag{6}$$

and fit it with the regression

$$\bar{y}_j = \beta_1\bar{x}_{j1} + \cdots + \beta_M\bar{x}_{jM} + \eta_j. \tag{7}$$

Aggregate summaries are defined by

$$\tilde{\beta}_m = \frac{\sum_{j=1}^J n_j\bar{x}_{jm}\beta_{jm}}{\sum_{j=1}^J n_j\bar{x}_{jm}} \quad \text{for } m = 1, \ldots, M. \tag{8}$$

As described in the previous paragraph, additional covariates can be added by applying regression models to the β_{jm}'s.

2.2 Assumptions and alternative models

Under what assumptions are the parameters α and β in (2) reasonable substitutes for the district-level parameters α_j and β_j in (1)? Ansolabehere and Rivers (1991a) identify two such models. The simplest assumption,

$$\text{constancy model:} \quad \alpha_j \equiv \alpha \text{ and } \beta_j \equiv \beta, \text{ for all } j,$$

corresponds to the regression (2) with errors η_j all equal to zero. More interestingly,

$$\text{zero-correlation model:} \quad \mathrm{E}(\alpha_j|\bar{x}_j) = \alpha \text{ and } \mathrm{E}(\beta_j|\bar{x}_j) = \beta \text{ for all } j,$$

corresponds to errors η_j that are uncorrelated with each other and with the $\bar{x}_j$'s.

Without making some major assumptions, ecological regression has a fundamental indeterminacy, which can be seen in (1) which, given the data, $\bar{x}_j, \bar{y}_j$, is one equation with two unknowns. This problem can be illustrated by a plot that takes advantage of the mathematical identity of the random-coefficient regression (1) and positron-emission tomography (Budinger et al. (1977); see also Shepp and Vardi (1982), Beran, Feuerverger, and Hall (1996), and Feuerverger and Vardi (1999)). In the parameterization (3), the axes of the plot represent the parameters β_{j1} and β_{j2}, and the linear constraint on each district j inherent in (3) is represented by a line with slope $-\bar{x}_j/(1-\bar{x}_j)$ that goes through the point $(\bar{y}_j, \bar{y}_j)$. This sort of plot is standard in tomography and was borrowed by King (1997) for ecological regression.

Figure 1 gives an example. Figure 1a plots the turnout vs. proportion black in 5035 election districts for the 1993 New York City mayoral election, an example we return to in Section 4. Figure 1b is the associated tomography plot, where each line corresponds to a point on Figure 1a and represents the possible values of β_{j1}, β_{j2} that are consistent with the data x_j, y_j. (For visual clarity, only one-tenth of the districts are displayed in Figure 1b.)

This picture corresponds to emission tomography with a limited range of angles (since, if x is a binary variable, all the lines must have negative slope). The key independence assumption, which is satisfied by the physics of emission tomography (but is not in general true in ecological regression applications), is that the emission angles (characterized by $\bar{x}_j$) are independent of the locations (that is, the parameters (β_{j1}, β_{j2})). The tomography lines do not go through a common point, which is equivalent to the fact that the points on the scatterplot do not all fall on a straight line, and implies that the within-district conditional expectations, β_{j1} and β_{j2}, *must* vary among districts.

Before continuing with the ecological regression model, it is important to realize that other models can be set up, consistent with ecological data, under which the ecological regression parameters α and β in (2) are *not* reasonable substitutes for the α_j's and β_j's in (1). The simplest such model is to assume $\beta_j \equiv 0$ in all districts j, which corresponds to $\beta_{j1} = \beta_{j2}$ and thus all points falling on the 45° line in the tomography plot. Freedman et al. (1991) call this the "neighborhood model" (because the assumption is that the expectation of y depends only on its district, or neighborhood, and does not otherwise vary with x) and note that it is consistent with *any* aggregate data set—this can be seen on the tomography plot by noting that all the lines intersect the 45° line of $\beta_{j1} = \beta_{j2}$. Ansolabehere and Rivers (1991b) derive systematic relationships between the estimates from the ecological and neighborhood models of the population proportions in (5).

The neighborhood model is important as a reminder that more than one model can be completely consistent with the data (or, to put it another way, that there are lots of rules one could use to put one point on each line in the tomography plot). However, unlike the ecological regression models, the neighborhood model has the serious flaw that it cannot be generally correct—it can only be valid at, at most, one level of aggregation. For example, if the neighborhood model is valid at the county level, then it will not be valid for towns, and it will not be valid for states. This problem arises because, if the variable x has systematic geographic variation (that is, if x is higher in some areas than others), then any neighborhood effects will aggregate up to correlations of y_j with x_j; and, conversely, zero correlation within neighborhoods (that is, $\beta_j = 0$) cannot be produced unless there is correlation of averaged x and y at smaller levels of aggregation.

2.3 Model checking

Any ecological regression model is based on a huge set of untestable assumptions, since there will be an infinite set of parameters α_j and β_j that are consistent with the observed $\bar{x}_j$'s and $\bar{y}_j$'s. It is perhaps surprising, then, to realize that in many datasets there is actually information available to reject the model.

First consider Figure 2, which displays hypothetical data that are consistent with the ecological regression model. Figure 2 displays a scatterplot of (x_j, y_j) and the associated tomography plot, corresponding to model (3) with (β_{j1}, β_{j2}) close to constant (near the value $(0.2, 0.4)$) with variation independent of x_j. The lines in Figure 2b all go near the point $(0.2, 0.4)$, indicating that the data are consistent with the ecological regression model with nearly-constant parameters β_{j1}, β_{j2}.

In contrast, the scatterplot in Figure 3a and tomography plot in Figure 3b are inconsistent with the zero-correlation assumption underlying ecological regression. This can be seen because an ecological regression (4) fit to these data would correspond to $\beta_1 = -0.2$ and $\beta_2 = 0.4$. We know these cannot be mapped directly to β_{j1} and β_{j2}, because with discrete data, a negative value of β_1 makes no sense. To put it another way, the data in Figure 3 are inconsistent with the assumption that $\mathrm{E}(y_{ij}|x_{ij} = 1)$ and $\mathrm{E}(y_{ij}|x_{ij} = 0)$ are independent of $\bar{x}_j$.

