Evaluating Risk Assessment: Finding a Methodology that Supports your Agenda

There are few issues facing criminal justice decision makers generating more interest than fairness and bias with risk assessments. Journalists, scholars, judges, and defendants have questioned the potential unfairness in risk assessments. A lively debate has emerged recently about the fair use of risk assessments to inform decisions about pretrial release or detention. Despite the widespread use, even mandatory use of risk assessments (e.g., Pretrial Integrity and Safety Act, 2017), criminal justice researchers have yet to develop agreement on standards of fairness or the methods to assess bias. The lack of standards has given rise to a series of studies that come to very different results regarding the presence of disparate impact.

Differential prediction in the criminal justice system received a lot of attention due, in part, to a ProPublica article that emphatically stated that the COMPAS risk assessment equated to "machine bias" (Angwin, Larson, Mattu, and Kirchner, 2016). This article set off something of a firestorm within the criminal justice research and practitioner communities because their analysis "turned up significant racial disparities" (Angwin et al., 2016). The ProPublica article, of course, was not the first to suggest biased prediction. Some criminologists have long argued that risk assessments are part of a "new penology" and a necessary element of mass incarceration (Feely and Simon, 1992).

Risk assessment proponents, however, argue that they "can scaffold efforts to unwind mass incarceration without compromising public safety" (Skeem and Lowenkamp, 2016: 705). But, questions of fairness remain. Flores, Bechtel, and Lowenkamp (2016: 45) analyzed the same data using a different methodology in a rejoinder, and they came to a nearly opposite conclusion of "no evidence of racial bias." They even claimed that ProPublica "strayed from their own code of ethics in that they did not present the facts accurately" and that the article provided "misleading information about the reliability, validity, and fairness" of risk assessments (Flores et al., 2016: 45). It appears, however, that Angwin et al. are less likely to have committed any ethical breach and more likely to have used a different criterion than Flores et al. to evaluate fairness.

Fairness is a subjective issue that has yet to be clearly defined by criminal justice scholars. Using several definitions of fairness for statistical models, Chouldechova (2017) conducted a third assessment of the COMPAS risk assessment data and concluded that, given the differences in the base rates for rearrest between African-Americans and Whites, it was impossible for Flores et al. and Angwin et al. to agree. Angwin et al. found imbalance in the error rate (i.e., equalized odds) with nearly twice as many African-Americans that were classified as higher risk did *not* fail (i.e., were not

rearrested) compared to Whites (42 percent vs. 22 percent, respectively). Flores et al., on the other hand, studied what is known as predictive parity and found that for those classified as higher risk failure rates were similar for both race groups. Chouldechova (2017) points out that it is mathematically impossible for a risk assessment to have both error rate balance and predictive parity when there are differences in base rates by races. In the COMPAS dataset, 52 percent of African-American defendants and 39 percent of White defendants were rearrested.

Mass incarceration is one of the most pressing issues facing the judicial system. There are currently more than 2 million incarcerated adults, over 10 million adults are admitted to jail annually, and more than 600,000 adults are sentenced to prison each year. Although at one time risk assessment was seen as part of mass incarceration, now it is seen by many as way to ease the burdens. There are criminal justice reforms at all levels of government and many of these efforts have risk assessment as a central component. Yet, there is little agreement about how to define and measure fairness, much less about how to address issues of unfairness. We discuss how fairness can be practically measured under the moderator regression methodology using a Bayesian approach and describe issues with the standard approach of moderator regression under a frequentist paradigm

Methods

Data

The data are a sample of individuals booked in the Broward County, Florida jail to assess two rearrest rates. The dataset is publicly available and is the same used for the ProPublica and Flores et al. studies. We followed their data processing steps by removing individuals with missing offense data, those without COMPAS risk scores, and those with only a traffic offense. Our dataset differs in one way in that we include all races to show the flexibility of our model procedures. This resulted in 894 more individuals included in the general recidivism dataset. While we build the model on all the data to obtain more accurate estimates, we focus our analysis on differences between African American, Caucasian, and Hispanic race/ethnic subgroups. The outcome variable is rearrest within two years of release from jail (n=6172).

