Supplementary Material for Alleviating Ecological Bias in Poisson Models using Optimal Subsampling: The Effects of Jim Crow on Black Illiteracy in the Robinson Data

June 24, 2014

Optimal Design

In this section, we use (6) from Appendix B of the paper to present the optimal designs for the estimation of β_{1i} in each state, using both the ecological data and stratified random samples, and also accounting for the uncertainty in the nuisance parameter β_{2i}. Because the Poisson model is non-linear, the optimal design depends on the true values of β_{1i}, β_{2i} and we consider different values of these parameters in Tables 1 and 2 below.

In Table 1 we consider values of β_{1i}, β_{2i} from (2.5,2.5) down to (0.0,0.0) at intervals of 0.5. This is meant to give a sense of the variability in the optimal design that the survey designer would face if they did not know the true parameter values. As the log ratios in illiteracy rates approach 0, the % of the sample foreign-born white approaches the ecological proportions: Indiana 5%, Kansas 5%, Wyoming 13%, Nevada 16%.

In Table 2, we present the optimal designs when β_{1i}, β_{2i} are set to the true parameter values for each of the states. This information would not be available to the survey designer, but

it provides the true optimal design. Notice that for Nevada, the log ratio for foreign-born whites is large enough that illiterate foreign-born whites are fairly common, so the optimal design has flipped in the sense that the optimal design includes native-born whites but not foreign born whites.

	Indiana	Kansas	Wyoming	Nevada
Log Illiteracy Relative Risk of 2.5				
% of Sample Black	69	71	37	32
% of Sample Foreign-Born White	31	29	63	68
Log Illiteracy Relative Risk of 2.0				
% of Sample Black	76	78	49	43
% of Sample Foreign-Born White	24	22	51	57
Log Illiteracy Relative Risk of 1.5				
% of Sample Black	83	84	60	55
% of Sample Foreign-Born White	17	16	40	45
Log Illiteracy Relative Risk of 1.0				
% of Sample Black	88	89	71	67
% of Sample Foreign-Born White	12	11	29	33
Log Illiteracy Relative Risk of 0.5				
% of Sample Black	92	93	80	77
% of Sample Foreign-Born White	8	7	20	23
Log Illiteracy Relative Risk of 0				
% of Sample Black	95	95	87	84
% of Sample Foreign-Born White	5	5	13	16

Table 1: Optimal sampling design to estimate the log ratios between black and native-born white illiteracy rates within each state conditional on the ecological data and assuming log ratios in illiteracy rates between the racial groups are between 2.5 (top of table) and 0 (bottom of table). Because we are conditioning on the ecological data, the optimal design only involves sampling two of the three racial categories. For these values of log ratios in illiteracy rates, the optimal samples do not include native-born whites. As the log ratios in illiteracy rates approach 0, the % of the sample foreign-born white approaches the ecological proportions: Indiana 5%, Kansas 5%, Wyoming 13%, Nevada 16%.

	Indiana	Kansas	Wyoming	Nevada
% of Sample Native-Born White	0	0	0	73
% of Sample Black	73	71	35	27
% of Sample Foreign-Born White	27	29	65	0

Table 2: Optimal sampling design to estimate the log ratios between black and native-born white illiteracy rates within each state conditional on the ecological data and setting log ratios according to the population illiteracy rates broken down by racial group: Indiana (black 1.9, foreign-born white 2.4), Kansas (black 2.5, foreign-born white 2.5), Wyoming (black 2.6, foreign-born white 2.6), Nevada (black 2.0, foreign-born white 3.6). Notice that for Nevada, the log ratio for foreign-born whites is large enough that the optimal design has flipped in the sense that the optimal design includes native-born whites but not foreign born whites.

Sampling Distributions (300 observations per state)

The following section presents sampling distributions with 300 observations per state. In Figure 1 we present box plots of the sampling distributions of β_{1i} for each of the four states with 100 respondents in each racial group. The true parameter values are represented by the horizontal line. For each state, the first box plot represents the sampling distribution for estimation using the combined likelihood (1) from the paper. The third box plot represents the sampling distribution using the likelihood for only the individual level data. Note that because β_{1i} represents a log ratio, the individual level estimates are highly unstable. The second and fourth box plots represent the sampling distribution for the combined and individual approaches where all racial categories are artificially forced to have at least one illiterate person in the sample. Notice that this restriction stabilizes the individual only sampling distributions and reduces the variance of the combined sampling distribution, but it also induces some negative bias in the estimators (more so in the individual estimator).

In Figure 2 we present box plots of the sampling distributions of β_{1i} for each of the four states with 150 native-born white and 150 black respondents. The true parameter values are represented by the horizontal line. For each state, the first box plot represents the sampling distribution for estimation using the combined likelihood (1) from the paper. The third box plot represents the sampling distribution using the likelihood for only the individual level data. Note that again, the individual level estimates are highly unstable, although the combined estimates are also somewhat unstable as well. The second and fourth box plots again represent the sampling distribution for the combined and individual approaches where all racial categories are artificially forced to have at least one illiterate person in the sample. Again, this restriction stabilizes the sampling distributions and reduces the variance of the combined sampling distribution, but it also induces some negative bias in the estimators.

