UCLA

Department of Statistics Papers

Title

Statistical Difficulties in Determining the Role of Race in Capital Cases: A Re-analysis of Data from the State of Maryland

Permalink

<https://escholarship.org/uc/item/4x91087c>

Authors

Richard A. Berk
Azusa Li

Publication Date

2011-10-25

Statistical Difficulties in Determining the Role of Race in Capital Cases: A Re-analysis of Data from the State of Maryland*

Richard Berk
Azusa Li
Department of Statistics, UCLA

Laura J. Hickman
RAND Corporation

July 26, 2004

*Raymond Paternoster provided us with the data on which this paper is based and made a number of very helpful suggestions on a earlier draft of this paper. We very much appreciate both.

Abstract

Statistical Difficulties in Determining the Role of Race in Capital Cases: A Re-analysis of Data from the State of Maryland

In this paper, we re-analyze data used to study the role of race in capital cases in the state of Maryland. We show that when alternative, and arguably more appropriate, statistical procedures are applied, the racial effects reported in early work turn out to be very fragile. The methodological point is more general: conventional causal modeling with observational data is not likely to produce robust results for a variety of criminal justice applications.

Keywords: race, death penalty, logistic regression, propensity scores, classification and regression trees, random forests

1 Introduction

Capital punishment in the United States remains highly controversial. Among the expressed concerns is that the application of death may be racially biased. There is little doubt that African Americans are over-represented on death rows across the country. In 2002, for example, Blacks made up 47 percent of those in state prison under a death sentence (Bonczar and Snell, 2003). Yet, according to the U.S. Census Bureau, African Americans made up 13 percent of the U.S. population in that year. The question is to what extent such disparities reflects racial bias.

Researchers have attempted to tackle this question over several decades, using increasingly complex statistical procedures. Kleck's (1981) review of early death penalty research found that most studies did not control for legally relevant case characteristics. Any findings about the role of race were, therefore, highly suspect. Much subsequent research has employed a variety of control variables. Typically a defendant's race has been found to be unrelated to charging or sentencing decisions (e.g., Bowers and Pierce, 1980; Radelet, 1981; Jacoby and Paternoster, 1982; Gross and Mauro, 1984; Paternoster, 1984; Barnett, 1985; Klein and Rolph, 1991; Morton and Rolph, 2000).

However, many studies have found race-of-victim effects. For example, the General Accounting Office (1990) found that a victim's race was a significant factor in case processing decisions for twenty-three of the twenty-eight studies it reviewed. The relationship was particularly strong at the prosecutorial stage (charging and plea bargaining decisions) when there is arguably more discretion. When crime victims are white, defendants are more likely to be charged with capital crimes, convicted, and sentenced to death than when the victims are black (e.g. Baldus, Pulaski, and Woodworth, 1983; Bowers, 1983; Paternoster, 1984; Radelet and Pierce, 1985; Kazyaka, 1988; Bienen et al., 1988; Gross and Mauro, 1989; Keil and Vito, 1990). However, not all studies agree that victim race is an important predictor of sentencing outcome (see Klein and Rolph, 1991; Morton and Rolph, 2000).

While the body of work focusing on the application of the death penalty has steadily grown, the credibility of these results depends on an appropriate set of included covariates and a convincing use of statistical adjustments. Despite the best efforts of death penalty researchers, there have long been concerns about "omitted variables." For example, we could find no studies that adequately measured witness credibility even though it is surely impor-

tant to charging and sentencing decisions (Klein, 1991). Accurate measurement of included covariates can also be a challenge. For instance, predictors measured when a defendant is arrested (e.g., relationship to the victim) can easily be different if measured when evidence is presented to a jury; comparisons between arrest reports and pre-sentence investigation reports can be very unsettling. But even if the data are rich and well measured, there will usually be grounds for concern. Most studies treat predictors in a linear and additive fashion with little justification. And, it is rare to find the appropriate application of regression diagnostics to help assess model quality. In short, it is extremely difficult to determine the role of race in capital cases. It is not surprising that compelling results are hard to find.

In this paper, we focus on the statistical procedures commonly used in death penalty research. As an illustration, we consider a set of statistical models used in a recent analysis of the death penalty in Maryland. In re-analyzing these data, we bring to bare a number of statistical tools, some of which are quite new, to evaluate the credibility of results previously reported. The lessons drawn from this reanalysis are not just for research on the death penalty, but for criminal justice research more generally when racial disparity is a central issue.

2 Race the the Death Penalty in Maryland

The Maryland study we will reanalyze is titled *An Empirical Analysis of Maryland's Death Sentence System with Respect to the Influence of Race and Legal Jurisdiction* (Paternoster and Brame, 2003). Hereafter referred to as the Maryland report, it was commissioned in September of 2000 by then Maryland Governor Parris N. Glendening. The research goal was to examine the role of race and geography (as measured by county or city) at four stages of the death penalty process:

1. prosecutor decision to formally seek a death penalty;
2. prosecutor decision not to withdraw the “death notification” once filed;
3. prosecutor decision after conviction to undertake a penalty trial; and
4. judge or jury decision to sentence the defendant to death.

The data used in the Maryland report include all “death eligible” homicides in state between July 1, 1978 and December 31, 1999. Death eligibility criteria required a first degree murder offense; a defendant who was the principal in the offense, of sufficient age, not “mentally retarded” at the time of the offense; and a prosecution filing with the court notice of intent to seek the death penalty. In addition, the presence of at least one statutorially-defined aggravating factor is required. (For a full description, see Paternoster and Brame, 2003: 16)

The statistical population includes a total of 1,311 death-eligible cases from a total of about 6,000 homicides. Of these death-eligible cases, 27% were charged with a capital crime. For 40% of those so charged, the decision to seek the death penalty was later withdrawn, typically as the result of a plea bargain. Of the defendants for which the decision to seek the death penalty was not reversed, 84% advanced to the penalty phase. And of these, a death sentence was handed down for 42%. Thus, about 6% of all death eligible defendants eventually were sentenced to death, and of the cases for which the initial decision to seek the death penalty was not reversed, about 36% of the death eligible defendants were eventually sentenced to death.