Model

We fit a Bayesian multilevel model with weakly informative priors that predicts recidivism based off an intercept value and COMPAS score, with varying slopes and varying intercepts by racial subgroup. This is equivalent to a logistic regression version of the models used in testing for differences in slopes and intercepts under the traditional Cleary methodology.

Why Bayesian?

A Bayesian framework allows for inferences based on a distribution of parameter values. Since moderator regression is inherently concerned with the parameter values for slopes and intercepts, a Bayesian lens allows us to ascribe credibility to different values of those parameters. Thus, a distribution of parameter estimates from a Bayesian model gives us a larger, quantitative vocabulary to discuss differential prediction. A full treatment of Bayesian statistics is outside the scope of this paper, but Kruschke and Liddell (2017) provide an introduction to the fundamentals.

Parameter Transformations

Since we're predicting a binary outcome with logistic regression and are focused on practical differences, we focus on comparing the odds ratios between values of intercepts and slopes for each racial/ethnic subgroup. Odds ratios account for differences in the base rate of the outcome. An odds ratio of 1 means that the odds are equal for the two groups being compared. An odds ratio of 1.1 can be interpreted as indicating for every 110 individuals that recidivate from the numerator group, 100 recidivate from the group in the denominator of the ratio.

Since the odds of the outcome change for each unit increase in the slope, we evaluate differences in odds ratios for slopes given an individual scores a 10 on the COMPAS (the highest possible score). This allows drawing comparisons between subgroup models with the same intercept, but differing slopes.

How to Assess Results

We evaluate differences in model slopes and intercepts by comparing the highest posterior density interval (HPD) to a region of practical equivalence (ROPE). The ROPE signifies a range of parameter estimates that we wish to consider practically equivalent to no difference. This methodology is sometimes referred to as equivalence testing.

Using the ROPE as a decision rule has three potential outcomes (in reference to a null value parameter):

1. If the ROPE lies outside the 95% HPD of the posterior, the parameter value is not credible and is rejected.
2. If the ROPE completely contains the 95% HPD of the posterior, the parameter value is accepted.
3. If the ROPE and HPD overlap, with the ROPE not completely containing the HPD, then we withhold a decision as the current data are insufficient to yield a clear decision.

One additional advantage of equivalence testing using the ROPE is the ability to infer that there is no practical difference between estimates (accepting the null), a form of analysis not possible with

traditional statistical analysis. We use the following rules to assess predictive differences among the race/ethnicity subgroups: if a 95% HPD falls completely within a ROPE, we can conclude there is no difference. If there is overlap, we withhold a conclusion until further data can collected and additional analyses performed. If the 95% HPD lies outside the ROPE, we conclude there are meaningful group differences in the ability of the COMPAS to predict rearrests.

Analyses

We begin the analyses with a visual inspection of the predicted probabilities of failure by race/ethnicity and COMPAS score (ranges from 1-10) with the actual outcomes.