Another model check uses the fact that, under the zero-correlation model underlying (4), $\mathrm{E}(\bar{y}_j|\bar{x}_j)$ should be a linear function of $\bar{x}_j$. For example, the scatterplot of turnout vs. proportion black in Figure 1a looks like a curved regression. We can fit a linear regression model and then plot the binned residuals, which, as can be seen in Figure 4, display a clear pattern, indicating that the data reject the hypothesis that the within-district rates of voter turnout among blacks and nonblacks are independent of the proportion of blacks in the district.

In considering model violations, we might focus our thinking by asking how we would interpret a plot of $\bar{y}_j$ vs. $\bar{x}_j$ with clear curvature and a low variance. In this hypothetical setting, we would know that β_{j1} and β_{j2} are correlated with $\bar{x}_j$, but we would suspect this variation to have a systematic pattern, either related to unobserved covariates (in the epidemiological examples) or possibly to community effects that are mediated by the ethnic mix of the district (in the voting example).

Different methods of fitting ecological models react differently to data that are inconsistent with the model. One strength of the straight linear regression fit to (2) is that nonlinear residuals indicate model misfit, as do estimated parameters outside the range $[0, 1]$ for model (4) fit to aggregates of binary data. In contrast, the Bayesian method reviewed by King (1997) always gives estimates for (4) that lie between 0 and 1; this is presented by as a benefit, but it is in fact misleading if it causes the user to ignore signs of model failure. We must distinguish between out-of-bounds estimates of the parameters (which indicate model violation) and bounds on the individual coefficients β_{jm} (which hold in any case, and can be used to increase the precision of estimates of aggregates).

Finally, as noted earlier, whether or not the model fits, the quantities $\tilde{\beta}_1$ and $\tilde{\beta}_2$ are well defined, and different assumptions that are consistent with the data can imply a range of possibilities for these population quantities (Duncan and Davis (1953), Freedman et al. (1991), Ansolabehere and Rivers (1991a)).

3 Home radon and lung cancer

3.1 Background

Radon is a radioactive gas whose decay products, which are also radioactive, are known to cause lung cancer if high concentrations are inhaled, as has been convincingly demonstrated in studies of highly exposed populations such as miners. Radon occurs naturally in air and can become concentrated indoors; most of the lifetime radiation exposure for most people is attributable to radon. However, the extent to which risk estimates based on highly exposed populations can be used to estimate risks for more typical exposures is unknown. A conventional linear-no-threshold extrapolation suggests that even fairly ordinary levels of radon might cause substantial risk, but some researchers believe that such extrapolation greatly overestimates the risks due to low exposures, and a few researchers even think that low levels of radiation are good for you.

The ideal way to address this issue would be to perform a study on a large random sample from the general population, measuring the lifetime radon concentration and smoking habits for many people and recording whether or not they die of lung cancer. Smoking must be included because it is a key confounder: even under the most pessimistic assumptions about radon risk, the vast majority of lung cancers should occur among smokers; moreover, radon is thought to pose a larger risk (per unit exposure) to smokers than to nonsmokers.

Unfortunately the actual data are much inferior to the ideal case. Indoor radon concentrations have been measured in random surveys including thousands of houses in the U.S., but the smoking habits of the occupants are unknown, as is the eventual cause of death. Additional complications include population mobility, temporal variability of radon concentrations, the fact that people spend only part of the day in their homes, and the existence of other factors that contribute to lung cancer. A few case-control studies have been performed to try to address these problems, basically by attempting to approach the ideal case discussed above. Significant effects appear at high doses (breathing large amounts of radioactive gas really does cause cancer), but the studies are too small to be helpful for low doses—the confidence intervals are so wide that, although they are consistent with the effect extrapolated from high doses using the linear-no-threshold model, they are also consistent with no effect or even with small protective effect from radon exposure at low doses.

One seemingly natural way to address the problem of insufficient individual-level data is to perform an ecological regression. Although radon concentrations, smoking habits, and causes of death are not all known for most individuals, county-average values can be measured or estimated. Perhaps a regression of county lung cancer death rate on county-average indoor radon concentration and county-average smoking intensity could shed some light on the matter?

Such an analysis has been done, by Cohen (1995, 1997), who finds that both before and after adjustment for smoking, county lung cancer rates are *negatively* correlated with county-average indoor radon concentrations. (Figure 5 illustrates for the states of Georgia and Iowa, which we selected to include one Southern and one non-Southern state, each with a reasonably large variation in county-average home radon levels.) Moreover, the results are the same for various subsets of the counties obtained by stratifying on a large number of socioeconomic variables. Cohen points out that even if the estimated county smoking rates are seriously in error, the *errors* would need to be (implausibly) strongly correlated with county mean radon concentrations for those errors to be the cause of the negative correlation between lung cancer and radon. Some people, including Cohen, take this as convincing evidence that the linear-no-threshold assumption is incorrect and radon is possibly harmless, or perhaps protective, at low doses (see also Bogen, 1996). Are they right: should this ecological regression convince us that radon is good for you?

We address this question in two parts. First, we look at whether the observed negative correlation between county lung cancer deaths and county-average radon concentrations could be due to some sort of aggregation effect; second, we perform some model checking on Cohen's results to see whether there is evidence of significant violations of the modeling assumptions.

3.2 Effects of aggregation

As mentioned in the introduction, warnings against naive interpretation of ecological regression are common. Yet sometimes, as with the radon case, the results seem so clear: counties with higher radon concentrations have lower lung cancer death rates, whether or not the death rates are adjusted for county smoking rates. Realistically, how could this *not* indicate a protective effect from radon? The answer to this question is addressed in a very interesting paper by Lubin (1998), which we briefly summarize here to allow discussion.