The multivariate analyses undertaken for the report show both racial effects and county effects, particularly at the early stages of the process. The racial findings are primarily for the race of the victim; defendants with white victims were more likely to be charged with a capital crime and more likely to be sentenced to death, especially if the defendant is black. The findings for location (city or county) showed important variability in the chances of a capital charge and a death sentence. The reported race and geography effects are statistically and substantively significant.

At the same time, the authors readily acknowledge a number of problems with their work. In addition to the usual difficulties obtaining data on requisite covariates, there are substantial complications from missing data and, for the later stages of the process, small sample sizes.

3 Some Broad Modeling Issues

In this reanalysis, we will focus primarily on the modeling efforts. We address three broad questions. First, how appropriate are the statistical models employed in the Maryland Report? Second, how would the findings be affected with a different choice of statistical methods? Third, what can one

ultimately conclude about the role of race in death penalty case processing in Maryland?

Before turning to the details, there are some general points to be made about the analyses on which the report was based. (See Berk, 2003, for the background.)

1. The models reported are presented as causal models. Yet, they really not causal models. In particular, they violate the requirement that each predictor can be manipulated independently of all other predictors (Heckman, 1999; Freedman, 2001). Race is not manipulable, and it is difficult to imagine how the various aggravators and mitigators could be independently altered. This is not just an academic point. It goes to the heart of how to interpret the empirical results. It is a stretch to ask what would happened if a victim's race had been different when it could not possibly be different. The counterfactual needs to be plausible.
2. Any attempt to model the steps after the initial decision to level a capital charge risks well-known sample selection biases because of selection on the response variable. Observations are systematically lost the deeper in the system a case goes. This is not a new issue (Klepper et al., 1983), but it is unaddressed in the report. A reasonable decision, however, might have been to live with the selection biases because the statistical cure can be worse than the statistical disease (Kennedy, 1998: section 16.3; Greene, 2003: 289). In effect, this is what is done in the Maryland analysis.
3. The data are a population or perhaps a convenience sample. In the first case, formal statistical inference is irrelevant (there is no sampling error), and in the second case misleading (the data are in fact not a probability sample). If the Maryland report's justification for statistical inference is model-based (Thompson, 2002: section 2.7), that case is not made. As a result, the p-values on which so much of the analysis in rest the Maryland report have no clear interpretation. For example, a model-based approach necessarily assumes a conditional binomial process in which each defendant's outcome is independent of all others. Given clustering of cases within judges and prosecutors, and given the sequential nature of how cases are processed (earlier cases help inform how later cases are handled), independence is not plausible.

4. The models reported are the result of ad hoc winnowing based on a large number of t-tests. For example, “insignificant” predictors are dropped and “significant” predictors retained. This is a risky strategy, and there are alternative model selection procedures (Cook and Weisberg, 1999, chapter 11). One consequence, is that the meaning of the p-values for each of the models is further muddled.
5. The set of models initially considered are dominated by “main effects.” Interaction effects are important for the racial variables, but not elsewhere. We will show later that there are promising alternatives.
6. If regression diagnostics were used, they were not reported. We will show later that this matters too. (See, for example, Cook and Weisberg, 1999, Chapter 22)

We turn now to our reanalysis of the Maryland data. We will at least touch on each of these six issues and others that are raised along the way. In the interest of space, however, some of them will be covered in far more depth than others.

4 Replication of Key Results

In this reanalysis, we will focus on two outcomes: the initial decision to seek the death penalty and the final outcome of a death penalty or no death penalty. The two response variables represent the extremes that capture the beginning and ending of the legal processes analyzed in the report. In both cases, we will work with the full sample insofar as that is possible, which eliminates the usual problems with sample selection bias. In effect, we will be working with reduced form equations. We can consider the issues most directly in this context, while concentrating on where the data analysis is the strongest.

We begin with Table 1, which shows our replication of the Maryland report’s “final model” of the capital charge. The response was coded so that “1” equals a capital charge and “0” equals no capital charge. The equations contain a rich set of predictors consistent with past research. The racial variables are defined as interactions: they are 1-0 indicator variables for:

1. white defendant and white victim;

2. black defendant black victim;
3. other combinations; and
4. black defendant and white victim as the reference category.

From Table 1, one can see that all of the factors that legitimately should be related to charging are, and their signs generally make sense. Therefore, the results would seem to have some genuine face validity. It is unclear, however, why defendants who took a long time to complete the homicide or whose victims were his or her children (or grandchildren) should be *less* likely to have a capital charge. But some anomalies are always to be expected.

All three of the included racial variables have the effect of reducing substantially the estimated odds of a capital charge with the other predictors held constant. The odds multipliers range from about .37 to about .48 (after exponentiating the regression coefficients). Put more simply, for cases with a black defendant and a white victim, the odds of a capital charge are about 2.5 times greater than for any other racial combination. The differences between each of the other racial variables are probably not very important.

Finally, there are large geographical effects. For example, the estimated odds of a capital charge are substantially lower in Baltimore City than almost everywhere else, other things equal. County/city level variation is to be expected, in part because local prosecutors are elected officials chosen by voters whose preferences across locales can differ.

Table 2 shows the comparable model for the sentencing outcome, with a death sentence coded “1” and no death sentence coded “0.” The overall story is much the same, with 1) virtually all of the signs for legitimate variables in a sensible direction, 2) even stronger racial effects, but 3) slightly smaller geographical effects. If anything, these results appear to be more convincing than for a capital charge.

In summary, using the same data and methods, we were able to reproduce the strong race and county effects reported in the Maryland Report. In the next section, we present additional analyses which assess the quality of the models used to derive these results.

