Discussion regarding the appropriateness of including covariates

Analyst 1

This makes it sound like this was just another choice that analysts could have made about the meaning of these variables. In truth, because the data was not structured in a way that when a player received a red card their club, league, or even their age could be adequately determined, these variables are not appropriate for inclusion in ANY analysis. It is critical to make this clear so that any future analysis done on the posted data set will take this into account as well.

Analyst 2

That is not entirely true. Players probably tend to stay withing the same league and club, and younger players could not have got a red card during their older age. There is of course a lot of noise, but it is not true that the variables are not appropriate for inclusion in any analysis.

Analyst 1

First, when is it appropriate to analyze data that you think is "probably" accurate? Second, there is much more transferring between leagues than you seem to appreciate. Weaker players wash out of, say, the EPL and get picked up by lesser teams in, say, the Bundesliga. Or breakout stars in one league get bought up by the top teams in another league. Given that the transfer system is a completely free market and all it takes for players to change teams is money, there is a lot of roster churn. But, again, the key point is that we don't know the level of misclassification in the data set (i.e.., whether a red card occurred when the player was in the listed league or not), so we can't properly analyze that data. The same is true of age. Third, these discrepancies are particularly problematic in the present data set because red cards are such a rare event. It would take very few misspecifications to create a major effect on the results (e.g., one or two players listed as 27 in 2012-2013 who actually got a red card when they were 22 or 23) for that to throw off the whole analysis. The same applies for league. So this isn't about having a "good enough" system of classification or introducing just a little noise into the analyses. It's a major issue.

Analyst 2

You have almost always some noise in measurements. The older players here were more likely to play in their higher age than younger players, so if age influences somehow probability of receiving a red card, than there is also some information in the covariate. I did not use any of the covariates in the final analysis as well, but it is not true that they are not appropriate for inclusion in any analysis.

Furthermore, you write in another comment that "If these are the final analyses, the continued use of invalid covariates is a HUGE problem." I did an analysis both with and without age as covariate and it made virtually no difference. I do not believe that use of those covariates changes much in results.

Analyst 1

The fact that age doesn't happen to change anything shouldn't provide any confidence that age doesn't matter. If the variable is misspecified, in reality age could be an important factor in red cards and we would falsely conclude that it isn't important. That's why the results involving those covariates shouldn't be interpreted at all. And, as I noted because red cards are rare, we can't know how big a factor the misspecification is - it would only take one non-typical case no matter what the underlying age distribution to seriously skew things.