We adapt the notation of Lubin (1998) and express the standard linear-no-threshold model in terms of the death rate for any individual as a function of smoking status s and radon exposure x, with

$$\text{probability of lung cancer death} = \Pr(y = 1) = c(1 + \theta s)(1 + \beta x), \tag{9}$$

where $s = 1$ for smokers and $s = 0$ for nonsmokers, and x is the cumulative exposure to radiation from radon decay products. The parameter c represents a sort of baseline risk for nonsmokers who are unexposed to radon; θ defines the relative risk in smokers compared to nonsmokers, and this can be well estimated from other studies (perhaps not adjusting for radon exposure, but this is a minor adjustment compared to the effect of smoking); and β determines the rate at which the relative risk increases as a function of exposure. In Cohen's (or any) county-aggregate model, the observed lung cancer death rate is used to estimate a county-average risk $\bar{y}_j$, which is modeled as linearly proportional to county-average radon $\bar{x}_j$, and also depends on the county's proportion of smokers, $\bar{s}_j$:

$$\bar{y}_j = c(1 + \theta \bar{s}_j)(1 + \beta \bar{x}_j) + \eta_j. \tag{10}$$

This is a slight elaboration of the models considered in Section 2, in that smoking comes in as a multiplicative rather than additive covariate.

Suppose the simple model in (9) is true: lung cancer risk for both smokers and nonsmokers increases linearly with radon exposure, and at a much higher rate for smokers than for nonsmokers. Under this model, wouldn't Cohen's regression necessarily reveal a positive correlation between county average radon and county lung cancer death rates? As Lubin shows, the answer is no. Built into the derivation of (10) from (9) is an important assumption: that the distribution of radon concentrations for smokers is the same as that for nonsmokers. In fact, this assumption is known to be false. Smokers tend to have somewhat lower indoor radon concentrations than nonsmokers (Cohen, 1991), perhaps partly because smokers open their windows more often, but also because of more subtle effects: smoking varies with age and socio-economic status, as do housing type, duration of home occupancy during the day, time of residence in a particular house, etc. If sufficient individual-level (that is, non-aggregated) data were available, then the correlation between smoking and radon could be incorporated into the model. However, eliminating the effects of radon-smoking correlation in the ecological regression would require enough individual-level data to obtain very precise estimates of the correlation within counties, for reasons we discuss next.

If the correlation between smoking and radon varies from county to county, even by a very small amount, this variation can cause a very large discrepancy between the underlying parameter β in (9) and the estimated β in the ecological regression (10); indeed, the two parameters can be of opposite signs, and their magnitudes can vary arbitrarily. (By the way, we have discussed this problem in terms of smoking, but as Lubin (1998) points out, the same argument holds for other potential confounders, such as age: "U.S. lung cancer mortality for ages 70–74 y is over 20 times the rate for ages 40–44 y (NRC 1988)."

In addition, any analysis of the relation between radon and lung cancer has the problem that the most important predictor of lung cancer—smoking status—is not available at the county level. Some attempts have been made to estimate county-average smoking levels from surveys, cigarette sales, or constructing proxies from demographic variables (Cohen, 1995), but none of these methods are very reliable. One could presumably construct a very accurate estimate of smoking levels using age-adjusted lung cancer rates, but this cannot be used in the ecological regression, of course, as it would eliminate any attempt to estimate the effects of radon conditional on smoking rate.

3.3 Internal evidence of problems with the model

Ecological regressions generate misleading results (in the sense that coefficients for aggregated data differ from those for individual data) if the causes of variation between aggregation units differ from the causes of variation within aggregation units. Since there is no way of investigating within-unit effects using only aggregated data, it might seem that there is no way of performing data-based model checking for this sort of problem, but that is not always the case, as we illustrate.

Cohen (1995) performed a large number of validation checks, mostly based on stratification: in addition to regressing lung cancer death rate on average radon concentration for every county for which data were available, regressions were also performed on various subsets of the data. For each of 56 county-level socio-economic variables taken from census data, counties were stratified into quintiles and the regression was performed on each stratum. For instance, separate regressions were performed for counties in each quintile of percent urbanization. Results of these regressions show remarkable stability for all the quintiles and all the socioeconomic variables. Thus the regression passes one test that could have revealed the presence of important confounding variables.

Many researchers would be tempted to stop there, but Cohen performed several additional checks. One of these was to stratify geographically, by both state and region (using the Census Bureau's division of the U.S. into nine regions). This is important because both radon and smoking vary geographically, as most confounding variables would be expected to do as well. In contrast to the stability of the results on stratification by socio-economic variables, regional stratification led to substantial variation, with the regression coefficient for radon varying from 0.03 in the East North-Central region to -0.39 in the West South-Central region (in units of % per Bq/m^3).

As Price (1995) pointed out, the regressions within individual states, though subject to substantial stochastic noise due to small sample sizes for radon and lung cancer deaths, show a remarkable pattern: of the 12 states with estimated coefficients below the overall average, 11 are in the South. Cohen (1995) suggests that this phenomenon is caused by a nonlinearity in the effect of radon: radon is strongly protective at low concentrations, leading to a substantial negative coefficient for Southern states (in which most counties have low radon concentrations), whereas it is less protective, or even harmful, at higher concentrations.

However, this is unlikely to be the real explanation: several non-Southern states have low radon concentrations but do not have large negative coefficients for risk from radon, and some Southern states have moderate radon concentrations, similar to those in non-Southern states, but still show large negative correlations. In short, there is a "Southern-ness" confounder: based on Cohen's results, it appears that radon is much more protective in the South than elsewhere in the U.S., an effect that can be only partially explained by a hypothesized nonlinearity in the risk from radon. This effect is apparent even in the county-aggregated data; of course, there may be additional within-county confounding that cannot be found this way.

Given the inherent problems with ecological regression, the fact that a protective effect from breathing radioactive gas seems unlikely based on both non-ecological studies at higher doses and biological understanding of the low-dose effects of alpha particles (which differ from other radiation for technical reasons outside the scope of this paper), and the model violation (the Southern-ness confounder) evident in this particular ecological regression analysis, we do not think that the regression provides meaningful evidence for a protective effect at low doses.

Of course, the regression does not provide evidence the other way, either. Suppose the ecological regression generated results more in line with the miner studies and expectations from cellular biology, showing a linear or near-linear increase in risk with exposure. Almost certainly, many of the same people who disparage Cohen's analysis in which radon appears protective would embrace it if radon appeared dangerous. This should be a caution to us when interpreting any ecological regression; in particular, rigorous model checks should be attempted no matter the results of the analysis, and results of any ecological analysis should be treated with caution.