5 Some Regression Diagnostics

Table 3 show the “Classification Table” for the model of charging decisions. The rows contain the actually outcomes, and the middle two columns contain

Predictor	Coeff	SE	Odds Multiplier
Intercept	0.34	0.28	—
Number of prior violent felony convictions	0.16	0.09	1.17
Multiple victim case	0.75	0.22	2.12
Any of the victims a stranger to defendant	0.27	0.18	1.31
Def has a history of alcohol abuse	-0.37	0.19	0.69
Def mental/emotional problems	0.80	0.24	2.23
Def admitted crime	-0.29	0.21	0.75
Def actively evaded arrest	0.46	0.26	1.59
Def implicated in other killing(s)	0.76	0.37	2.13
Def persisted in attack even after death was certain	0.80	0.24	2.22
Any of the vics forced to beg/plead for their lives	0.66	0.32	1.93
Any of the vics murder took a long time to complete	-0.54	0.30	0.58
Any of the vics has children or grandchildren	-0.41	0.20	0.67
Def made full confession to aggravating circumstances	0.73	0.26	2.07
County = Anne Arundel	-1.37	0.35	0.25
County = Baltimore City	-2.92	0.29	0.05
County = Baltimore County	0.37	0.26	1.44
County = Harford	-0.10	0.43	0.90
County = Montgomery	-1.72	0.44	0.18
County = Prince Georges	-0.15	0.26	0.86
Other Counties (Reference Category)	—	—	—
Race = White Defendant—White Victim	-0.75	0.24	0.47
Race = Black Defendant—Black Victim	-0.96	0.24	0.38
Race = Other Combinations	-1.01	0.34	0.36
Black Defendant—White Victim (Reference Category)	—	—	—

Table 1: Logistic Regression Model for Capital Charges (N=1061, AIC=946)

Predictor	Coeff	SE	Odds Multiplier
Intercept	-4.24	0.59	—
Number of prior violent felony convictions	0.51	0.13	1.67
Multiple victim case	1.58	0.36	4.86
Any of the victims a stranger to defendant	0.47	0.34	1.60
Def was sexually abused as a child	1.87	0.48	6.47
Def implicated in other killing(s)	1.34	0.50	3.83
Def forced way into place of victims	0.99	0.33	2.70
Def tried to hide or dispose of bodies	1.05	0.39	2.87
Any of the victims sexually abused after killing	0.03	0.82	1.03
Def confessed to first-degree murder to police	-0.80	0.48	0.45
Def confessed to aggravating circumstances	1.48	0.43	4.39
One eyewitness to the event testified	0.87	0.33	2.37
An informant testified against defendant	0.96	0.40	2.61
County = Anne Arundel	0.47	0.64	1.61
County = Baltimore City	-1.37	0.60	0.25
County = Baltimore County	1.11	0.44	3.04
County = Harford	0.68	0.77	1.97
County = Montgomery	-1.36	0.97	0.26
County = Prince Georges	-0.50	0.59	0.60
Other Counties (Reference Category)	—	—	—
Race = White Defendant—White Victim	-0.93	0.41	0.40
Race = Black Defendant—Black Victim	-1.32	0.45	0.27
Race = Other Combinations	-2.43	1.08	0.09
Black Defendant—White Victim (Reference Category)	—	—	—

Table 2: Logistic Regression Model for Death Sentences (N=1061, AIC=376)

the outcome predicted by the model. The number “1” stands for a capital charge and the number “0” stands for no capital charge. A predicted probability greater than .50 translates into a prediction of a capital charge. A predicted probability equal to or less than .50 translates into a prediction of no capital charge.

Relying on the marginal distribution of the response variable and no model whatsoever, one can simply classified by the modal charging category (i.e., no capital charge). Then, one would be correct nearly 70% of the time. From Table 3, one can see that with the model correctly classifies cases about 79% of the time. This is a modest improvement. Moreover, about 40% of the cases with a capital charge are misclassified as are about 12% of the cases with no capital charge. The model is far more likely misclassify cases with a capital charge than cases without a capital charge.

	0 (Pred)	1 (Pred)	Misclassification percentage
0 (Actual)	646	92	0.12
1 (Actual)	128	195	0.40

Table 3: Classification Table for Logistic Regression Estimating Effects of Defendant and Victim Race on Capital Charges (23 Predictors)

Figure 1 provides insight into the source of this misclassification. It is a scatter plot with two lines overlaid. The vertical axis of the plot is the response variable in 1/0 indicator variable form. The horizontal axis of the plot is the linear combination of predictors used in the logistic regression analysis in their logit units. The blue line shows a lowess smooth between the linear combination of predictors and the binary response variable. The red line shows the actual fitted values from the logistic regression against the linear combination of predictors.¹

In a satisfactory model, the two lines should fall on top of one another or at least be quite close. In this plot, however, the two lines are very different. The blue lowess smooth is far more steep than the red line for the fitted values from the model. A key implication is that the model does not distinguish as well as it should between cases in which a capital charge is likely and cases

¹There are in fact no observations for the response variable greater than 1 or less than 0. The appearance of such values in the plot is an artifact of the plotting algorithm.

Figure 1: Model Checking Plot for Capital Charges

in which it is not. In other words, the model depicts charging decisions as more ambiguous than they really are.

One possible explanation is that some important predictors that can more effectively distinguish between capital and non-capital cases are missing from the model. Because it is almost certain that these omitted variables are correlated with the predictors in the model, the estimated regression coefficients for the included predictors are probably capturing in part effects that more properly should be attributed to other factors.

Another possible explanation is that the functional forms used are incorrect. Either the conventional logit link function is substantially wrong, or predictor transformations are needed. Because most of the predictors are indicator variables, any problems with functional form are most likely caused by the logit link function. Once again, the model represents prosecutor charging decisions as more ambiguous than they really are. Consequently, the estimated regression coefficients may reflect far more (or less) than the role of their associated predictors.

	0 (Pred)	1 (Pred)	Misclassification percentage
0 (Actual)	976	14	0.01
1 (Actual)	55	16	0.77

Table 4: Classification Table for Logistic Regression Estimating Effects of Defendant and Victim Race on Death Sentences (22 predictors)

Table 4 is the classification table for the regression model of whether a death sentence is imposed (i.e., Table 2) For this outcome, the logistic regression model does not do well at all. With such a skewed response variable, it is very easy to classify cases with considerable accuracy even if no model whatsoever is employed. If one simply classified all cases a priori as not receiving a death sentence, this classification would be correct over 93% of the time. Table 4 shows that classifications based on the logistic regression model are correct 94% of the time. Clearly, the model is not help much. Moreover, what success the model has is explained by predicting that a given defendant will not receive the death penalty. These classifications are correct about 99% of the time. The model is correctly classifying death sentence cases about 23% of the time.