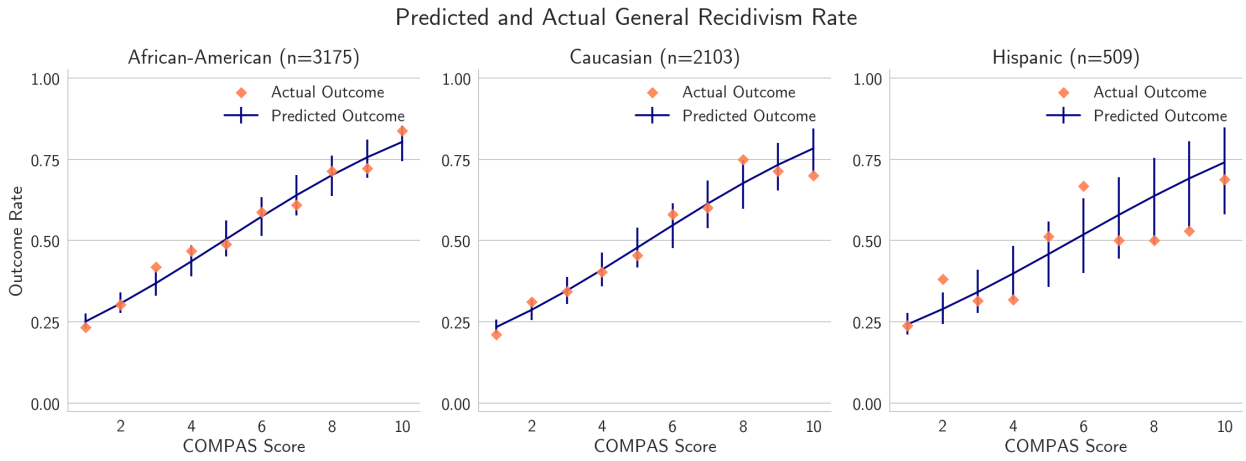

Figure 1: Predicted Probabilities and actual outcome rates by score for each racial/ethnic subgroup

From this visual inspection, we can see that, in general, higher risk scores are associated with higher predicted probabilities. The model appears to be less calibrated towards prediction of Hispanics, showing higher variation of actual outcomes and a smaller effect of increasing COMPAS score.