4 Ethnicity and voting

4.1 Ecological regression in voting studies

Ecological regression and correlation has been used for many years to estimate individual political behavior given aggregate information at the level of precincts or election districts. Gosnell (1937), for example, used ecological regressions to estimate how different ethnic groups in the city of Chicago voted for different candidates, parties, and political machines, and how newspaper readership and endorsements affected individual level voting behavior. In the voting rights field, experts today rely heavily on ecological regression because the United States Supreme Court concluded in *Thornburgh v. Gingles* that ecological regression provides evidence of the voting behavior of individuals of different ethnic groups. Ansolabehere and Rivers (1991a) provide further review of this literature. Ecological regression is also used to study the transition probabilities from aggregate voting data (Brown and Payne, 1986; McCue, 1995).

We focus on the problem of estimating votes and turnout for different ethnic groups: $y_{ij} = 1$ if person i in district j turns out to vote in a given election, and $x_{ij} = 1$ if that person is in a given ethnic group. The units j are electoral districts; the aggregate election numbers, $\bar{y}_j$, are public information, and the demographic proportions by district, $\bar{x}_j$, can be approximated by putting together Census numbers.

We begin by fitting a standard ecological regression, but our diagnostic tools reveal model errors, and so we go beyond the usual model, which in turn requires us to abandon the individualistic political assumptions that underlie it. We illustrate with an example.

4.2 New York City mayoral election

The 1993 New York City mayoral election between the (black) Democratic incumbent, David N. Dinkins, and his (white) Republican challenger, Rudolph W. Giuliani, was racially polarized, as indicated by campaign issues, rhetoric, and opinion polls. This is the sort of election for which political scientists would be interested in the voting patterns of different ethnic groups, which here we categorize as black, non-black latino, white, and other. These terms are of course somewhat arbitrary, and we follow standard practice for Census and opinion polls and rely on self-categorization.

The margins of Table 1 give the total votes and population for New York City adults, and the goal of the analysis is to estimate the interior cells. The data analyzed come from two sources: the 1990 Census of Population and Housing Public Law 94-171, and computerized voting returns for the 1993 mayoral general election obtained from the New York City Board of Elections (see Park, Slotwiner, and Minnite, 1998). The PL94-171 data present counts of the population by ethnicity at the Census block level. The voting returns file contains votes cast for each candidate in the 1993 mayoral general election in the 5694 election districts in the city. To estimate the population by ethnicity in each district, we aggregated the census block information to the election district level and used a conversion file from the New York City Districting Commission to aggregate census block fragments created where election district lines split the census block geography. In some cases, the census data were obviously flawed (for example, more registered voters listed than adults in the district). After removing these from our dataset we were left with 5133 election districts representing 96% of the adult population and 94% of the voters. This large sample size means that, in interpreting our model results below, essentially all observed differences are statistically significant. Finally, for our model fitting we exclude the districts with fewer than 100 voters (because we did not want to have to worry about variability due to discreteness within districts, since that is peripheral to the main concerns of this paper); these 525 districts contained fewer than 0.1% of the population and voters in the city.

4.3 Model fitting

To estimate the cells of Table 1 using ecological regression, we follow the standard approach and separately estimate the voter turnout and the proportion of Giuliani voters by ethnic group. In our discussion of model checking, we shall focus on the models for turnout, but for filling in Table 1 (as is done in different ways in Table 2), we apply models to both turnout and Giuliani vote.

We first discuss the basic ecological regression model and then a nonlinear generalization that fits the data better (but, as we shall see, still has serious problems).

4.3.1 The basic ecological regression model

We begin by fitting model (7) with $M = 4$, where $y_{ij} = 1$ if adult i in district j turns out to vote, and $x_{ijm} = 1$ if adult i in district j belongs to ethnic group m. Figure 6 displays scatterplots of $\bar{y}_j$ vs. $\bar{x}_{jm}$, for each $m = 1, \ldots, 4$. The estimates of the ecological regression parameters β_m, representing the expected turnout rates among the four ethnic groups according to the linear model, are displayed in the first row of Table 2. The fourth coefficient, representing the average proportion of "others" who turn out to vote, is outside the range $[0, 1]$, which indicates a model violation. We continue, however, by re-running the model with the parameters constrained to fall in $[0, 1]$; the constrained parameter estimates are displayed in the second row of Table 2.

The next two rows of the table display the analogous estimates for the proportion of adults who vote for Giuliani, fitted as before but with the additional constraint that $\beta_m^{\text{Giuliani}} \leq \beta_m^{\text{turnout}}$ for each m, to reflect that Giuliani voters are a subset of voters. (We iterate the turnout and vote-choice regressions to get the best fit under these constraints.) We use the turnout and Giuliani-vote models to get estimates of the proportions of adults within each district j who make each vote choice. King (1997) reviews a formal Bayesian method for combining the inference from the regression model with the bounds on the individual β_{jm}'s; however, this method has been developed only for $M = 2$, and so here, with $M = 4$, we perform a simpler procedure based on imputing $(\beta_{j1}, \ldots, \beta_{j4})$, for each j, to be as close as possible to the fitted linear model parameters $(\beta_1, \ldots, \beta_4)$, subject to the equality constraint (6) and the bounds of 0 and 1. We wrote an iterative computer program to perform this four-dimensional constrained optimization within each district. Finally, we repeat the procedure to estimate the proportions of Giuliani voters as the best fit to the regression model under the constraint of being no greater than the voter turnout in each district and ethnic group.

Once we have estimates of Giuliani vote and voter turnout for each ethnic group in each district, we average over districts as in (8) to estimate population proportions, which are displayed in the rows 5–7 of Table 2. These numbers look dubious, especially the estimate of 0% Giuliani vote among blacks and 1% turnout among others, both of which are clearly influenced by the parameter estimates on the boundary.

4.3.2 Ecological regression model with interactions

In addition, we can check the ecological regression model by comparing the observed district turnouts $\bar{y}_j$ to the model (7). Since there are more than two groups, we no longer expect the regression line of $\bar{y}_j$ on any $\bar{x}_{jm}$ to be a straight line, but we still would expect the residuals to be patternless if the model were indeed true. Figure 7 displays the residuals, which display clear patterns, implying that voter turnout within groups is correlated with the ethnic composition of districts.