In Figure 3 we present box plots of the sampling distributions of β_{1i} for each of the four states using combined estimation, comparing optimal sampling (according to the first two rows of Table 1) and equal samples (e.g., 300 observations within a state are allocated as 100 native-born white, 100 foreign-born white, and 100 black). Again, the second and fourth

estimators restrict the number of illiterate respondents in each racial group to be greater than zero. Notice that optimal sampling (first box plot) is always preferable to equal sampling (third box plot). However, when samples are restricted to include at least one illiterate individual in each racial category (second and fourth box plots), the optimal samples have lower variance for Indiana and Kansas while the equal samples approach has a lower variance for Wyoming and Nevada. Again, this reduction in variance comes at the cost of of some negative bias for the equal samples estimator. It is not entirely surprising that the optimal samples are not always optimal once we have forced each racial sample to have one illiterate person in it. This restriction changes the estimator and the optimal sample sizes were not calculated while taking this restriction into account.

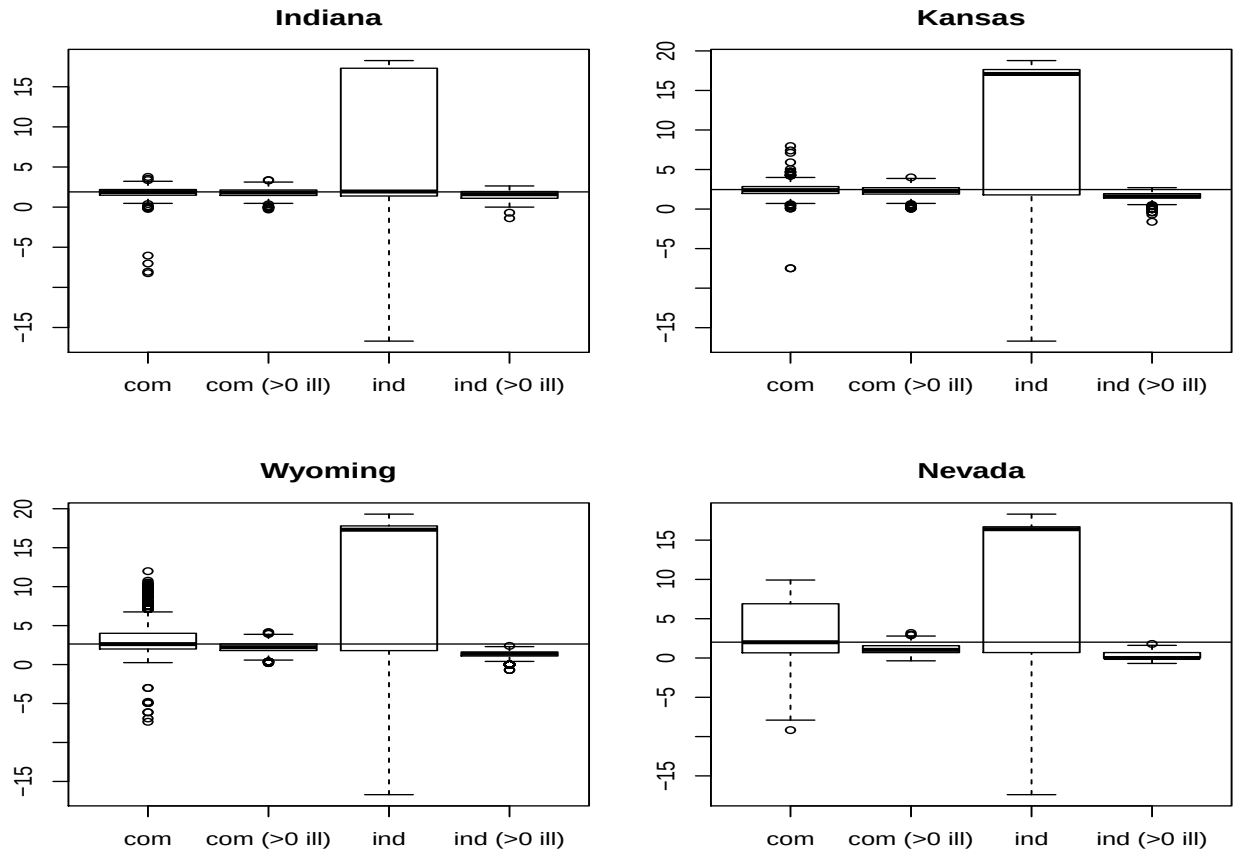

Figure 1: *Combined versus Individual with equal racial samples* Sampling distributions for the estimators of log relative risks in each of four states, β_{1i}, $i = 1, ..., 4$, based on stratified random samples stratified on state and race with equal within-strata sample sizes. (e.g., 300 observations within a state are allocated as 100 native-born white, 100 foreign-born white, and 100 black). The first two distributions are for the combined estimator. The last two are for the individual estimator. The second and fourth estimators restrict the number of illiterate respondents in each racial group to be greater than zero.