Figure 2 shows the diagnostic plot for Table 2. The red line shows the

Figure 2: Model Checking Plot for Death Sentences

values predicted by the logistic regression model. Only a few have a predicted probability greater than .5, and none have a predicted probability greater than .65. The logistic regression model is not doing a very good job in classifying death sentence cases correctly. The jagged blue line is the smooth of the model free link between the death sentence outcome and the linear combination of predictors. Because there are so few death sentences, the usual window size produces a flat, straight line falling on top of the response value of 0 (i.e., no death sentence). But smaller window sizes either lead to the same result, or if made sufficiently small, produce the jagged line shown. Still, the message is clear: for values of the linear predictors greater than about -1, the logistic regression model grossly underestimates the chances of a death sentence. At the same time, the statistical challenge is great because the response is so unbalanced. Conventional smoothing procedures either show no relationship or one that is very rough.

The classification tables and model checking plots clearly demonstrate that the apparently plausible results from the logistic regression are actually somewhat suspect. But what does that say in particular about the race and county effects?

6 More Diagnostics Focusing on Race

It is common in observational studies of program impacts to use propensity scores to adjust for biases resulting from non-random selection into the various treatment groups (Rosenbaum and Rubin, 1983; Rosenbaum, 2002, Chapter 10). Propensity scores are the estimated probability of membership in each of the treatment groups (or in the most simple case, experimental and control group). If one really thinks that the death penalty regression equations with racial predictors are causal models, one can view race as a “treatment.” (See Berk, 2003 for a discussion of race as a causal variable.) It follows that one can construct propensity scores for any racial variable. Propensity scores can then be used, just in observational studies of program impacts, to control for confounded variables.

When propensity scores are computed, there is a separate equation (typically a logistic regression equation) employed that can have different predictors and different functional forms from the equation for the outcome of interest. This allows for greater flexibility than provided by conventional regression adjustments which in turn, more effectively reduces the biases

resulting from confounding.

In death penalty research, there is no theory or empirical work specifying what the selection equation(s) for racial predictors should be.² Therefore, we proceeded inductively using classification and regression trees (Breiman et al., 1984) and all predictors included in the earlier logistic regressions to compute the requisite propensity scores. The added flexibility provided by classification and regression trees (CART) further improves the prospects for reducing biases in the estimated racial effects.

With CART, however, there is always a risk of serious over-fitting. Ensemble methods that build on CART will generally perform better and not be subject to over-fitting (Breiman, 2001). Consequently, we undertook the propensity score analysis again, but with random forests. Although the propensity score analysis based on CART was similar to the propensity score analysis based on random forests, we report the latter results as the more credible. (For an accessible introduction to ensemble methods see Berk, 2004.)

Predictor	Coefficient	Std. Error	Multiplier
Intercept	-2.31	0.24	—
White Defendant–White Victim	-0.35	0.24	0.71
Black Defendant–Black Victim	-0.55	0.26	0.58
Other Combinations	-0.66	0.35	0.52
Propensity Score White Defendant–White Victim	0.05	1.44	1.05
Propensity Score Black Defendant–Black Victim	1.96	1.07	7.08
Propensity Score Other Combinations	3.47	1.61	32.23

Table 5: Propensity Scores Analysis of Capital Charges Based on Random Forests (AIC= 889.8, N=1061)

Table 5 shows the results for the decision to seek a death sentence with the propensity scores used as the only predictors besides race. In this instance,

²And this makes sense. A person does not have his or her race “assigned.” At the same time, this underscores the problem mentioned earlier of treating race as a causal variable.

a comparison between the fit statistics is not especially helpful for deciding which model provides a more accurate picture of the role of race.³ Still, the model using propensity scores fits the data about as well as the logistic regression model (even a bit better), once one takes model complexity into account. The model represented in Table 5 has an AIC of 890 compared to an AIC of 946 for the model represented in Table 1. The classification tables are comparable as well.

These results imply that a large fraction of explanatory power of the original logistic regression for charging decisions can be attributed to predictors that are associated with race; what is being held constant really matters if racial effects are to be properly isolated. With the more simple model using propensity scores, which likely does a better job at removing confounding, the racial effects are reduced substantially. Even with no new covariates and even with measurement error in the covariates we have, it is possible to adjust better for confounding, and the racial effects are substantially attenuated. Moreover, if we use the .05 level of statistical significance and take the tests seriously, the null hypothesis is rejected only for the coefficient for black defendant and black victim is the null hypothesis rejected.⁴

We undertook the same exercise for the response variable of a death sentence. Table 6 shows the results. With model complexity taken into account, the model using propensity score adjustments fits about the same as the earlier “full” model. The model represented in Table 6 has an AIC of 431 compared to an AIC of 376 for the model represented in Table 2. Once again, the racial effects are dramatically reduced, and if statistical inference is to be taken seriously, only the black defendant and black victim regression coefficient is more than twice its standard error. As before, even with no new covariates, stronger covariances adjustments reduce any racial effects.⁵

³With observational data, a large regression model producing a good fit to the data will often have seriously biased regression coefficients. In contrast, a very simple regression model for a properly implemented randomized experiment may not fit the data very well, but is far more likely to produce unbiased estimates. The key is that on the average the randomized experiment eliminates confounding. Just as in randomized experiments, adjustments in observational studies using propensity scores are not meant to produce a close fit to the data. The goal is to eliminate confounding. Whether the fit is good is not usually an important issue.

⁴If we discount the p-value for multiple tests, the p-value would be far larger than .05.

⁵Because of the highly unbalanced response variable, the results for the propensity scores variables in Table 6 are unstable. The problem is that for many values of the predictors, the number of death sentences cases is very small or even zero. This can lead

Predictor	Coefficient	Std. Error	Multiplier
Intercept	-2.64	0.25	—
White Defendant–White Victim	-0.31	0.32	0.73
Black Defendant–Black Victim	-1.19	0.37	0.30
Other Combinations	-1.87	1.03	0.15
Propensity Score White Defendant–White Victim	9.54	4.58	13904.95
Propensity Score Black Defendant–Black Victim	-2.44	4.54	0.09
Propensity Score Other Combinations	-1.27	4.85	0.28

Table 6: Propensity Scores Analysis Death Sentences based Random Forests (AIC= 431.8, N=1061)

One potential weakness with the two propensity score analyses is that the functional forms used for the propensity scores are assumed to be linear. In principle, a more powerful set of adjustment could be obtained by allowing the data to determine the functional form. We turned to the generalized additive model and B-spline smoothers for the propensity score covariates (Hastie and Tibshirani, 1990).