It is then natural to fit a random-coefficient model for β_{jm}'s in (7) of the form,

$$\mathrm{E}(\beta_{jm}|x_j) = \gamma_m \bar{x}_{jm} + \delta_m (1 - \bar{x}_{jm}), \quad \text{for } m = 1, \ldots, 4. \tag{11}$$

The parameters γ_m and δ_m represent the expected turnout for group m, in districts composed of 100% and 0%, respectively, of group m residents. Ansolabehere and Rivers (1991a) discuss the rationale behind and some difficulties with this model.

A key difference compared to the usual ecological regression models is that the new model (11) is inherently a group model, not just an individual model, in that it explicitly models the ethnic composition of a district as a predictor or influence on individual voting behavior. This sort of model may be more appropriate when studying social behavior (such as voting) than when studying non-contagious disease (such as cancer).

Continuing with our example, the second section of Table 4 display the estimated parameters for model (11), fit to turnout, in raw and constrained form, followed by the estimated parameters for the model, fit to Giuliani vote with the parameters constrained to be no greater than the corresponding parameters for the turnout model. Incidentally, the big differences between the unconstrained and constrained parameters can be viewed as yet another model check, showing that the fitted model is inappropriate to the data in this respect (since, if the model were true, the fitted parameters would automatically be consistent with each other, with any exceptions being minor and caused by small-sample estimation variation).

We then combine the models for turnout and Giuliani vote to obtain estimated proportions of adults of each ethnic group in each district who voted for Giuliani, voted for others, or did not vote. As described at the end of Section 4.3.1, we constrain these proportions as appropriate and then perform the weighted averages over districts to get population proportions within each ethnic group, which are displayed in the last three rows of the second section of Table 2. By comparison, the estimated population proportions under the neighborhood model are displayed in the third section of Table 2. These estimates tend to understate the differences between ethnic groups, as is typical with this model (see Ansolabehere and Rivers, 1991b) and, as discussed at the end of Section 2.2, this model does not make logical sense; however, many of its population proportions are reasonable in this example.

Also, as before we examine model fit by plotting binned residuals; these plots (not shown here) do not look too bad, although we have already seen (from the parameter estimates that violate the constraints) that the model still has problems.

4.4 Comparison to exit polls and directions for further work

We can compare our results to those obtained from exit polls taken from respondents at 40 polling places in New York City. Unfortunately, we have not been given the election districts of these polling places and so we can only use the totals from these polls. The estimated vote decisions from these polls are displayed in the final three lines of Table 2. These proportions can be compared to the estimated proportions from the models, as displayed in the earlier segments of the table. There are some discrepancies that indicate probable errors in the models, even in the better-fitting nonlinear model, most notably in the vote assignments for the "other" category. We must be careful in interpreting these comparisons, however, because the exit poll is far from a random, equal-probability sample and is subject to many biases.

Natural next steps for this problem are to include demographic information as predictors of the γ_{jm} and δ_{jm} parameters in the expanded ecological regression model, and to combine in the information from the exit poll, following the ideas of Ansolabehere and Rivers (1991b).

5 Discussion

What have we learned? Ecological regression requires lots of assumptions, and we have considered various kinds of model error:

- Theoretical inconsistencies (as in the discussion of the neighborhood model in Section 2.2 of this paper, or the theoretical examination of Lubin (1998) of the implicit model of Cohen's (1995) ecological regression for radon);
- Parameter estimates that are outside the bounds, thus invalidating the zero-correlation model (as in Figure 3) or more complicated models (as in the negative unconstrained estimate for $\delta_4^{\text{turnout}}$ in Table 2)
- Patterns in binned residuals from the regression (as in Figure 4), which can suggest interactions or other extensions of the model.

The rejection of an ecological regression model can have interesting substantive interpretations, as in the voting example in Section 4.3.2 where an interaction can be interpreted as a community effect. But, even if a model *could* be correct and we cannot reject it given the observed aggregate data, it still relies on an individual-level assumptions.

What, then, can be done? One approach is to combine ecological regression with individual data—as discussed by Ansolabehere and Rivers (1991b), even a small amount of individual-level survey data can be used to check or modify the zero-correlation model of ecological regression in a voting study. Even where individual data are not directly put in the ecological analysis, this sort of comparison is done informally; for example, in the radon study, Lubin (1998) and Goldsmith (1999) evaluate ecological studies of radon risk with reference to individual-level studies of radon and lung cancer.

What if individual-level data are not available; then why should we study ecological regression? The short answer is that, as Cho (1998, 1999) points out, researchers are interested in estimating individual-level coefficients such as β_j and averages such as $\tilde{\beta}_1$ and $\tilde{\beta}_2$, and often only aggregate data are available. People will be doing ecological regressions whether we like it or not, so we should try to understand this method. A longer answer is that ecological regression sometimes makes a lot of sense; for example, in the radon study, if smoking were not such a dominant confounder, and if accurate smoking rates were available on the county level, then the ecological regression could be informative. In the election example, it has been a struggle to set up an ecological regression model that gives reasonable parameter estimates and fit to data, but this struggle is itself informative in that it reveals that certain natural-seeming models are not, in fact, appropriate. Potential confounders arise with observational studies in general, and it would be a mistake to somehow think of this as a special problem of ecological regression.

References

Ansolabehere, S., and Rivers, R. D. (1991a). Bias in ecological regression estimates. Technical report, Department of Political Science, Massachusetts Institute of Technology.

Ansolabehere, S., and Rivers, R. D. (1991b). Combining aggregate and survey data. Technical report, Department of Political Science, Massachusetts Institute of Technology.

Beran, R., Feuerverger, A., and Hall, P. (1996). On nonparametric estimation of intercept and slope distributions in random coefficient regression. *Annals of Statistics* **24**, 2569–2592.

Bogen, K. T. (1996). Do U.S. county data disprove linear no-threshold predictions of lung-cancer risk for residential radon? A preliminary assessment of biological plausibility. *Human and Ecological Risk Assessment*.

Brown, P., and Payne, C. (1986). Aggregate data, ecological regression, and voting transitions. *Journal of the American Statistical Association* **81**, 452–460.