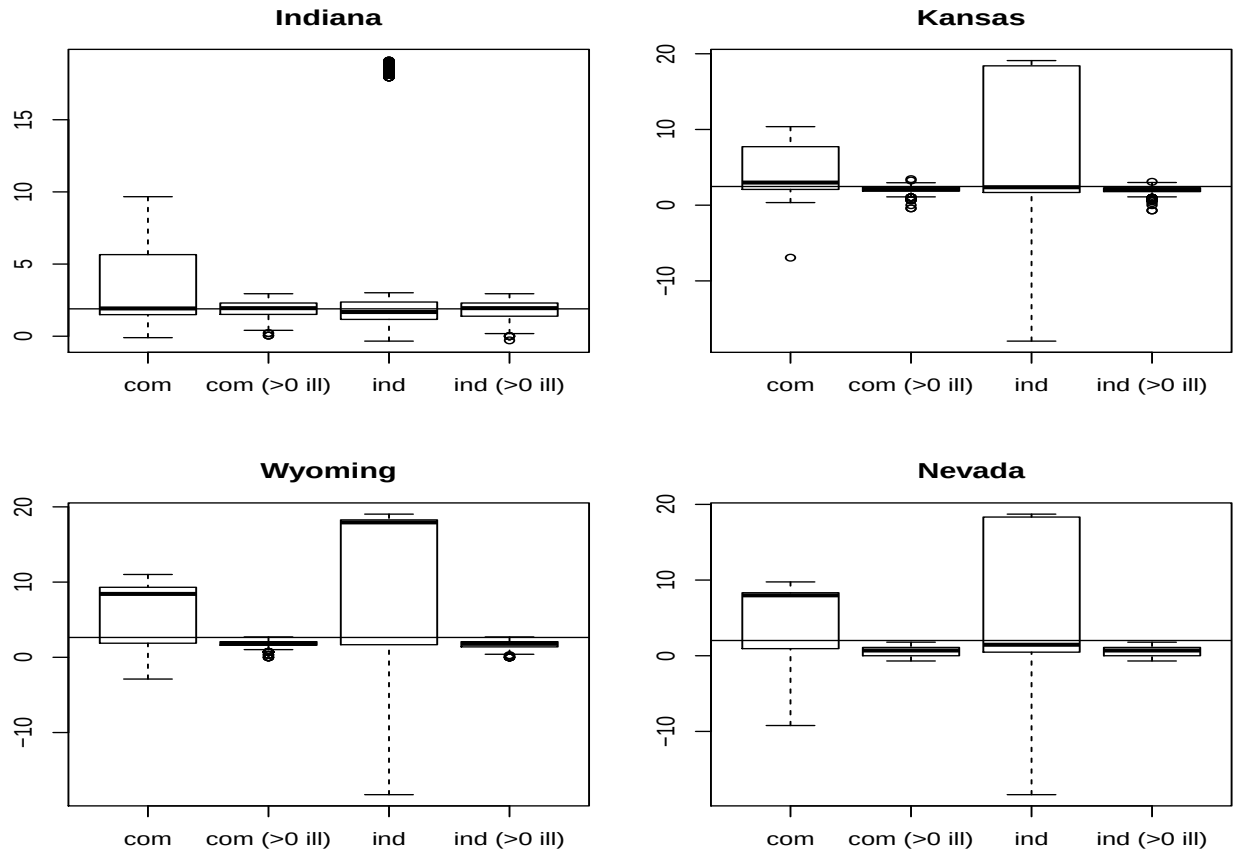

Figure 2: *Combined versus Individual with native-born white and black samples* Sampling distributions for the estimators of log relative risks in each of four states, β_{1i}, $i = 1, ..., 4$, based on stratified random samples stratified on state and race with only the native-born white and black racial categories (e.g., 300 observations within a state are allocated as 150 native-born white and 150 black). The first two distributions are for the combined estimator. The last two are for the individual estimator. The second and fourth estimators restrict the number of illiterate respondents in each racial group to be greater than zero.

