From: **Johannes Ullrich** j.ullrich@psychologie.uzh.ch
Subject: Re: CS1: Covariate Discussion
Date: 11 Feb 2016 13:11
To: Silberzahn, Raphael RSilberzahn@iese.edu
Cc: ssyoon@temple.edu, Rickard Carlsson rickard.carlsson@lnu.se, Christoph Spörlein ch.spoerlein@gmail.com, Erikson Kaszubowski erikson84@yahoo.com.br, molden@northwestern.edu, Alicia Hofelich Mohr hofelich@umn.edu, Garret S. Christensen garret@berkeley.edu, mathew.evans@manchester.ac.uk, b.a.nijstad@rug.nl, Tom Stafford t.stafford@sheffield.ac.uk, Michelangelo Vianello michelangelo.vianello@unipd.it, Kent Hui huingan@msu.edu, dkennedy@uwb.edu, feng.bai10@rotman.utoronto.ca, fabia.hoegden@uni-koeln.de, Frederik Aust frederik.aust@uni-koeln.de, eawtrey@uw.edu, SOMMERa@hec.fr, egidio.robusto@unipd.it, brysonrpope@gmail.com, Eric Luis Uhlmann eric.luis.uhlmann@gmail.com, Brian Nosek nosek@virginia.edu, Dan Martin dpmartin42@gmail.com

---

Dear all,

to get started, I just quote what we (Team 5) said in our report:

We did not use any covariates such as player position and decided to stick with this approach even though reviewers of our approach suggested that we should do so. As already noted in the project description by Silberzahn, Martin, Uhlmann, & Nosek, the data cannot be used for causal inference. Thus, if the goal is to come up with a generalizable *descriptive* statement (i.e., effect size), it does not matter why a player ends up getting more red cards (e.g., being a tall, heavy defense player). In fact, such information when included as covariates might even bias the result. For example, it is unclear what happens when you include „League Country“ because the data span a player’s entire career, often spent in multiple leagues, but the entry for „League Country“ is only for the league the player was in during the year of data retrieval.

Best regards,

Johannes

--
University of Zurich
Prof. Dr. Johannes Ullrich
Department of Psychology
Social Psychology
Binzmühlestrasse 14 / 15
CH-8050 Zürich
Switzerland
Tel. +41 44 635 72 76
Fax +41 44 635 72 79
Website: t.uzh.ch/10

> Am 11.02.2016 um 12:20 schrieb Silberzahn, Raphael <RSilberzahn@iese.edu>:
>
> Dear all,
>
> I am writing you regarding our Crowdsourcing project. You indicated that you’d be willing to help in an assessment of additional aspects of the dataset - namely the covariates.
> In addition to the peer review regarding the feasibility of approaches, this analysis is important to address reviewers’ concerns.
>
> What I would like to do is to initiate an e-mail discussion about the appropriateness of the following covariates in the context of our data. There are two aspects:
> 1. Is it justified theoretically to include the variable.
> 2. Is the data in our dataset accurate enough to capture the theoretical element.
>
> We will exchange our different opinions. Feel free to actively share your thoughts and/or follow others’ comments. Finally you will receive a short survey to capture your opinion following this discussion. As far as I can see this will be the last element in our peer review before we are ready to re-write our manuscript. I look forward to hearing your thoughts! Please use the reply-all button to get started.
>
> All the best, Raphael
>
> Age
> Club
> Defeats

From: **Garret S. Christensen** garret@berkeley.edu
Subject: Re: CS1: Covariate Discussion
Date: 11 Feb 2016 19:18
To: Johannes Ullrich j.ullrich@psychologie.uzh.ch
Cc: Silberzahn, Raphael RSilberzahn@iese.edu, ssyoon@temple.edu, Rickard Carlsson rickard.carlsson@lnu.se, Christoph Spörlein ch.spoerlein@gmail.com, Erikson Kaszubowski erikson84@yahoo.com.br, molden@northwestern.edu, Alicia Hofelich Mohr hofelich@umn.edu, mathew.evans@manchester.ac.uk, b.a.nijstad@rug.nl, Tom Stafford t.stafford@sheffield.ac.uk, Michelangelo Vianello michelangelo.vianello@unipd.it, Kent Hui huingan@msu.edu, dkennedy@uwb.edu, feng.bai10@rotman.utoronto.ca, fabia.hoegden@uni-koeln.de, Frederik Aust frederik.aust@uni-koeln.de, eawtrey@uw.edu, SOMMERa@hec.fr, egidio.robusto@unipd.it, brysonrpope@gmail.com, Eric Luis Uhlmann eric.luis.uhlmann@gmail.com, Brian Nosek nosek@virginia.edu, Dan Martin dpmartin42@gmail.com