From the estimated degrees of freedom, it was apparent that the linear form is probably appropriate for two of the three propensity scores. And in fact, there were no important changes when the nonparametric regression was applied. The conclusions were also effectively the same for the death sentence response variable. In short, whatever problems there are with our propensity score covariance corrections, they do not stem from the functional forms used for the predictors. The problems are likely caused by omitted variables and measurement error in those predictors.

to unstable estimates for a wide variety of estimation procedures (Hastie and Tibshirani, 1990, section 6.10). Note the enormous odds multiplier for the predicted probability of a case having a white defendant and a white victim.

7 Competing Models with CART

Classification trees provide a more flexible way than logistic regression to model directly both the decision to level a capital charge and the final outcome of a death sentence. For us the key question is what roles race and county might play. However, it is important to emphasize that CART models cannot usually be interpreted like regression models. CART models are constructed in a stagewise fashion so that associations represented by splits and any given level in the tree are unadjusted for variables used afterwards.

Figure 3 coupled with the legend in Table 7 shows the CART results for capital charges drawing on the same predictors used in the logistic regression analysis.⁶ Location is clearly an important predictor, providing the first optimal split of the data. The chances of a capital charge are less for Baltimore City, Baltimore County, and Prince George’s County. Note that Baltimore County stood out the place with the *greatest* chance of a capital charge in the logistic regression model. Within these three areas, which contain a bit more than half of the cases, racial indicators are selected for the second split. When either the defendant or the victim (or both) is non-white, about 6% of the defendants receive a capital charge. No other predictors alter this percentage in a meaningful way. When both the defendant and victim are white, approximately 14% the defendants who murdered a single individual receive a capital charge. But nearly 90% of such defendants receive a capital charge if there are multiple victims. However, because the number of multiple victim cases with white defendants and white victims is very small, the 90% figure is not very firm. Note also that whatever the value of these findings, they apply only to Baltimore City, Baltimore County and Prince George’s County, not more generally. But perhaps the major point is that while the logistic regression model singles out cases with a black defendant and a white victim

⁶The split at each break point indicates which cases go to the left, the direction of a lower likelihood of a capital charge. For binary predictors, the split is always at .5. The top number associated with each terminal node indicates how the node was defined (1 = a capital charge, 0 otherwise) using the majority vote criterion. The bottom two numbers (from left to right) show the number of actual non-capital charges and the number of capital charges. The proportion of capital charges generally increases for each terminal nodes moving from left to right. The analysis was done with `rpart` in R, using all the default settings with one exception. The complexity parameter was set so that there would be about 20 terminal nodes so that the degrees of freedom in the CART model would be about the same as the comparable logistic regression model. Pushing the data any harder would have led to many terminal nodes with very few cases.

Figure 3: CART Results for Capital Charges

pvfelony*	Number of prior violent felony convictions
multvic	Multiple victim case
vstrang	Any of the victims a stranger to defendant
alchist*	Defendant has a history of alcohol abuse
afac4*	Defendant unable to control conduct due to mental/emotional problems
afac16	Defendant admitted crime
afac31*	Defendant actively evaded arrest
afac34	Defendant implicated in other killing(s)
afac41	Defendant persisted in attack even after death was certain
pc5*	Any of the victims forced to beg/plead for their lives
pc25	Any of the victims murder took a long time to complete
pc37	Any of the victims has children or grandchildren
se3	Defendant made full confession to aggravating circumstances
race*	a = White Defendant–White Victim b = Black Defendant–Black victim c = Other Combinations d = Black Defendant–White Victim (Reference Category)
counties*	a = Anne Arundel b = Baltimore City c = Baltimore County d = Harford e = Montgomery f = Prince Georges g = Other Counties (Reference Category)

Table 7: Legend for CART Analysis of Capital Charges

as most likely by far to receive a capital charge, the CART model singles out cases with white defendants and white victims.

By and large, the direction of the other splits in Figure 3 make sense. There is also some evidence for additional racial effects (near the bottom middle of the tree), but these are based on very small samples and are inconsistent with each other and the earlier racial effects. But if one were to take them seriously, they surface primarily for crimes whose “death worthiness” is not high. For those cases with the most serious aggravating factors, race would seem to play no role.

Table 8 shows the classification table for the CART analysis. One can see that while only about 8% of the true negatives were misclassified, about 44% of the true positives were misclassified. Even with the more flexible approach provided by CART, nearly half of the cases with a capital charge are classified as not having a capital charge. The results in Table 8 are in part a result of over-fitting.

Recall Table 3 which showed the classification table for the logistic regression in Table 1. The results are much the same although there is a bit higher percentage of false positives and a bit lower percentage of false negatives. One implication may be that some of the modeling problems have less to do with the model itself and more to do with the omission of important predictors.

To summarize, the geographic effects for charging decisions come through loud and clear in the logistic regression results and the CART results. Racial effects found using logistic regression are by and large not replicated by the CART model. But this does not imply that meaningful racial effects do not exist. It implies that with these data it is extremely difficult to tell whether meaningful racial effects exist or not.

	0 (Pred)	1 (Pred)	Misclassification percentage
0 (Actual)	802	73	0.08
1 (Actual)	143	184	0.44

Table 8: Classification Table for CART Model of Capital Charges (17 Terminal Nodes)

Figure 4 coupled with Table 7 show the CART results for the imposition

of a death sentence.⁷ In this analysis, counties again produce the first split. Harford County stands apart from the rest as having death eligible homicides most likely to lead to a death sentence. This is not the same county split found for capital charges. The subsequent splits in Figure 4 generally make sense. For example, within Harford County, the two next splits are for the number of previous felonies. With a greater number of previous felonies, the likelihood of a death sentence increases substantially.

Race enters late in the process and toward the middle of the tree. In both instances when race surfaces, cases with a black defendant and white victim or “other” racial combinations are *less* likely to have a death sentence imposed. This contradicts the logistic regression results. Moreover, the cases for which race plays a role represent a small fraction of total — about 5% of all the cases (65 cases). These cases are also the ones for which the outcome is more uncertain. In short, the CART results for race differ greatly from the logistic regression results. Again, this does not mean that there are no racial effects. Rather, it means that one really cannot tell with these data and modeling procedures.

Table 10 is the relevant classification table. CART does a good job in classifying true negatives and a poor job classifying true positives. About two-thirds of the capital cases are incorrectly classified as not having a capital charge. Given the lack of balance in the response variable, the relatively large number of false negatives is not surprising.