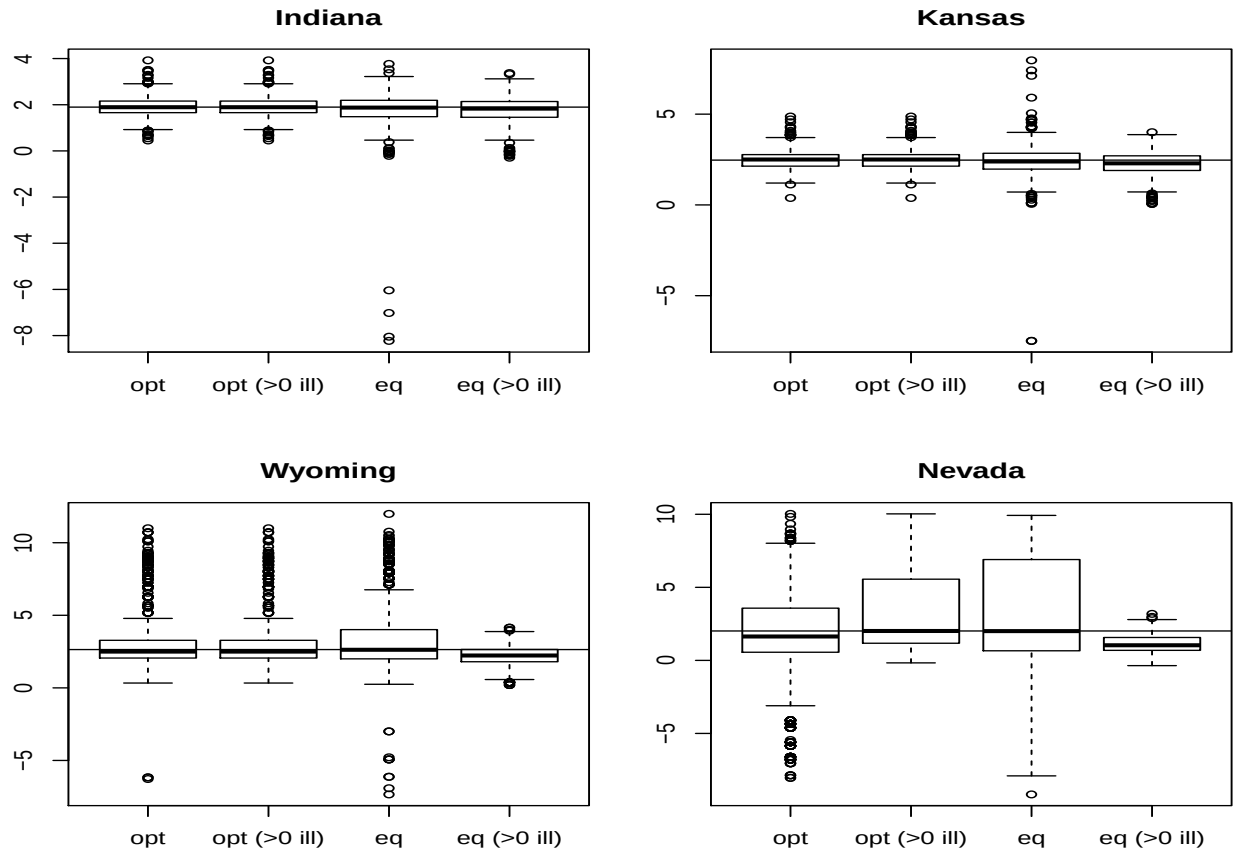

Figure 3: *Optimal Samples versus Equal Samples for Combined* Sampling distributions for the estimators of log relative risks in each of four states, β_{1i}, $i = 1, ..., 4$, based on stratified random samples stratified on state and race. The first two distributions are for the combined estimator with optimal samples (according to the first two rows of Table 1). The last two are for the combined estimator with equal samples (e.g., 300 observations within a state are allocated as 100 native-born white, 100 foreign-born white, and 100 black). The second and fourth estimators restrict the number of illiterate respondents in each racial group to be greater than zero.

Sampling Distributions (600 observations per state)

Figures 4, 5, 6 present similar box plots but with 600 observations within each state. The results are very similar to those of the previous section, although we note that the skew of the unrestricted sampling distributions has been reduced.

