Journal of Data Science **3**(2005), 233-241

On the Radical Views of Coal Miners in the Early Twentieth Century in Southern Illinois

Stephane E. Booth and David E. Booth
Kent State University

Abstract: There has been great interest in the Southern Illinois mine war by historians. An explanation has been that this war was caused by miners who had radical political beliefs. We examine this view by applying four methods of ecological inference to estimate the proportion of coal miners who were socialist voters in this time period. Based on these results (especially considering the assumptions of the methods) we conclude that miners were politically less radical than previously thought.

Key words: Ecological inference, robust regression, Southern Illinois mine war.

1. Introduction

It is often the case in historical research that data needed for a particular study were never recorded. If such data are cell frequencies from a contingency table, and the marginal frequencies were recorded, then under some conditions, it may be possible to recover the cell frequencies. Such a procedure goes under the heading of ecological inference even though that term is more general. We stress that the assumptions are critical. If they are not met then it is not possible to recover the cell frequencies. Following Goodman (1959), we use the term "ecological" to refer to the grouped state. Desired contingency table cell frequencies will be recovered from the aggregate data.

In this paper we consider a problem discussed previously by Booth and Booth (1988). In the early 1930s, a civil war erupted in the coalfields of southern Illinois. This shooting war was between coal mine managers, owners, their soldiers (mine guards, strikebreakers, local law enforcement officers, and often the Illinois National Guard) and mine workers. As has been discussed previously (Booth 1983), the conventional wisdom has been that the war pitted very radical coal miners against the more conservative owners and managers. In this view, miners were considered radical because of their strong union activity. Alternative views are possible. One such might be that miners banded together in unions to protest and ameliorate low wages as well as poor and unsafe working conditions in the mines. The purpose of the present study is to examine exactly how politically radical the miners were by using ecological inference methods to look at the strength of miners support of Socialist candidates in southern Illinois.

To be specific, we wish to consider a problem discussed previously by Booth and Booth (1988). We wish to consider whether or not there was a difference in the proportion of support for socialist candidates between coal miners and non-coal miners in a ten-county region of Southern Illinois in the early twentieth century. If there was such a difference, we wish to know if there was any variation by geographic location. Such an approach can be based on historical voting records. A contingency table for answering this question is shown in Table 1, where the letters indicate the cell and marginal frequencies (Freedman *et al.* 1998) (the total row and column contain the marginal frequencies). The major quantities of interest are p, the proportion of coal miners that are also socialist voters, and r, the proportion of non-miners that are also socialist voters.

Table 1: Basic contingency table

	Coal Miners	Non-miners	Total
Socialist Voters	a	b	$a+b$
Non-socialist Voters	c	d	$c+d$
Total	$a+c$	$b+d$	1

The marginal totals and frequencies are available for Illinois counties, but the cell frequencies are not (Booth and Booth 1988). We will attempt to recover these frequencies using ecological inference techniques. Further, we will stress the importance of the assumptions and attempt to suggest some ways in which the assumptions may be checked and hence which of the estimates from the competing estimation methods are likely to be best. We begin by describing the methods. Thus we wish to estimate p and r as described below.

2. Ecological Regression

We follow Goodman (1959) and Booth and Booth (1988). Let us consider the case where we have some marginal values for a series of contingency tables of the form of Table 1, with each table from a different county. We wish to estimate $p(= a/(a+c))$ and $r(= b/(b+d))$. In ecological regression this is done under the assumption that p and r are constant over the counties. This will be discussed later in the paper.

Let y be the proportion of socialist voters and x the proportion of coal miners in the population. Following Goodman, we may write

$$y = xp + (1 - x)r \tag{2.1}$$

and by algebra

$$y = r + x(p - r) \tag{2.2}$$

Since y and x are known marginal values we can (if the assumptions are met) regress y on x using the county data in order to estimate p and r. The y-intercept of such a regression will be a point estimate of r, and the slope will be a point estimate of $p - r$. Hence the point estimates for r and p are given by

$$r = r\text{-intercept} \tag{2.3}$$

and

$$p = \text{slope} + r \tag{2.4}$$

The variance of p and r are given by Goodman (1959). Since this is a straight line regression all of the usual least squares assumptions must be met if the least squares method is used to estimate p and r with the above equations. Those assumptions are given in Booth and Booth (1988) and many other places as well. As we mentioned, Goodman (1959) indicates that an additional assumption must be met as well. This additional assumption is that p and r (or at minimum $E(p|x)$ and $E(r|x)$) must be constant over the groupings that provide the y and x values needed to perform the regression indicated by equation (2.2). Because p and r are population values the assumptions can be difficult to check. A perusal of the literature suggests that this assumption is often ignored. This is a very dangerous thing to do. Examples of the problems caused by ignoring the additional assumption are illustrated in Booth and Booth (1988). Further discussion is given by Freedman (1999). Freedman refers to the extra assumption as the constancy assumption. In the language of our problem, Freedman would state the constancy assumption as: the voting preferences of miners or non-miners do not systematically depend on the make up of the area of residence. As both references indicate the constancy assumption as well as the usual assumptions of ordinary least squares must be satisfied for ecological regression to be successful in estimating p and r. Full details on ecological regression (and the use of an outlier resistant (robust) regression with it) can be found in Goodman (1959), Booth and Booth (1988) and Freedman (1999).