---

All,
I think most, if not all, of these variables, including the much-maligned "League Country" variable, should probably be included, but that ultimately we can't be sure given the way the question was framed. Especially club, height, position, and weight seem like they could obviously be correlated with both skin tone and red cards, and thus there would be omitted variable bias if the variable were not included. I could probably come up with a similar story for all the other variables as well. But that's the economist in me with my emphasis on causal inference, and the data isn't really capable of answering a causal question.

It was clearly stated upfront that we wouldn't be answering a causal question, but then it seems to me like we all still tried to do a causal exercise regardless. I found PNAS Reviewer #1 fairly convincing in this regard. (See the 3/9/15 e-mail from Raphael with subject line "Our Crowdsourcing Paper - Update")

The bits from Reviewer #1 I find the most convincing are:

"What really surprised me was the absence of model comparison. Here you have 29 models that could be compared to each other and identify which model seems to do best. This comparison could be done with information criterion methods (AIC, BIC, DIC) or cross-validation methods (build the model on a subset of the data and then test on a holdout set). Not doing this analysis is a missed opportunity."

And then his/her long suggestion on how to reframe the question as a prediction contest with direct model comparison, which I will paste below. Basically, I think there's not much to be gained from debating covariate inclusion if we're not going to directly compare the prediction value of the models, but if we're still in our sort-of-but-not-quite causal framework, I lean towards including more rather than less.

Best,
Garret

PNAS Reviewer #1 Comments:

To be more constructive, I thought I would describe how I think the basic idea of the project could be implemented in a way that makes sense.

1. Frame the research questions in terms of predicting outcomes relative to some cost function. For example, if the prediction is whether a player-referee dyad will result in a red card during a match (X: 1-yes, 0-no), then the cost function might be

C = sum_players/refs sum_games (X-Y)^2

where Y is what actually happened in the data set (e.g., whether the ref gave a red card to the player in a game: 1-yes, 0-no), you can divide C by the number of data points to get Mean Squared Error and take the square root to get Root Mean Squared Error (there are many other cost functions of course, and maybe this is not the best one for this type of data, but the flavor is the same).

If the task is framed this way, then the research questions become much more direct. For example, the original research questions might be reframed as:

Research Question 1: Does including skin tone improve model performance?
Research Question 2: Does including information about skin-tone prejudice for a referee's country improve model performance?

It seems to me that this is how many teams implicitly interpreted the original research questions. An even better research task would be:

Research Task: Use this data set to build a model to predict red card assignment.

Then, any effects of skin tone or skin-tone prejudice appear automatically, rather then being the focus of the project. I think developing such models would be quite challenging because many variables in the data set are (probably) correlated, red cards are rather rare (most matches have none), and there are all sorts of complicated dependencies (e.g., if one player in a match gets a red card, his teammates will play more carefully so as to not get another one). My guess is that none of the models will do very well because the information simply is not in the data set.

2. Analyses might proceed largely as before, with a focus on the resulting model rather than on statistical significance. Various types of regression are perfectly good ways to develop a model. However, one is not just looking for the best fit to the data (the smallest C cost function) because there is the risk of overfitting the data by including lots of covariates. Instead, you need to estimate C for data that was not used to build the model.

One way of doing this is cross validation. For example, divide the data set into 10 equal subsets. Build the model on 9 of the subsets and then compute C for the remaining subset. Repeat this process by rotating which subset is used to compute C. You now have 10 estimates

of C, and an average should give an estimate of what C would be for a new data set (there are many assumptions here about the representativeness of the dataset, but that is always true). The details of such a cross-validation approach could be complicated because of the hierarchical structure of the data set (which makes it difficult to divide it into 10 equal subsets). There are Bayesian methods that get to roughly the same thing from a different perspective. The motivation for any of these methods is that an overfit model (that uses covariates that happen to do well for the training data by chance) will do poorly at predicting data that were not part of the model-construction data set.

Note that statistical significance is not part of the analysis. Covariates that are not statistically significant might still contribute to model prediction; and covariates that are statistically significant might not contribute to model prediction.

I should mention that several of the teams did something similar to what I just described (using AIC, for example, instead of cross validation). That is nice work, it's just not what was asked for the project.