Posterior Distributions of Intercept and Slope Parameters

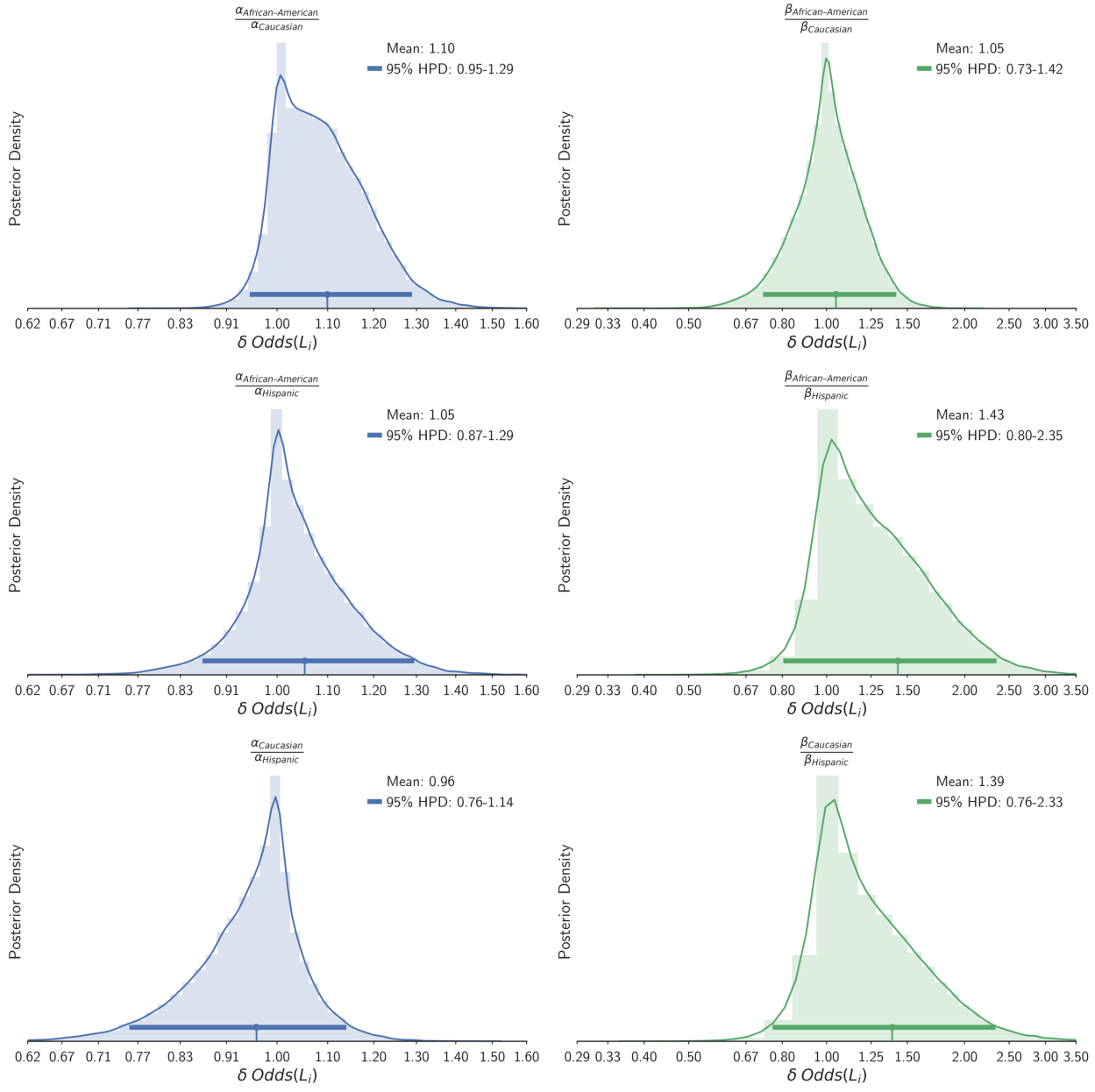

Figure 2: Posterior Distributions of Odds Ratios between intercepts and slopes. Slope odds ratios are calculated at a score of 10 on the COMPAS.

Analysis of Intercept and Slope Parameters

Analysis of Intercepts

The posterior distributions on the left of Figure 2 display the odds ratios between the intercepts of the race/ethnicity subgroup dyads. For these analyses, we are comparing subgroup recidivism for a

score of 1 (lowest) on the COMPAS. Since we are using the full distribution of parameter estimates, we can fully understand the uncertainty around our parameter estimates by using the 95% HPD.

In comparing intercepts between African American and Caucasians, it appears there are differences in slope values, as the mean of the posterior distribution is 1.10. The probability of an odds ratio difference greater than 1, indicating any difference is 87%.. However, there is large uncertainty surrounding that measurement, with a 95% HPD of 0.95–1.29. From a point estimate perspective, this could be indicative of bias, but within the context of our uncertainty, it is also plausible there's no bias, since 1 is included in the 95% HPD.

Examining the comparisons to Hispanics, intercept values for African Americans are typically higher (mean OR: 1.05), and Caucasians are lower (mean OR: 0.96). The 95% HPD also includes 1, so it's still also possible that there are no true differences between values.

At this point, we're approaching a paradox: How can it be possible that there is bias when looking at point estimates but no evidence of bias when considering the uncertainty? It may be helpful at this point to make a comparison to traditional hypothesis testing. In such a framework, we would not have enough evidence to reject the null hypothesis in any of our comparisons, since all distributions have an odds ratio of 1 within their 95% HPD. This could be construed as a win for those who would conclude no bias. However, failing to reject the null hypothesis is not the same as accepting the null hypothesis of no difference. To accept a hypothesis of no differences, we employ a region of practical equivalence.

Assessing Meaningful Intercept Differences: ROPE

Something missing from prior studies is to establish a clear criterion (outside of significance testing with p-values) to determine fairness. We defined bias using the ROPE to determine when difference between subgroups falls completely outside of the 95% HPD.