Budinger, T. F., Derenzo, S. E., Gullberg, G. T., Greenberg, W. L., and Huesman, R. H. (1977). Emission computer assisted tomography with single-photon and positron annihilation photon emitters. *Journal of Computer Assisted Tomography* **1**, 131–145.

Cho, W. K. T. (1998). Iff the assumption fits ...: a comment on the King ecological inference solution. *Political Analysis* **7**, 143–163.

Cho, W. K. T. (1999). Latent groups and cross-level inferences. Technical report, Department of Political Science, University of Illinois.

Cohen, B. L. (1995). Test of the linear no-threshold theory of radiation carcinogenesis for inhaled radon decay products. *Health Physics* **68**, 157–174.

Cohen. B. L. (1995). Response to Price (1995). *Health Physics* **68**, 578–579.

Duncan, O., and Davis, B. (1953). An alternative to ecological correlation. *American Sociological Review* **18**, 665–666.

Feuerverger, A., and Vardi, Y. (1999). Positron emission tomography and random coefficients regression. *Annals of the Institute for Statistical Mathematics*, to appear.

Freedman, D., Klein, S., Sacks, J., Smyth, C., and Everett, C. (1991). Ecological regression and voting rights. *Evaluation Review* **15**, 673–711.

Goldsmith, J. R. (1999). The residential radon - lung cancer association in U.S. counties: a commentary. *Health Physics* **76**, 553–557.

Goodman, L. A. (1953). Ecological regressions and behavior of individuals. *American Sociological Review* **18**, 663–669.

Goodman, L. A. (1959). Some alternatives to ecological correlation. *American Journal of Sociology* **64**, 610–625.

Gosnell, H. F. (1937). *Machine Politics Chicago Model*. University of Chicago Press.

Greenland, S., and Robins, J. (1994). Ecologic studies: biases, misconceptions, and counter examples. *American Journal of Epidemiology* **139**, 747–760.

Hanushek, E., Jackson, J., and Kain, J. (1974). Model specification, use of aggregate data, and the ecological correlation fallacy. *Political Methodology*, 89–107.

King, G. (1997). *Reconstructing individual behavior from aggregate data: a solution to the ecological inference problem*. Princeton University Press.

Lubin, J. H. (1998). On the discrepancy between epidemiologic studies in individuals of lung cancer and residential radon and Cohen's ecologic regression. *Health Physics* **75**, 4–10.

McCue, K. F. (1995). Individual choice and ecological analysis. Technical Report, California Institute of Technology.

Park, D. K., Slotwiner, D. M., and Minnite, L. C. (1998). White, black, and latino voter turnout in 1993 New York City mayoral election: a comparison between King's ecological regression and exit poll data. Paper presented at the American Political Science Association annual meeting, 1998.

Price, P. N. (1995). Letter. *Health Physics* **68**, 577–578.

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

Shepp, L. A., and Vardi, Y. (1982). Maximum likelihood reconstruction for emission tomography. *IEEE Transactions on Medical Imaging* **MI-1**, 113–122.

Shively, W. P. (1991). A general extension of the method of bounds, with special application to studies of electoral transition. *Historical Methods* **24**, 81–94.

Ethnicity of	Voting decision			
voting-age person	Giuliani	not Guiliani	no vote	Total
Black	?	?	?	1,314,520
Latino	?	?	?	1,237,121
White	?	?	?	2,667,127
Other	?	?	?	404,574
Total	929,981	896,909	3,796,452	5,623,342

Table 1: 2×3 contingency table for ecological regression model for turnout and vote in 1993 New York City mayoral election

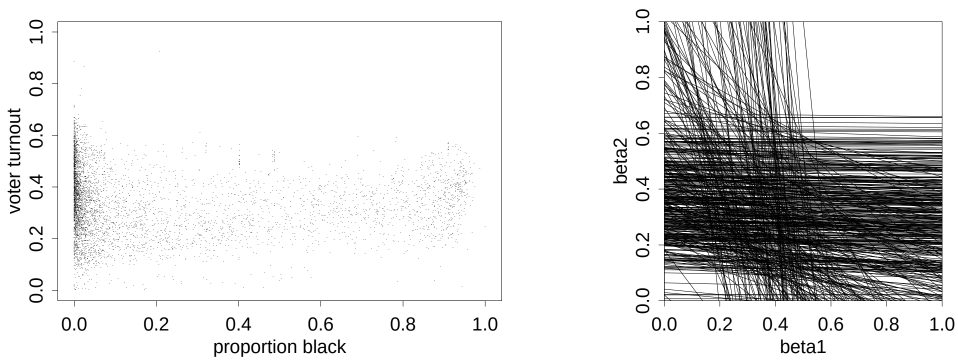

Figure 1: (a) Voter turnout, y_j, (as a proportion of estimated adult population) vs. proportion black, x_j, for districts in the 1993 New York City mayoral election. (b) Tomography plot; each line corresponds to a point on the scatterplot and represents the possible values of β_{j1}, β_{j2} that are consistent with the data x_j, y_j. (For visual clarity, only one-tenth of the districts are displayed in the tomography plot.)

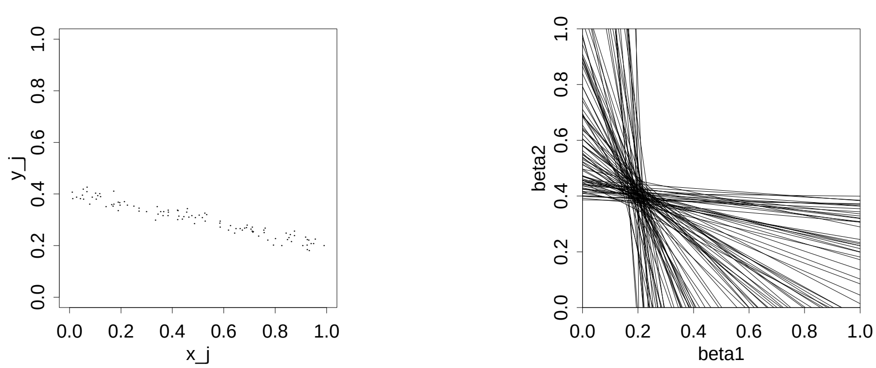

Figure 2: Scatterplot and tomography plot corresponding to simulated data where the ecological regression model is appropriate.