Table 4, discussed earlier, is the relevant classification table for the logistic regression analysis reported in Table 2. The logistic regression does somewhat worse than CART; there is even a higher percentage of false positives. Despite over-fitting, over 77% of the true positives are misclassified. If classification accuracy is an important criterion, the CART results are at least as credible as the logistic regression results and lead to somewhat different conclusions about the role of race.

⁷As before, `rpart` in R was used with the default settings, save for the complexity parameter. For comparability to the logistic regression model used for death sentences, we were aiming for about 10 terminal nodes, although with the highly unbalanced response, some terminal nodes with very few cases were anticipated.

Figure 4: CART Results for Death Sentences

pvfelony*	Number of prior violent felony convictions
multvic	Multiple victim case
vstrang	Any of the victims a stranger to defendant
afac9*	Defendant was sexually abused as a child
afac34	Defendant implicated in other killing(s)
pc1*	Defendant forced his/her way into place of murder of any of the victims
pc19	Defendant tried to hide or dispose of bodies of any of the victims
pc53	Any of the victims sexually abused after killing
se1	Defendant made full confession to first-degree murder to police
se3	Defendant made full confession to aggravating circumstances
se7	One eyewitness to the event testified
se13	An informant or jail-house snitch testified against defendant
race*	a = White Defendant–White Victim b = Black Defendant–Black victim c = Other Combinations d = Black Defendant–White Victim (Reference Category)
counties*	a = Anne Arundel b = Baltimore City c = Baltimore County d = Harford e = Montgomery f = Prince Georges g = Other Counties (Reference Category)

Table 9: Legend for CART Analysis of Death Sentences

	0 (Pred)	1 (Pred)	Misclassification percentage
0 (Actual)	1115	16	0.01
1 (Actual)	47	24	0.66

Table 10: Classification Table for CART Model Death Sentences (12 Terminal Nodes)

8 Competing Models with Random Forests

Random forests is an ensemble procedure that builds on CART. It will generally perform better (Breiman, 2001), but, is somewhat harder to interpret. We will use it here to counter the tendency of CART to over-fit the data.

In random forests, a large number of trees is constructed. Each tree is built from a bootstrap sample of the data and at each node, a random sample of predictors is evaluated for the best split. Then, classifications are determined by a majority vote over the set of trees constructed. One key asset of random forests is that by averaging over trees, over-fitting can be effectively addressed. Another key asset is that observations not included in the bootstrap sample (called “out-of-bag” observations) can serve as a test data set in model assessments.

The results from random forest were not dramatically better than the results from CART. As before, the results for death sentences were especially problematic. Once the marginal distribution was taken into account, it was difficult to do much better. The main conclusion is that the earlier CART results are not subject to substantial over-fitting.

8.1 Taking Costs into Account

Implicit in all of the analyses above has been the assumption that the costs of false negatives were the same as the costs of false positives. All of our analyses above treated the two equally. But suppose it is more important to correctly identify cases charged with a capital crime than cases not charged. And suppose it is more important to identify cases that result in a death sentence than those without a death sentence. Using this approach, all of the procedures can be re-evaluated with differential costs.

Unfortunately, we have no insider knowledge of what the different costs should be. It is ultimately a matter for the political process to determine. But it is relatively easy to illustrate the role that costs can play and how one’s conclusions can be affected by differential costs.

For the logistic regression, a .30 classification threshold can be used instead of a .50 classification threshold. Roughly speaking, one false negative is now worth about two false positives. When this is applied to the predicted probabilities from the logistic regression, the proportion of capital charge cases correctly classified increases from .40 to .85. At the same time, the proportion of non-capital charge cases correctly classified decreases from .92

to .72. There are 79 more capital charge cases correctly identified and 114 fewer non-capital charge cases correctly identified. The tradeoff turns out to be about 1.5 to 1.

When the same approach is applied to death sentences, the proportion of death sentence cases correctly classified increases from .33 to .85 while the proportion of non-death sentence cases correctly classified decreases from .99 to .84. There are 54 more death sentence cases correctly identified and 145 fewer non-death sentence cases correctly identified. The trade off turns to be about 2.7 to 1.

The logistic regression model is the same regardless of the decision threshold used because estimation does not take classification errors explicitly into account. If classification and prediction is an important goal, this is a major limitation. For CART and random forests, changing the relative costs of misclassification changes the trees that result. We present the results for random forests because they are the most credible.

Tables 11 and 12 are the classification tables for the analysis of capital charges and death sentences. True prediction errors are shown based on the out-of-bag observations. Because of the highly unbalanced nature of the death sentence response, stratified random sampling was used in the tree construction so that equal numbers were drawn for cases in which there was a death sentence and cases in which there was not. The goal was to compensate for the lack of balance in the response.⁸ For comparability, we employed the same balancing strategy for the capital charges.

For capital charges, this implies that instead of assuming equal costs for false negatives and false positives, failing to identify a single capital charge cases is about the same as failing to identify two non-capital charge cases. For death sentences this implies that instead of assuming equal costs for false negatives and false positives, failing to identify a single death sentence case is about the same failing to identify nine non-death sentence cases. The results from both tables are the best so far when one remembers that these tables are based on real prediction.⁹

But what about role of race? An important drawback to random forests is

⁸At the suggestion of Dr. Andrew Liaw, who ported random forests to R, the sample sizes were taken to be 60% of the number of death sentences. This allows for a sufficient number of out-of-boot observations used to construct the classification votes.

⁹The results for logistic regression superficially look about as good, but the data used to test the model are the same data used to build the model. Thus, the performance of the logistic regression is overstated.

	0 (Pred)	1 (Pred)	Prediction Error
0 (Actual)	575	163	0.22
1 (Actual)	62	261	0.19

Table 11: Classification Table for Random Forest Model of Capital Charges

	0 (Pred)	1 (Pred)	Prediction Error
0 (Actual)	855	135	0.14
1 (Actual)	14	57	0.20

Table 12: Classification Table for Balanced Bootstrap Random Forest Model of Death Sentences

that there is no longer a tree to interpret. And there seems to be no sensible way to extract information about individual predictors from the set of trees.