On Thu, Feb 11, 2016 at 4:11 AM, Johannes Ullrich <j.ullrich@psychologie.uzh.ch> wrote:

> Dear all,
>
> to get started, I just quote what we (Team 5) said in our report:
>
> We did not use any covariates such as player position and decided to stick with this approach even though reviewers of our approach suggested that we should do so. As already noted in the project description by Silberzahn, Martin, Uhlmann, & Nosek, the data cannot be used for causal inference. Thus, if the goal is to come up with a generalizable *descriptive* statement (i.e., effect size), it does not matter why a player ends up getting more red cards (e.g., being a tall, heavy defense player). In fact, such information when included as covariates might even bias the result. For example, it is unclear what happens when you include „League Country“ because the data span a player’s entire career, often spent in multiple leagues, but the entry for „League Country“ is only for the league the player was in during the year of data retrieval.
>
> Best regards,
>
> Johannes
>
> --
> University of Zurich
> Prof. Dr. Johannes Ullrich
> Department of Psychology
> Social Psychology
> Binzmühlestrasse 14 / 15
> CH-8050 Zürich
> Switzerland
> Tel. +41 44 635 72 76
> Fax +41 44 635 72 79
> Website: t.uzh.ch/10

> Am 11.02.2016 um 12:20 schrieb Silberzahn, Raphael <RSilberzahn@iese.edu>:
>
> Dear all,
>
> I am writing you regarding our Crowdsourcing project. You indicated that you’d be willing to help in an assessment of additional aspects of the dataset - namely the covariates.
> In addition to the peer review regarding the feasibility of approaches, this analysis is important to address reviewers’ concerns.
>
> What I would like to do is to initiate an e-mail discussion about the appropriateness of the following covariates in the context of our data. There are two aspects:
> 1. Is it justified theoretically to include the variable.
> 2. Is the data in our dataset accurate enough to capture the theoretical element.
>
> We will exchange our different opinions. Feel free to actively share your thoughts and/or follow others’ comments. Finally you will receive a short survey to capture your opinion following this discussion. As far as I can see this will be the last element in our peer review before we are ready to re-write our manuscript. I look forward to hearing your thoughts! Please use the reply-all button to get started.

From: **Alicia Hofelich Mohr** hofelich@umn.edu
Subject: Re: CS1: Covariate Discussion
Date: 23 Feb 2016 15:54
To: Garret S. Christensen garret@berkeley.edu
Cc: Johannes Ullrich j.ullrich@psychologie.uzh.ch, Silberzahn, Raphael RSilberzahn@iese.edu, ssyoon@temple.edu, Rickard Carlsson rickard.carlsson@lnu.se, Christoph Spörlein ch.spoerlein@gmail.com, Erikson Kaszubowski erikson84@yahoo.com.br, molden@northwestern.edu, mathew.evans@manchester.ac.uk, b.a.nijstad@rug.nl, Tom Stafford t.stafford@sheffield.ac.uk, Michelangelo Vianello michelangelo.vianello@unipd.it, Kent Hui huingan@msu.edu, dkennedy@uwb.edu, feng.bai10@rotman.utoronto.ca, fabia.hoegden@uni-koeln.de, Frederik Aust frederik.aust@uni-koeln.de, eawtrey@uw.edu, SOMMERa@hec.fr, egidio.robusto@unipd.it, brysonrpope@gmail.com, Eric Luis Uhlmann eric.luis.uhlmann@gmail.com, Brian Nosek nosek@virginia.edu, Dan Martin dpmartin42@gmail.com

---

Hi all,

Apologies for coming late to this conversation.

Our team took a theoretical approach to choosing covariates, selecting only the ones that we thought would likely limit or enhance opportunity for red cards. We took this approach because we were interested in controlling for as many confounding factors as we could, but ultimately, our goal was to test whether skintone was a reliable predictor, rather than creating a model that overall explained the most variance in getting red cards. Therefore, including and excluding variables until we got a model with the lowest information criterion wasn't our overall goal. This is especially true in the sense that many of the variables were not ideal measures of what they represented, and could bias the results, as Johannes said (like the League Country). Further, they could also be correlated with skin tone, and since the regression coefficient of skintone was what we were interested in, we thought that adding many potentially collinear measures would not be ideal.

So we ended up including player (as individuals differ in their tendencies to display behavior that may elicit red cards), referee (as individuals differ in their thresholds at which they give a red card), games (because this directly relates to opportunities for getting a red card), and position (as some people the field, such as the goal keeper, maybe inherently less likely to receive red cards - again, it seemed related to opportunity for getting a red card).