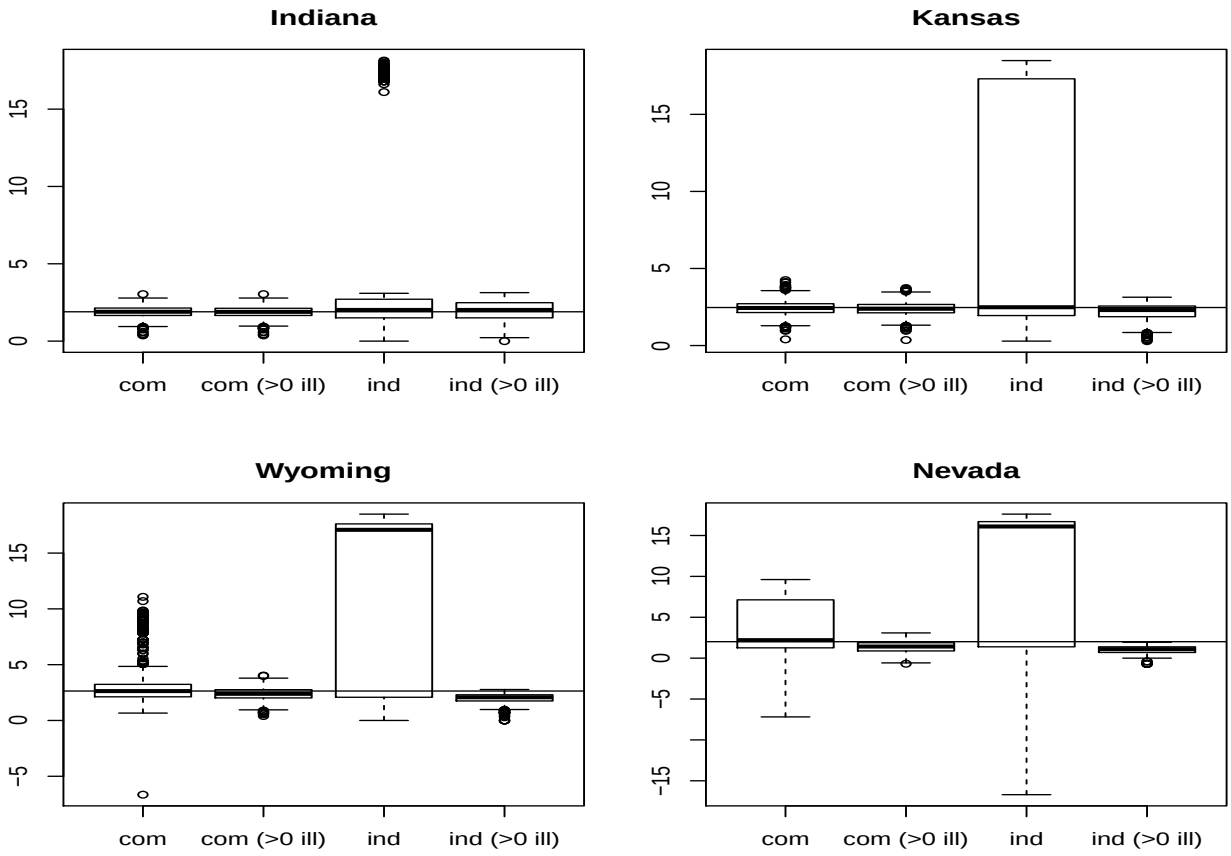

Figure 4: *Combined versus Individual with equal racial samples* Sampling distributions for the estimators of log relative risks in each of four states, β_{1i}, $i = 1, ..., 4$, based on stratified random samples stratified on state and race with equal within-strata sample sizes. (e.g., 600 observations within a state are allocated as 200 native-born white, 200 foreign-born white, and 200 black). The first two distributions are for the combined estimator. The last two are for the individual estimator. The second and fourth estimators restrict the number of illiterate respondents in each racial group to be greater than zero.