If the goal is to determine if race should be included in the model at all, the problem is one of model selection. One way to consider if race is needed is to replace each of the racial indicators in turn by a shuffled version in which the 1's and 0's are put in random order. In the shuffled form, the racial indicators should be unrelated to the response and to other predictors. If the racial indicator variables belong in the model, turning each into "noise" should degrade the perform of the fitting process and prediction errors should increase.

There are three ways one can examine the impact of the shuffling. One can look at the increase in prediction error overall, the increase on false positives or the increase in false negatives. With an unbalanced response variable, the increase in overall prediction error will tend to mask increases in less common response category. If these are important, the results could be misleading. So, we will focus on the increase in failures to correctly predict capital charge cases and death sentence cases.

Table13 reports the results of the shuffling for capital charges. The first row shows the proportion of capital charge cases incorrectly predicted when all of the explanatory variables used in the earlier analyses are included. About 19% of these cases are predicted incorrectly (i.e. false negatives). The following rows show how the prediction error increase when each variable

(alone) is shuffled.¹⁰ Precisely how to take chance into account with random forests is at this point unresolved, but examining a number of random forest results for capital charges suggests that increases less than about 2% could easily represent random variation.

Predictor Shuffled	prediction error
None	0.19
White Def - White Vic	0.20
Black Def - Black Vic	0.19
Other Race	0.23
Baltimore City	0.33
Previous Felony	0.19
More than one Victim	0.21

Table 13: Prediction Error for Cases Charged with a Capital Crime

One can see in Table 13 that race does not have much of an impact on prediction accuracy. Recall that the baseline category is black defendants and white victims. From this baseline, false negatives increase a negligible amount for white defendants and white victims and for black defendants and black victims. The increase from .19 to .23 for the “other” race category is substantial, but even if one takes that seriously, it represents a small fraction of the cases and contradicts the logistic regression results. To provide a relative sense of the importance of race, three other illustrative predictors are included. Clearly the indicator for Baltimore city dominates the story and is far more important than any of the race variables. The presence of multiple victims matters too, while having a previous felony does not.

Table 14 show the results for death sentence cases. Now the likely margin of error is perhaps around .03. Race appears to make little difference with the exception again of the “other” race variable. But the role of the other predictors changes somewhat. Having a case brought in Baltimore City is still important, but not as important as having a previous felony and far less than having multiple victims.

¹⁰We also applied the same overall approach but dropped each variable in turn, rather than shuffling it. The results were much the same. The (small) advantage of the shuffling strategy is that the number of explanatory variables does change; the degrees of freedom lost is constant.

Predictor Shuffled	Prediction Error
None	0.20
White Def - White Vic	0.22
Black Def - Black Vic	0.21
Other Race	0.23
Baltimore City	0.25
Previous Felony	0.25
More than one Victim	0.30

Table 14: Prediction Error for Death Sentence Cases

Several conclusions follow. First, the predictive accuracy for capital charges and death sentences is about the same once we allow for differential costs of false positives and false negatives. Second, however, the predictors that weigh in are a bit different. Third, if race is playing a role, it is small and hard to specify. Finally, the location in which a case is brought has a very large impact.

These conclusions are roughly consistent with the CART results but differ importantly from the results for the logistic regressions. Part of the explanation is that CART and random forests arrived at very different models than used in the logistic regressions. A key departure was allowing for different predictors in different locales. Another part of the difference is that CART and random forests use prediction error as a way to arrive at a final model. Logistic regression maximizes the likelihood function, which essentially maximizes the fit with the data on hand. Maximum likelihood estimation has a long and venerable history, but if the goal of science is to make predictions and then test them, minimizing prediction error has enormous appeal. Finally, CART and random forests can respond to the relative costs of false negatives and false positives. Logistic regression cannot, and implicitly each case is given equal weight. In this application, equal weighting may be less desirable.

9 Summary and Conclusions

The goal of this work was to examine the role of race in death penalty charging and sentencing when alternative statistical methods are applied. Using

the Maryland data, we found that logistic regression models did not fare especially well if accurate classification is important, and that when stronger adjustments for confounding were employed, the racial effects were substantially diminished. Moreover, the analysis of capital charges using CART showed different racial effects from those found in the logistic regression results, and the analysis of death sentences using CART revealed no important racial effects, or small and unconvincing ones. Finally, the most credible competition for the logistic regression results came from random forests. For both capital charges and death sentences, race either played no role or a small one that is very difficult to specify.¹¹ In short, it is very difficult to find convincing evidence for racial effects in the Maryland data.

The methodological point is not that different statistical methods can lead to different results. The methodological point is that the routine use of conventional causal modeling to understand the role of race in capital cases risks very fragile results. And this lesson generalizes to a wide variety of other applications (Berk, 2003)

There are at least three important caveats. First, with better covariates it is possible that stronger racial effects could be found, not just weaker effects. If, for example, black defendants are more likely to show contrition, including a measure of contrition as a covariate might well show stronger racial effects. Second, if the concern is about illegitimate factors affecting capital cases, the impact of location needs to be studied in much greater depth. As now measured, a county or city is just a proxy for processes that are not analyzed. Finally, perhaps more important than the role of illegitimate factors is pure capriciousness. It is difficult to imagine that a few covariates exist that if included as predictors would lead to clear and justified distinctions between defendants who are charged with capital crime and defendants who are not; likewise for death sentences. As we have observed before (Berk et al., 1993; Weiss et al., 1996), if idiosyncrasies associated with the case, the defendant, or the adjudication process seem to determine a substantial part of the outcome, the adjudication process is suspect whether race is important or not.

¹¹However, random forests is an ensemble procedure that compared to logistic regression requires very different approach to data analysis. A discussion of these differences are beyond the scope of this paper, but a number of accessible references exist (e.g., Breiman, 2001b; Berk, 2004).

