LS 123 Data, Prediction, and Law

Meeting 11

**COMPAS Rejoinder, Regression for Prediction**

1. **COMPAS prediction instrument and its defenders--Flores**
   1. two really different conclusions on the accuracy (but maybe not on the predictive utility) of the COMPAS prediction tool
   2. more free form discussion: I will carry around the microphone
      1. how do Flores et al criticize the ProPublica analysis? they critique it from a number of angles
      2. how good are their critiques? say why you think that
   3. today it is a rebuttal to the ProPublica critique of COMPAS –how would you characterize the paper from Flores, Lowenkamp, and Bechtel?
      1. not just a scholarly critique of the data, tests, or models that ProPublica used to make its claims
      2. instead it is a full throated defense of actuarial risk assessment instruments
         1. they are better at selecting the right people to punish than are the judges
         2. they work, and if we want to have a different punishment policy, we should debate that
         3. authors want to defend actuarial instruments against the Bernard Harcourts of the world
         4. there’s no empirical evidence that using ARAI’s result in racial bias—it is all theorizing (again, a dig at Harcourt, I think)
   4. they make several critiques
      1. ProPublica used the wrong dataset to test COMPAS
      2. they used the wrong methods to detect bias
      3. they made errors in their methods anyway (e.g., throwing away data by only using the three groupings of COMPAS scores instead of all 10)
   5. wrong population to use the COMPAS instrument
      1. recidivism prediction instruments supposed to be for people who have already been convicted
      2. here we are talking about? pretrial detainees in Broward County
      3. there is a different instrument for pretrial detention that predicts risk of nonappearance
      4. the instrument was never validated on this population so you should not make inferences here
   6. your analysis methods were wrong
      1. there is an established way to measure bias for psychometric testing and you didn’t use it
      2. to measure test bias you should have used AUC-ROC (which measures the degree to which a test distinguishes true positives from false positives—comes from signals detection for radar operators in WWII)
         1. imagine in medical testing: pick one person at random from those known to have the disease, and one from those known not to have it, and pair them. Each person has a test result. The AUC-ROC score indicates how frequently in the pool of test pairs the person with more abnormal test result is the one with the disease
         2. so an AUC-ROC of .71 means that in any pair (defined as one known recidivist and one known non-recidivist, drawn at random), the algorithm will show the higher score for the recidivist 71% of the time
      3. even if we use two year recidivism logistic model, you did it wrong
         1. by using only two risk categories in your model, you threw away data
         2. we use all ten deciles in our model and find that the logistic model behaves similarly for black and white defendants in Broward county
         3. when we included an interaction term for race and COMPAS score it was not statistically significant
         4. you guys were cagey with your hypothesis testing; even with the big sample you didn’t get a p < .05 so you cannot reject the null hypothesis that there is no interaction between the COMPAS instrument and race of the defendant
            1. this is not as bad as they make it out to be; frequentist statistics settled on that value for the probability somewhat arbitrarily
            2. again, p < .05 means that in repeated sampling, the sample value will be within a particular range of the true parameter 95 times out of 100 (so here it means that 94 times out of 100 the coefficient estimate would not have zero in its confidence interval)
            3. which means, there really could be a true effect, since there is nothing that magic about 95 times out of 100
         5. for neither set of models (the ProPublica and the Flores et al) is much of the variation actually accounted for by the model
   7. so who is right?
      1. maybe the measure of predictive accuracy is really important (although reading just a bit on AUC-ROC makes me think that the .71 figure they cite is not really impressive accuracy, and in diagnostic medicine would be mediocre)
      2. so perhaps we need to think more about measures of predictive accuracy
      3. we do need to ask how robust the predictions are, given the very different conclusions just based on model specification (and the not so good fit between model and data)
   8. to me, asking whether the instrument is biased seems like the wrong question
      1. what about the process that generates criminal records, the COMPAS score, and re-arrest? that is, bias against Black people could be built into all the measures
      2. once you have been selected into being “criminal” then the model is not going to behave too differently for you,
         1. even if being selected into the category is not at all random or
         2. being selected is not proportional to your group’s true rate of participation in criminal activity (Harcourt)
   9. so, if the bias is elsewhere, does this vindicate predictive tools like COMPAS?
   10. since all the data are available from the ProPublic people, we can reanalyze it (and we may want to if we have the opportunity)
2. **Dressel & Farid: COMPAS as snake oil**
   1. The algorithm is not all that accurate, and has problems in predicting risk levels incorrectly for black and white defendants
   2. MTurkers did just as well, using the AUC-ROC measure, as COMPAS
   3. So a small crowd of amateurs, with only seven predictors, is just as good as COMPAS
   4. Race information did not generate different results—there was enough bias in the “race blind” condition for the MTurkers
   5. You can also do just as well as COMPAS with a logistic regression model with the same 7 features the MTurkers used
   6. Support Vector Machine did not really do better at separating the data than the linear (logistic regression) model—maybe you cannot do this accurately
   7. More on that next time with **Kleinberg** “Human Decisions and Machine Predictions”—maybe we can train models appropriately to keep humans from making worse decisions than they otherwise might
3. **Lab 12—Data Splitting; regression for prediction**
   1. Difference between regression for causal inference and regression for prediction
      1. Causal inference: we have a theory that gives us an alternative hypothesis to the null hypothesis, and the theory specifies how the various causal variables are expected to affect the outcome variable
      2. So we really care about having the variables in the model and keeping them there—the model is driven by theory; we also care about interpreting the effect size of each variable so we tend not to normalize like we do in ML models (see the SciKit learn documentation)
      3. Prediction: we are interested in getting close to reality, and we do not really have a model based on theory that we are trying to investigate
      4. Note that regression models have some advantages
         1. Modified appropriately, they can work for classification problems—we’ve already talked about this with logistic regression, where we transform the “y hat” (what we’re trying to estimate) into the natural log of the odds ratio so that we are estimating something continuous
         2. They can provide estimates for continuous outcomes, of course
         3. It is relatively easy to see what is going in the model, even when it is a logistic regression model
         4. Refinements like ridge and lasso can help remove predictors that increase the variance of the estimate (at the cost of bias)
   2. Problem of preventing overfitting to training data
      1. Training—validation—test split to let you know whether or not you are overfitting
      2. Techniques like ridge and lasso shrink less important predictor coefficients and so the predictors that are really just noise are zeroed (or almost zeroed) out
   3. Change loss functions to minimize the less influential (once they’ve been *standardized* – or normalized, in Data Science lingo -- by setting the mean to zero and variance/distance from origin/L2 norm to one) features: OLS loss function is sum of squared residuals, and for ridge & lasso we add an additional term to the loss function
   4. Ridge: minimize the sum of the sum of squared residuals plus the sum of the squared coefficients (all the betas or all the terms in that vector we called theta in last Wednesday’s lab)
   5. LASSO: minimize the sum of the sum of the squared residuals plus the sum of the absolute value of the coefficients