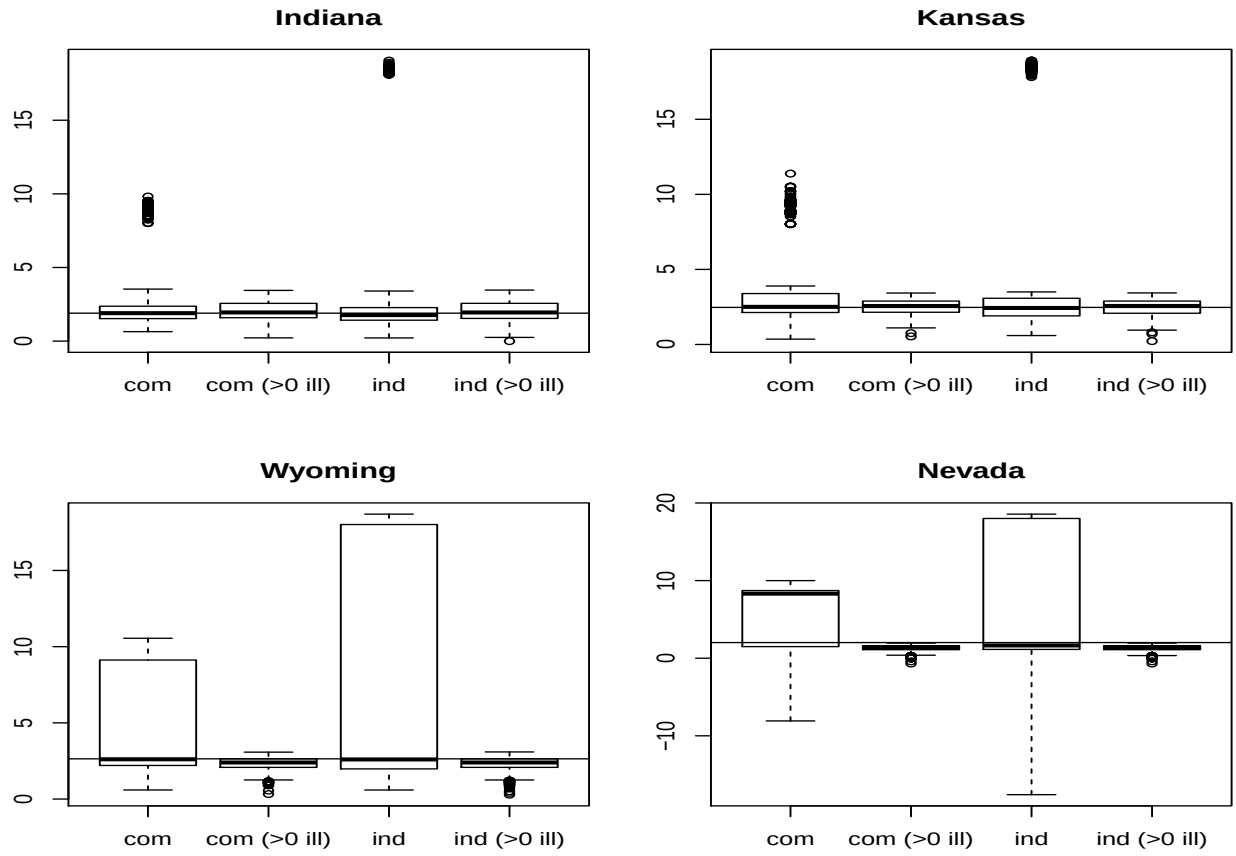

Figure 5: *Combined versus Individual with native-born white and black samples* Sampling distributions for the estimators of log relative risks in each of four states, β_{1i}, $i = 1, ..., 4$, based on stratified random samples stratified on state and race with only the native-born white and black racial categories (e.g., 600 observations within a state are allocated as 300 native-born white and 300 black). The first two distributions are for the combined estimator. The last two are for the individual estimator. The second and fourth estimators restrict the number of illiterate respondents in each racial group to be greater than zero.