10 References

- Baldus, D., Pulaski, C., and G. Woodworth (1983) "Comparative Review of Death Sentences: An Empirical study of the Georgia Experience," *Journal of Criminal Law and Criminology*, 74 (3): 661-753.
- Barnett, A. (1985) "Some Distribution Patterns for the Georgia Death Sentence," *University of California-Davis Law Review*, 18 (4), 1327-1374.
- Berk, R.A., Weiss, R.E., and J. Boger (1993) "Chance and the Death Penalty (with commentary and rejoinder)," *Law & Society Review*, 27: 89-110 1993.
- Berk, R.A. (2004) "Data Mining with a Regression Framework," UCLA Department of Statistics Preprint Series # 371.
- Berk, R.A. (2003) *Regression Analysis: A Constructive Critique*. Newbury Park, CA: Sage Publications.
- Bienen, I., Weiner, N., Denno, D., Allison, P., and D. Mills (1988) "The Reimposition of Capital Punishment in New Jersey: The Role of Prosecutor Discretion," *Rutgers Law Review* 41 (1): 27-372.
- Breiman, L. (2001b) "Statistical Modeling: Two Cultures," (with discussion) *Statistical Science* 16: 199-231.
- Breiman, L. (2001a). "Random Forests." *Machine Learning* 45, 5-32.
- Breiman, L., Friedman, J.H., Olshen, R.A., and C.J. Stone (1984) *Classification and Regression Trees*. Monterey, Ca: Wadsworth
- Bonczar, T. P. and T.L. Snell (November 2003) *Capital Punishment, 2002*. Washington, DC: U.S. Department of Justice, Bureau of Justice Statistics.
- Bowers, W. (1983) "The Pervasiveness of Arbitrary and Discrimination under Post-Furman Capital Statutes," *Crime and Delinquency*, 74 (3): 563-635.
- Bowers, W. J. and G. Pierce (1980) "Arbitrariness and Discrimination under Post-Furman Capital Statutes," *Crime and Delinquency*, 26: 563-572.

- Cook, R.D. and S. Weisberg. (1999) *Applied Regression Including Computing and Graphics*. New York: John Wiley.
- Freedman, D.A. (2001) "On Specifying Graphical Models for Causation, and the Identification Problem." U.C. Berkeley Department of Statistics Preprint Series #601.
- General Accounting Office (1990) *Death Penalty Sentencing: Research Indicates Pattern of Racial Disparities*. Report to the Senate and House Committees on the Judiciary, GGD-90-57. Washington, DC: General Accounting Office.
- Greene, W.H. (2000) *Econometric Analysis*, fourth edition. New York: Prentice Hall.
- Gross, S. R. and R. Mauro (1984) "Patterns of Death: An Analysis of Racial Disparities in Capital Sentencing and Homicide Victimization," *Stanford Law Review*, 37(1): 27-153.
- Gross, S. R. and R. Mauro, R (1989) *Death and Discrimination: Racial Disparities in Capital Sentencing*. Boston, MA: Northeastern University Press.
- Heckman, J. (1999) "Causal Parameters and Policy Analysis in Economics: A Twentieth Century Retrospective," *Quarterly Journal of Economics*, February: 45-97.
- Hastie, R., and R. Tibshirani (1990) *Generalized Additive Models*. New York: Chapman & Hall.
- Hastie, R., Tibshirani, R. and J. Friedman. (2001) *The Elements of Statistical Learning*. New York: Springer-Verlag.
- Jacoby, J. E., and R. Paternoster (1982) "Sentencing Disparity and Jury Packing: Further Challenges to the Death Penalty," *Journal of Criminal Law and Criminology*, 73 (1): 379-387.
- Kennedy, P. (1998) *A Guide to Econometrics*, fourth edition. Cambridge, Massachusetts: MIT Press.

- Klepper, S., Nagin, D., and L. Tierney. (1983) "Discrimination in the Criminal Justice System: A Critical Appraisal of the Literature," in A. Blumstein, J. Cohen, S.E. Martin, and M.H. Tonrey (eds.) *Research on Sentencing: A Search for Reform*, Volume II, Washington,D.C.:National Academy Press.
- Keil, T. J. and G.F. Vito (1990) "Race and the Death Penalty in Kentucky Murder Trials: An Analysis of Post-Gregg Outcomes," *Justice Quarterly*, 7 (1): 189-207.
- Kleck, G. (1981) "Racial Discrimination in Criminal Sentencing: A Critical Evaluation of the Evidence with Additional Evidence on the Death Penalty," *American Sociological Review*, 46 (6): 783-805.
- Klein, S. (1991) "Racial Disparity in Sentencing Decisions." Unpublished Manuscript of Testimony before the Subcommittee on Civil and Constitutional Rights hearing on Race and Arbitrariness in Capital Sentencing, Committee on the Judiciary, U.S. House of Representatives, July 10, 1991. Santa Monica, CA: RAND.
- Klein, S. and J. Rolph (1991) "Relationship of Offender and Victim Race to Death Penalty Sentences in California," *Jurimetrics*, 32, 33-48.
- Morton, S. C. and J.E. Rolph (2000) "Racial Bias in Death Sentencing: Assessing the Statistical Evidence," In S. C. Morton and J. E. Rolph (eds), *Public Policy and Statistics: Case Studies from RAND*, pgs. 94-116. New York, NY: Springer-Verlag.
- Paternoster, R. (1984) "Prosecutorial Discretion in Requesting the Death Penalty: A Case of Victim-Based Racial Discrimination," *Law and Society Review*, 18 (3): 437-478.
- Paternoster, R. and A.M. Kazyaka (1988) "The Administration of the Death Penalty in South Carolina: Experiences over the First Few Years," *South Carolina Law Review*, 39(2): 245-411.
- Paternoster, R., and R. Brame. (2003) *An Empirical Analysis of Maryland's Death Sentence System with Respect to the Influence of Race and Legal Jurisdiction*. Department of Criminology, University of Maryland.

- Radelet, M. (1981) "Racial Characteristics in the Imposition of the Death Penalty," *American Sociological Review*, 46: 918-927.
- Radelet, M. and G. Pierce (1985) "Race and Prosecutorial Discretion in Homicide Cases," *Law and Society Review*, 19 (4): 587-621.
- Rosenbaum, P.R., and D.B. Rubin (1983) "The Central Role of Propensity Score in Observational Studies of Causal Effects." *Biometrika* 70: 41-55.
- Rosenbaum, P. (2002) *Observational Studies*, second edition. New York: Springer-Verlag.
- Thompson, S.K. (2002) *Sampling*, second edition. New York: John Wiley.
- Weiss, R.E., Berk, R.A., and C. Lee (1996) "Assessing the Capriciousness of Death Penalty Charging," *Law & Society Review*, Vol. 30, No. 3, pp. 607-626, 1996.
- Zhang, H., and B. Singer. (1999) *Recursive Partitioning in the Health Sciences*. New York: Springer-Verlag.