		Black	Latino	White	Other
Ecological regression model	β_m^{turnout} (raw)	.38	.16	.45	$-.09$
	β_m^{turnout} (constrained)	.38	.15	.44	.00
	$\beta_m^{\text{Guiliani}}$ (raw)	$-.01$	.03	.35	$-.03$
	$\beta_m^{\text{Guiliani}}$ (constrained)	.00	.02	.35	.00
	vote for Guiliani	.00	.02	.34	.01
	vote not for Guiliani	.37	.12	.09	.00
	no vote	.63	.83	.57	.99
Expanded ecological regression model	$\gamma_m^{\text{turnout}}$ (raw)	.37	.21	.50	.18
	$\gamma_m^{\text{turnout}}$ (constrained)	.37	.21	.50	.16
	$\delta_m^{\text{turnout}}$ (raw)	.55	.07	.17	$-.05$
	$\delta_m^{\text{turnout}}$ (constrained)	.56	.07	.16	.00
	$\gamma_m^{\text{Guiliani}}$ (raw)	.00	.07	.40	.09
	$\gamma_m^{\text{Guiliani}}$ (constrained)	.01	.07	.40	.16
	$\delta_m^{\text{Guiliani}}$ (raw)	.03	.04	$-.01$	.18
	$\delta_m^{\text{Guiliani}}$ (constrained)	.00	.07	.05	.00
	vote for Guiliani	.01	.07	.30	.03
	vote not for Guiliani	.42	.05	.10	.00
	no vote	.57	.88	.60	.97
Neighborhood model	vote for Guiliani	.04	.10	.25	.16
	vote not for Guiliani	.27	.16	.10	.09
	no vote	.69	.75	.64	.75
Exit poll	vote for Guiliani	.02	.07	.28	.07
	vote not for Guiliani	.35	.11	.08	.07
	no vote	.63	.83	.64	.86

Table 2: Parameter estimates and estimated population summaries for various models for vote choice and turnout in 1993 New York City mayoral election.
(a) For the ecological regression model, parameters represent expected rates of voting for Giuliani and turning out to vote, as proportions of adult population of each ethnic group. Parameters are then constrained to fall between 0 and 1. Proportions of adults in each ethnic group who voted for Giuliani, voted for other candidates, or did not vote, are estimated from the constrained models, also applying to the constraints to individual district parameters β_{jm}, then summing over all districts as in (8). The estimated population proportions, especially for "other," are not plausible and indicate a problem with the model.
(b) For the expanded ecological regression model, parameters are as described in (11). The estimated population proportions, especially for "other," are not plausible and indicate a problem with the model.
(c) The "neighborhood model" assumes $\beta_{j1} = \cdots = \beta_{j4}$ within each district j. This model tends to understate the differences between ethnic groups, and the model is implausible at the district level; nonetheless, many of the resulting aggregate estimates are reasonable.
(d) From an exit poll, we display the proportion of adults in each ethnic group who voted for Giuliani, voted for other candidates, or did not vote. These exit poll data are themselves imperfect as they are far from an equal-probability sample.

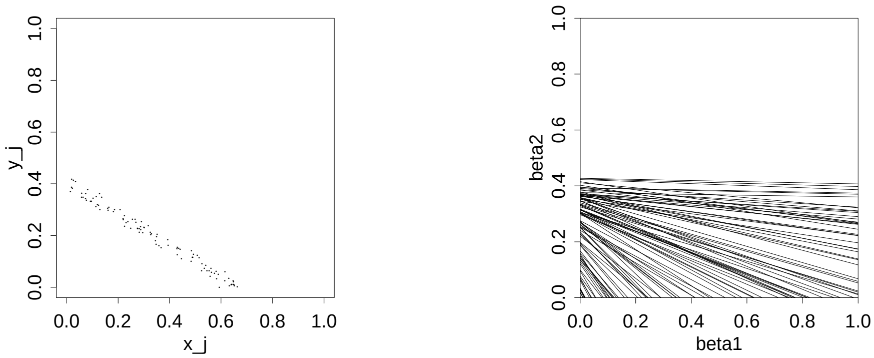

Figure 3: Scatterplot and tomography plot corresponding to simulated data where the ecological regression model is inappropriate.

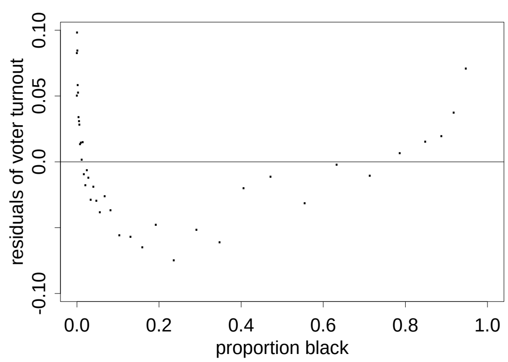

Figure 4: Average residuals of the voter turnout y_j from the ecological regression (2), with x_j divided into 40 bins. The systematic discrepancies from a mean of zero indicate a model misfit: this plot reveals that the proportions β_{j1} and β_{j2}, defined in (3), cannot be uncorrelated with x_j.

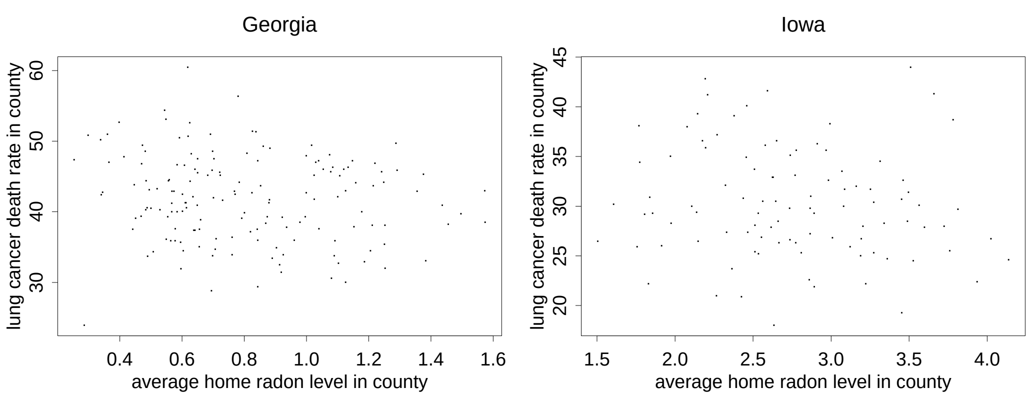

Figure 5: Age-adjusted annual lung cancer death rates (per 100,000) vs. estimated average home radon levels (in pCi/L), by county, for Georgia and Iowa. The (slight) negative correlation in these plots is at first surprising, given that most experts believe that, for any individual, an increase in radon exposure increases the probability of lung cancer.