3. The Method of Bounds

Another method that has been proposed for ecological inference (the estimation of p and r of equations (2.1) and (2.2)) is the method of bounds (Freedman 1999). This method was originally proposed by Duncan and Davis (1953). This method is based on the use of equation (2.1) on a one county at a time basis. As Freedman (1999) remarks, we have one equation in two unknowns, "...which is the problem with ecological inference". In this case, we are primarily interested in p which is, in our example, the fraction of coal miners that were socialist voters in each county. We know that since r is a proportion it must satisfy

$$0 \leq r \leq 1 \tag{3.1}$$

Thus, by substituting values of zero and one for r in equation (2.1) we can obtain upper and lower bounds on the value of p. As Freedman (1999) remarks, these bounds are often too wide to be helpful.

4. The Neighborhood Model

In two papers, Freedman *et al.* (1998)and Freedman (1999) propose a third approach to ecological inference, the neighborhood model. The basic idea is to choose an assumption that is completely opposed to the constancy assumption of ecological regression and see what the effect of the assumption is. The constancy assumption assumes that p and r are constant over county. In the neighborhood model, Freedman proposed an assumption that we will call the neighborhood assumption. The neighborhood assumption is that p and r depend on which neighborhood (county) is being considered. In the language of Table 1 we can use the following equations to estimate p. Recall that these equations are calculated for each county. Let $H =$ the number of socialist voters that are also coal miners, $I =$ number of votes for the socialist candidate in a county, $J =$ number of coal miners in a county, and $N =$ total number of voters in a county. Then we have the estimates,

$$H = \frac{IJ}{N} \tag{4.1}$$

and thus

$$p = \frac{H}{N} \tag{4.2}$$

5. The King Model

Another solution to the ecological inference problem was proposed by King (1997). Freedman *et al.* (1998) showed that this model was problematic and hence we will not consider it here.

6. The Coal Miners

We now consider the specific problem of Illinois coal miners. (The data for this study and the summary statistics are given in Table 2). As mentioned earlier, this data set is important because it speaks to an issue of importance to historians of radical movements during this period. A clear analysis would help to settle a controversy over how radical coal miners were in this period. We will use ecological regression to determine p and r, the portion of coal miners and non-miners respectively that cast votes for the Socialists in a ten-county region of southern Illinois in 1912. The variables in the regression model are y, Socialist vote/total males, and x, number of coal miners/total males. Note that there were virtually no female miners at this time. The number of Socialist votes in each county consisted of the number of votes cast in each county for the Socialist party and Socialist Labor party candidates for the office of State Treasurer. (These data were taken from the *Blue Book of the State of Illinois*, 1912; the State of Illinois Department of Mines and Minerals' *Annual Coal Report*, 1912; and *The relationship between Radicalism and Ethnicity in Southern Illinois Coalfields*, 1870-1940, see references)

Table 2: Ecological regression results using ordinary least squares and robust regression

Illinois County	Socialist Vote Total Males	Number of Coal Mines Total Males	Residual	Weight
Christian	.0367	.1985	−.0122	.71
Franklin	.0485	.5092	−.0043	1.00
Macoupin	.0561	.3085	.0058	1.00
Madison	.0587	.1457	.0104	.83
Montgomery	.0448	.2492	−.0048	1.00
Perry	.0413	.4341	−.0106	.82
Saline	.0789	.5398	.0257	.34
Sangamon	.0352	.2603	−.0145	.60
St. Clair	.0651	.1471	.0168	.51
Williamson	.0578	.6370	.0034	1.00

Notes: least squares estimates: (1) y-intercept$=.0452 = r$, slope $=.0206 = p-r, p = .0658$ (2) robust regression results using Huber's ψ-function ($c = .8$): y-intercept $= .0465 = r$, slope $= .0125 = p - r, p = .059$.

The scatter plot of y versus x is increasing. We are, of course, not prevented from determining a p and any r even if all of the assumptions are not met. In fact, least squares regression gives estimates of $p = .0658$ and $r = .0452$. From these estimates we would infer that the rate of Socialist support among coal miners was greater than among non-miners in this ten-county area. As yet however, no assumptions have been checked.

We then performed a robust (outlier resistant) regression on the y and x data from Table 2. (ecological regression and robust regression are described in Booth and Booth (1988)). The purpose of robust regression is to discount the effect of outlying observations (which pull least squares-based estimates toward them) and thus generate better estimates than least squares for data sets that contain outlying observations. In addition, the discounting procedure provides a weight value for each observation that helps identify outlying observations (i.e., nonrepresentative sample points). The method used here (based on the program cited in Booth and Booth (1988)) gives weight values for each observation between 0 and 1. The smaller the weight, the more a particular observation is an outlier (i.e., the more nonrepresentative of the rest of the data). The robust regression weights reported in Table 2 clearly show that several counties are not representative of the rest of the data (i.e., are outliers). Further, after applying robust regression to these data, the values of the estimates of p and r both change . In fact, we can elaborate further.