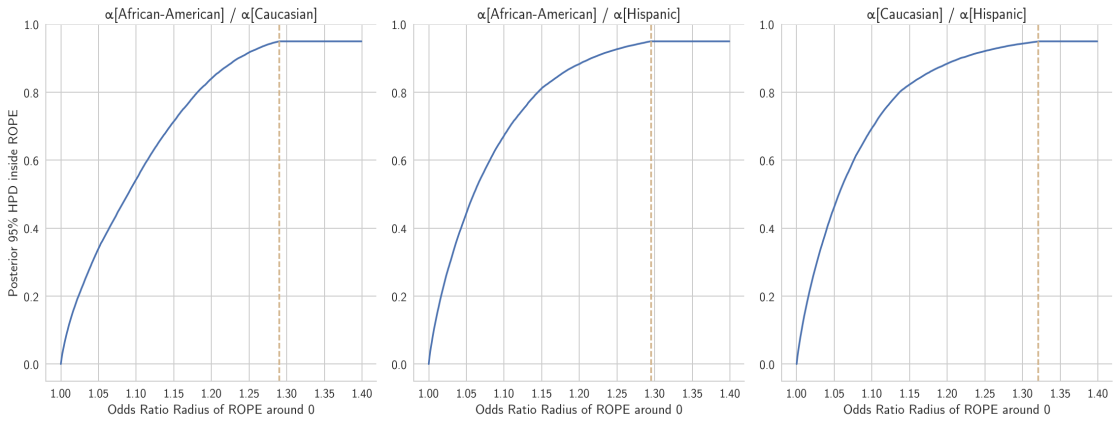

Figure 3: ROPE radius thresholds required to accept the null hypothesis of no difference in intercepts.

The vertical lines in Figure 4 indicate the value of a ROPE for each intercept odds ratio that would be required to conclude that there is "no difference" in intercepts. For all subgroups, it appears that difference is between around 1.30. This would mean defining no practical difference in intercepts as anywhere an odds ratio less than 1.30 exists. Or, for every 100 White defendants that are rearrested there are 130 African-American defendants (or a roughly 30% difference).

Analysis of Slopes

Since the odds of the outcome change for each unit increase in the slope, we evaluate differences in odds ratios given an individual scoring a 10 on the COMPAS (the highest possible score). This allows drawing comparisons between subgroup models with a hypothetically identical intercept, but differing slopes.

Slope differences between African-American and Caucasian subgroups appears to be centered around 1 (mean OR: 1.05). There is large uncertainty around the estimate for the Hispanic subgroup and resulting in a distribution of odds ratios results in underprediction for Hispanics. The probabilities that odds ratios for African American and Caucasian compared to Hispanics is larger than 1 are 0.87 and 0.84 respectively. However, the probability that the odds ratios are greater than 1.25 are only 0.57 and 0.51, indicating less certainty that the differences are large. This would suggest that analysis of risk assessment instruments that focus on slope differences relative to the Hispanic subgroup, but that those differences may be small and not practically significant.

Assessing Meaningful Slope Differences: ROPE

If we are willing to accept practical differences at an odds ratio of 1.42, then we could definitively conclude that there is no differential prediction between African American and Caucasian subgroups for slope values. We would have to set a radius around 2.35 to conclude that there are no differences in comparison to the Hispanic subgroup.