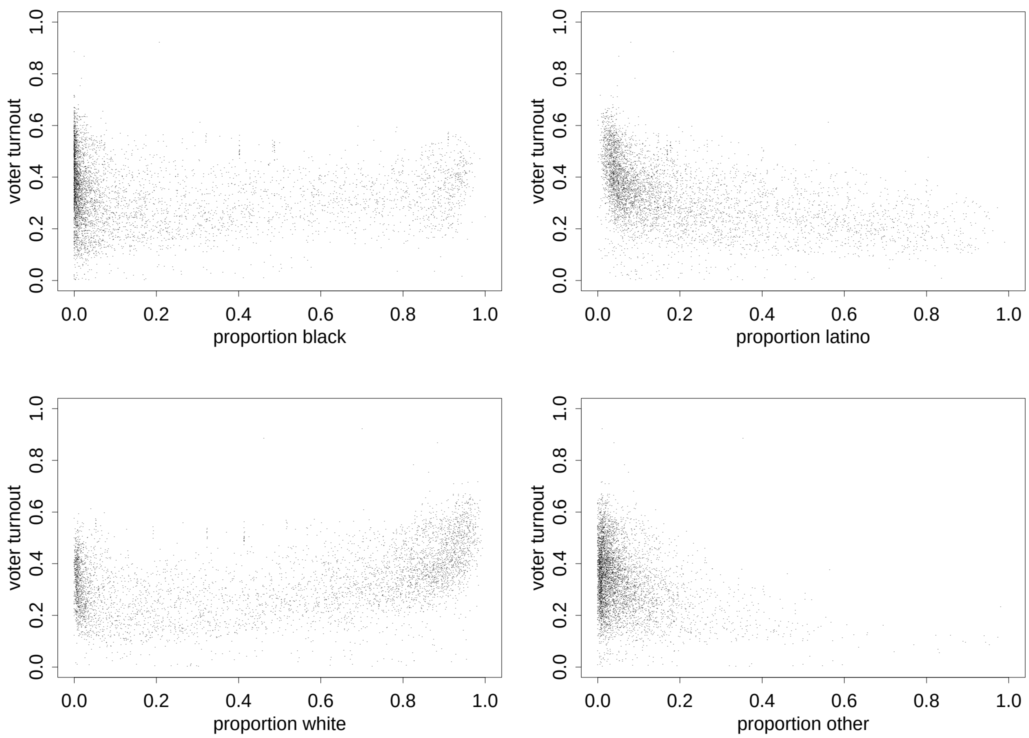

Figure 6: Voter turnout vs. proportion (a) black, (b) latino, (c) white, (d) other, for districts in the 1993 New York City mayoral election.

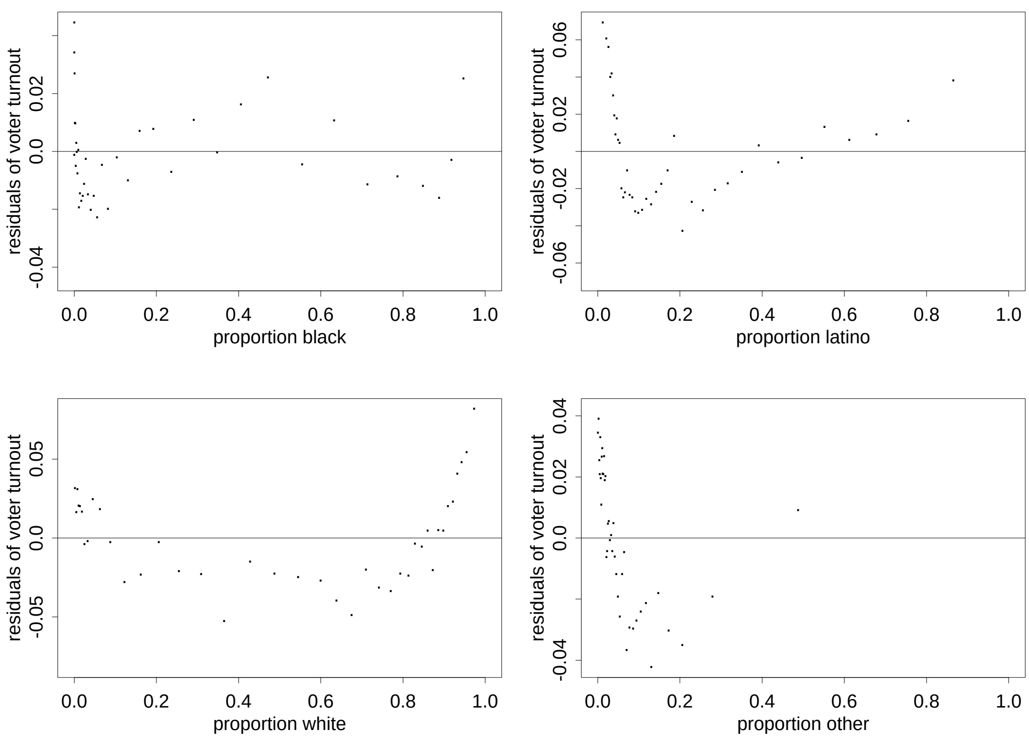

Figure 7: Residuals of voter turnout from the ecological regression model (7), plotted vs. proportion (a) black, (b) latino, (c) white, (d) other, for districts in the 1993 New York City mayoral election. The systematic discrepancies of $\mathrm{E}(y|x)$ from the line in each plot indicate a model misfit: the proportions β_{jm}, defined in (6), cannot be uncorrelated with the x_{jm}'s.