Table 3: Method of bounds estimates of p.

County	Maximum p	County	Maximum p
Christian	0.1849	Perry	0.0951
Franklin	0.0952	Saline	0.1462
Macoupin	0.1818	Sangamon	0.1352
Madison	0.3948	St.Clair	0.4425
Montgomery	0.1798	Williamson	0.0907

For those counties that are outliers (e.g., county seven, Saline County) and that have a negative residual sign (i.e. , $y - \hat{y} < 0$), we can conclude that the socialist vote in the county was less than would have been predicted based on the number of coal miners located in that county. A similar but opposite result holds for positive residuals. In addition, *independent qualitative historical research suggests nonconstant values for p and r over this grouping.* (Booth 1983) We thus are forced to conclude that it is unlikely that p and r can be reliably estimated from Table 2 data. Using ecological regression as we have seen, however, we can determine by means of robust regression which counties have more or less socialist support than would be expected based on the proportion of coal miners in the county, using the fitted regression equation as the measure of an "average county." We conclude that the ecological regression constancy assumption is not valid and thus the p and r estimates are suspect.

Second, we consider the method of bounds, using equations (2.1) and (3.1). We know that p must be between zero and one inclusively.

Using the data from Table 2 gives the results reported in Table 3.

As we can see from Table 3, we get a wide disparity in maximum values of p. We can safely conclude from Table 3 is that $p < 0.5$ for each county. We can, however, conclude more. In particular, we can conclude that the evidence again suggests that the constancy assumption of ecological regression is not valid. Again we have information that suggests the p and r estimates from ecological regression are unlikely to be very good. Unfortunately because the bounds are so wide we have yet to answer our question precisely about the radicalism of southern Illinois coal miners.

Thirdly, we now consider the use of the neighborhood model with the Table 2 data. We perform the calculations with equations (4.1) and (4.2). The results are given in Table 4.

Table 4: Neighborhood model estimates of p and r

County	p	r
Christian	0.0073	0.0294
*Franklin	0.0247	0.0238
Macoupin	0.0173	0.0388
Madison	0.0086	0.0501
Montgomery	0.0111	0.0336
Perry	0.0179	0.0234
*Saline	0.0426	0.0363
Sangamon	0.0092	0.0260
St. Clair	0.0096	0.0555
*Williamson	0.0368	0.0210

Let us compare the Table 2 results with the Table 4 results. In Table 2 the robust outlier resistant regression low weights all come from counties that have low p value in Table 4.

From Table 4, we see that with the exception of Franklin, Saline and Williamson counties, the p values are less than the r values. These results would seem to indicate that southern Illinois coal miners were not particularly strong supporters of the socialist candidates in 1912. Further, these results, as well as the method of bounds results of Table 3, indicate that the neighborhood assumption is more likely to be correct than the constancy assumption and hence the Table 4 estimates are probably better than the other two sets of methods.

7. Conclusion

It is clear that the assumptions of the different methods are critical in using ecological inference methods. Based on our analysis it appears that the proportion of coal miners that are also socialist voters (p) is not likely constant over the counties considered. Thus, it appears that the method of bounds (Table 3) and the neighborhood model (Table 4) provide the best estimates of p. Both suggest that the proportion is much smaller than most historians have thought in the past, even though the method of bounds estimates are, in places, somewhat wide. This, and especially Table 4, suggests that explanations other than the coal miners being politically radical are required to explain the mine war.

Acknowledgement

We thank a reviewer whose comments greatly improved the exposition.

References

Booth, D. E. and Booth, S. E. (1988). An introduction to ecological and robust regression in historical research. *Historical Methods* **21**, 35-44.

Booth, S. E. (1983). *The relationship between radicalism and ethnicity in southern Illinois coal fields, 1870-1940.* Doctor of Arts dissertation, Illinois State University.

Duncan, O. D. And Davis, B. (1953). An alternative to ecological correlations. *American Sociological Review* **18**, 665-666.

Freedman, D. A. (1999). Ecological inference and the ecological fallacy. Technical Report No. 549, Department of Statistics, University of California, Berkeley.

Freedman, D. A., Klein, S. P., Ostland, M., and Roberts, M. (1998). On "solutions" to the ecological inference problem. Technical Report No. 515, Department of Statistics, University of California, Berkeley.

Goodman, L. A. (1959). Some alternatives to ecological correlation. *American Journal Of Sociology* **64**, 610-25.

King, G. (1997). *A solution to the ecological inference problem.* Princeton University Press.

State of Illinois (1912). *Blue Book of the State of Illinois, (1912).* State of Illinois. Bureau of Labor Statistics, Department of Mines and Minerals, 1912, Annual coal report.

Received August 22, 2003; accepted February 27, 2004.

Stephane E. Booth
Office of the Provost
Kent State University
sbooth@kent.edu

David E. Booth
Department of Management and Information Systems
Kent State University
Kent, OH 44242, USA
dbooth@bsa3.kent.edu