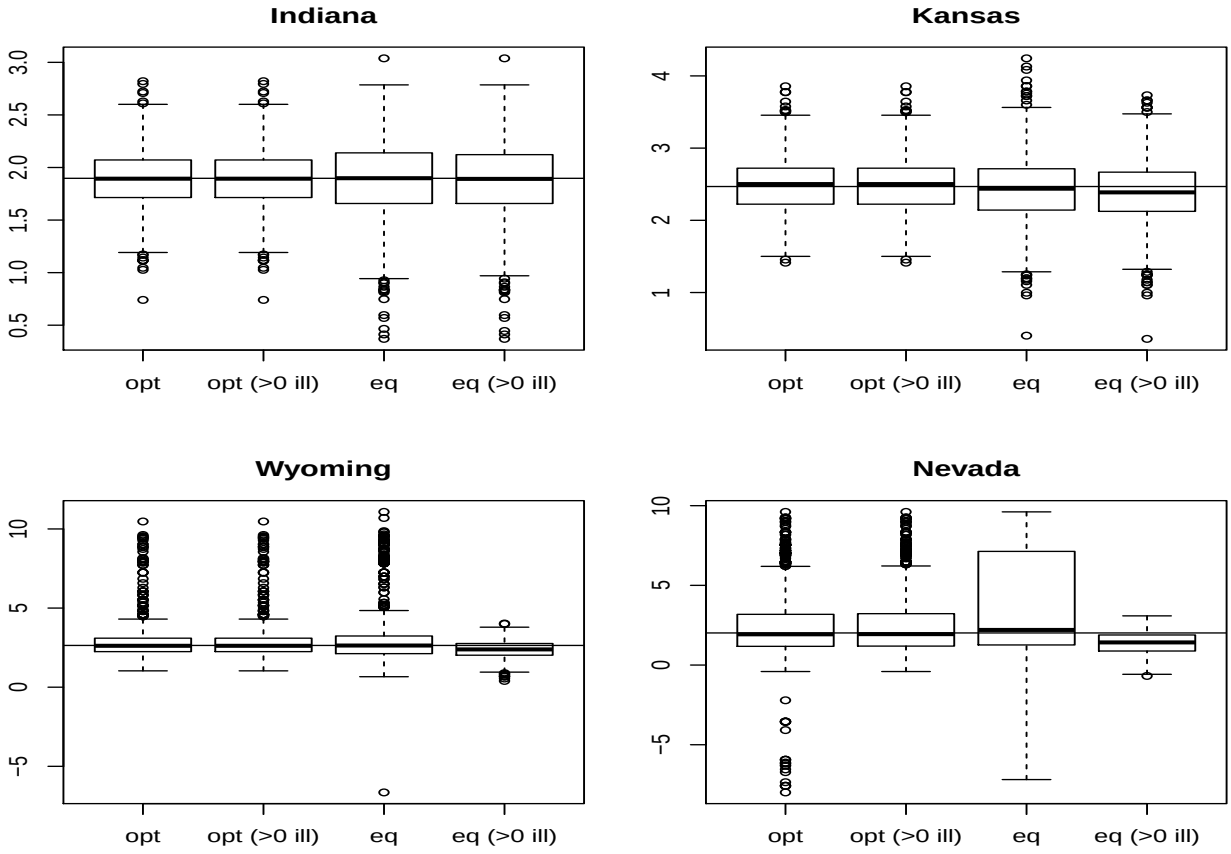

Figure 6: *Optimal Samples versus Equal Samples for Combined* Sampling distributions for the estimators of log relative risks in each of four states, β_{1i}, $i = 1, ..., 4$, based on stratified random samples stratified on state and race. The first two distributions are for the combined estimator with optimal samples (according to the first two rows of Table 1). The last two are for the combined estimator with equal samples (e.g., 600 observations within a state are allocated as 200 native-born white, 200 foreign-born white, and 200 black). The second and fourth estimators restrict the number of illiterate respondents in each racial group to be greater than zero.

Alternative Matches and Expanded Analysis

As described in the Discussion of the paper, Figure 7 summarizes the associations that result if we match on geography: Iowa matched to Kansas and Montana to Wyoming. We see that the Jim Crow states now have lower literacy rates than their geographically-matched counterparts, a reversal of the association observed in the main analysis of the paper in which the matching was carried out on the basis of the sizes of the black population and of the foreign-born white population. This demonstrates that the subsampling and inference procedures discussed in the paper do not solve fundamental modeling decisions such as which variables to match on, or which variables to include in the regression model.

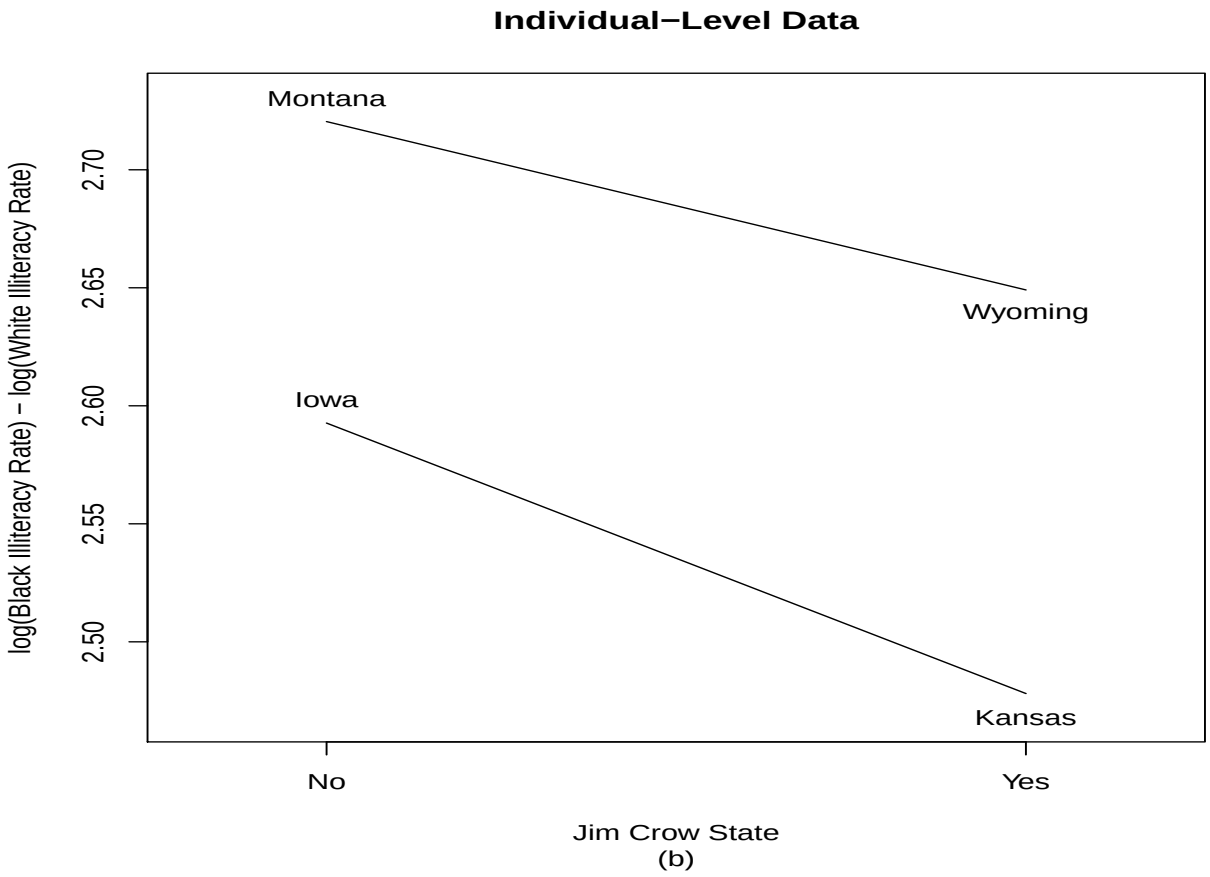

Figure 7: *Individual Level Analysis with Alternate Matches*

Additionally as described in the Discussion of the paper, Figure 8 summarizes the associations that result if we do not leave out the states that were part of the Confederacy during