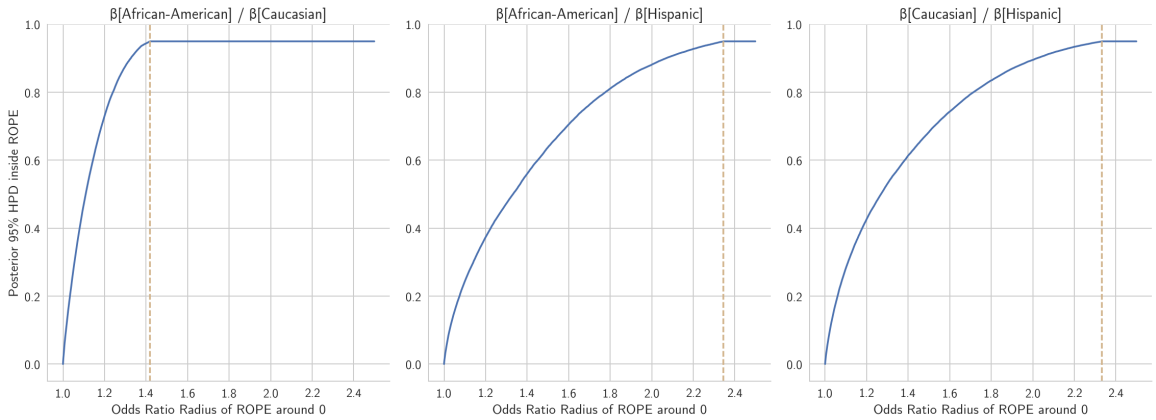

Figure 4: ROPE radius thresholds required to accept the null hypothesis of no difference in slopes.

These results indicate that to conclude no subgroup differences would produce an assessment instrument that for every 100 Whites there are 140 African-American defendants rearrested.

Conclusions

It is no longer a matter of *if* criminal justice agencies will use risk assessment instruments. Rather, risk assessment instruments increasingly are being required at various stages of the criminal justice system (e.g., sentencing, pretrial release). There is ample evidence that communities of color, the poor, and mentally ill have disproportionately felt the effects of mass incarceration. The former U.S. Attorney General Eric Holder (2014) warned that risk assessments could "inadvertently undermine our efforts to ensure individualized and equal justice…they may exacerbate unwarranted and unjust disparities that are already far too common in our criminal justice system" (Holder, 2014; cited in Angwin et al., 2016). The criminological community has yet to define fairness, nor provide clear direction about the tools needed to assess fairness. Our paper advances current quantitative methods to assess potential disparate impacts for communities of color with a multilevel Bayesian approach.

The analyses show there are differences between African-American and White defendants with low scores. For those with a score of 1 on the COMPAS, African-Americans had higher failure rates than White and Hispanic defendants. Explaining why these differences exist is beyond our purposes, but prior research demonstrates that African-Americans are more likely to be stopped by the police, when stopped, they are more likely to be searched, and when searched, they are more likely to be arrested than Whites. Differential selection is a well-known feature of crime control practices that has the potential to influence risk assessment validations.

The findings show no difference between African-American and White defendants with high scores, but we found that Hispanic defendants had meaningful differences with African-American and White defendants. There is the cliché of "garbage in, garbage out" that is used to suggest that

when poorly collected or questionable data are used to model outcomes, one will end up with questionable predictive models. Criminal justice agencies are well-known for struggling to collect accurate demographic information, which is especially troublesome regarding Hispanicity. It is possible that data quality may influence the posterior distributions for Hispanic defendants. Interestingly, the lack of slope differences between African-American and White defendants potentially suggests that the COMPAS is better at identifying risk than non-risk. Simply, criminal justice decision makers are most concerned with false negatives in which someone that commits a crime was predicted not to do so and the COMPAS may be normed with acceptance of higher false negative rates. Northpoint Inc. owns the COMPAS and they have refused to be transparent in the development and ongoing refinement of the assessment.

Our purpose for this paper is methodological. The recent studies using the same dataset have made starkly different yet strong declarations about disparate impacts. It appears that confirmation bias may be at work, with proponents of risk assessments analyzing the data in one way and the opponents analyzing it in another way, and producing opposing interpretations. We are approaching our findings with far more caution to suggest that more data are needed, additional analyses should be conducted, and a spirit of collaboration among researchers is needed to find common ground to define key concepts, such as: What is a fair risk assessment? What are the appropriate methods to assess fairness? How much difference in prediction is practically significant? The consequences for public safety, the administration of justice, and the equal treatment for all defendants requires some agreement on fairness.