the Civil War. In this scenario, Jim Crow states that were part of the Confederacy during the Civil War are included in the analysis, and matched to non-Jim Crow states that were not members of the Confederacy. One-to-one, without replacement, Mahalanobis distance matching on sizes of the black population and of the foreign-born white population produces the analysis pictured in Figure 8, which includes all states with the exception of Massachusetts, Rhode Island, Connecticut, New York, and New Jersey. We see that the Jim Crow states now have lower literacy rates on average than their matched counterparts, a reversal of the association observed in the main analysis of the paper. This again demonstrates that the subsampling and inference procedures discussed in the paper do not solve fundamental modeling decisions such as which variables to match on, or which variables to include in the regression model.

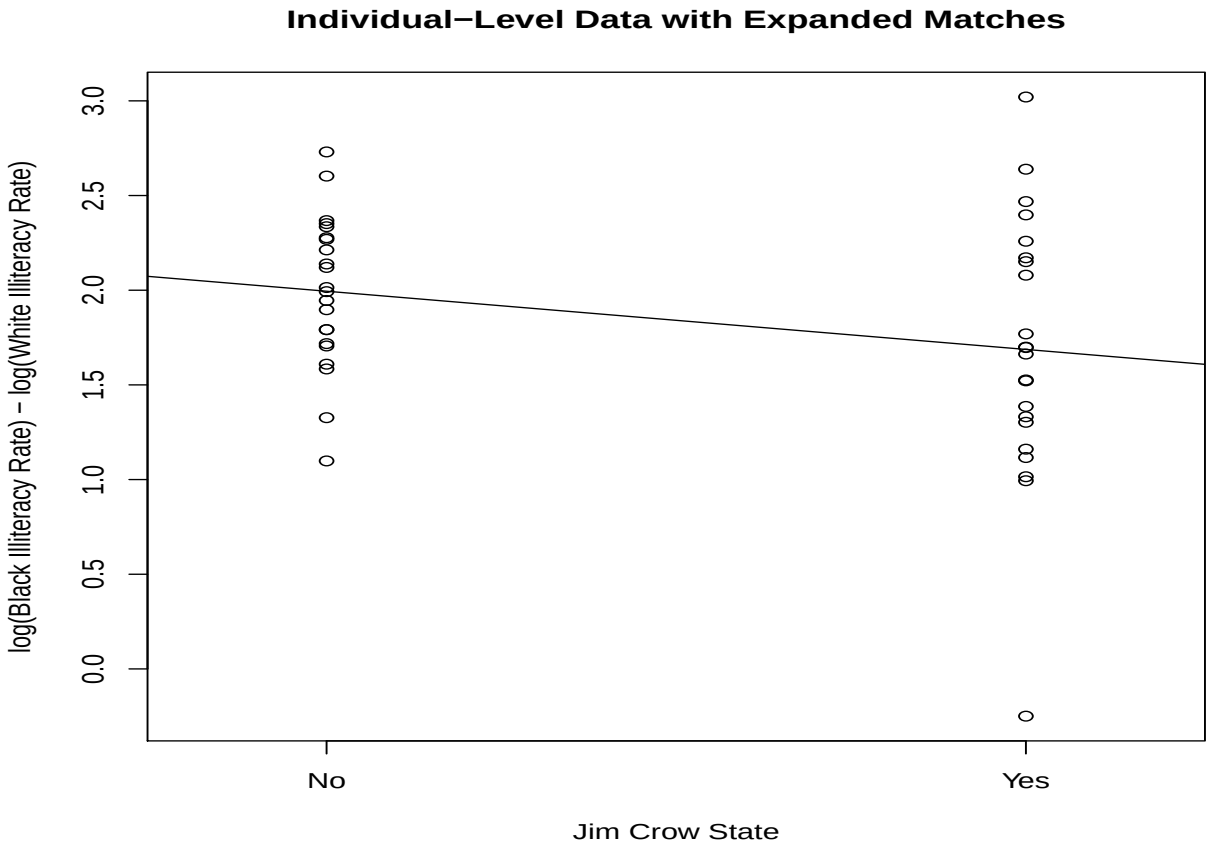

Figure 8: *Individual Level Analysis with Expanded Analysis featuring states that were part of the Confederacy during the Civil War.*