Best,
Alicia

---------- Forwarded message ----------
From: **Garret S. Christensen** <garret@berkeley.edu>
Date: Thu, Feb 11, 2016 at 12:18 PM
Subject: Re: CS1: Covariate Discussion
To: Johannes Ullrich <j.ullrich@psychologie.uzh.ch>
Cc: "Silberzahn, Raphael" <RSilberzahn@iese.edu>, "ssyoon@temple.edu" <ssyoon@temple.edu>, Rickard Carlsson <rickard.carlsson@lnu.se>, Christoph Spörlein <ch.spoerlein@gmail.com>, Erikson Kaszubowski <erikson84@yahoo.com.br>, "molden@northwestern.edu" <molden@northwestern.edu>, Alicia Hofelich Mohr <hofelich@umn.edu>, "mathew.evans@manchester.ac.uk" <mathew.evans@manchester.ac.uk>, "b.a.nijstad@rug.nl" <b.a.nijstad@rug.nl>, Tom Stafford <t.stafford@sheffield.ac.uk>, Michelangelo Vianello <michelangelo.vianello@unipd.it>, Kent Hui <huingan@msu.edu>, "dkennedy@uwb.edu" <dkennedy@uwb.edu>, "feng.bai10@rotman.utoronto.ca" <feng.bai10@rotman.utoronto.ca>, "fabia.hoegden@uni-koeln.de" <fabia.hoegden@uni-koeln.de>, Frederik Aust <frederik.aust@uni-koeln.de>, "eawtrey@uw.edu" <eawtrey@uw.edu>, "SOMMERa@hec.fr" <SOMMERa@hec.fr>, "egidio.robusto@unipd.it" <egidio.robusto@unipd.it>, "brysonrpope@gmail.com" <brysonrpope@gmail.com>, Eric Luis Uhlmann <eric.luis.uhlmann@gmail.com>, Brian Nosek <nosek@virginia.edu>, Dan Martin <dpmartin42@gmail.com>

All,
I think most, if not all, of these variables, including the much-maligned "League Country" variable, should probably be included, but that ultimately we can't be sure given the way the question was framed. Especially club, height, position, and weight seem like they could obviously be correlated with both skin tone and red cards, and thus there would be omitted variable bias if the variable were not included. I could probably come up with a similar story for all the other variables as well. But that's the economist in me with my emphasis on causal inference, and the data isn't really capable of answering a causal question.

It was clearly stated upfront that we wouldn't be answering a causal question, but then it seems to me like we all still tried to do a causal exercise regardless. I found PNAS Reviewer #1 fairly convincing in this regard. (See the 3/9/15 e-mail from Raphael with subject line "Our Crowdsourcing Paper - Update")

The bits from Reviewer #1 I find the most convincing are:

"What really surprised me was the absence of model comparison. Here you have 29 models that could be compared to each other and identify which model seems to do best. This comparison could be done with information criterion methods (AIC, BIC, DIC) or cross-validation methods (build the model on a subset of the data and then test on a holdout set). Not doing this analysis is a missed opportunity."

And then his/her long suggestion on how to reframe the question as a prediction contest with direct model comparison, which I will paste below. Basically, I think there's not much to be gained from debating covariate inclusion if we're not going to directly compare the prediction value of the models, but if we're still in our sort-of-but-not-quite causal framework, I lean towards including more rather than less.

Best,
Garret

From: **Daniel C. Molden** molden@northwestern.edu
Subject: Re: CS1: Covariate Discussion
Date: 23 Feb 2016 21:45
To: Silberzahn, Raphael RSilberzahn@iese.edu, ssyoon@temple.edu, Rickard Carlsson rickard.carlsson@lnu.se, Christoph Spörlein ch.spoerlein@gmail.com, Johannes Ullrich j.ullrich@psychologie.uzh.ch, Erikson Kaszubowski erikson84@yahoo.com.br, Alicia Hofelich Mohr hofelich@umn.edu, Garret S. Christensen garret@berkeley.edu, mathew.evans@manchester.ac.uk, b.a.nijstad@rug.nl, Tom Stafford t.stafford@sheffield.ac.uk, Michelangelo Vianello michelangelo.vianello@unipd.it, Kent Hui huingan@msu.edu, dkennedy@uwb.edu, feng.bai10@rotman.utoronto.ca, fabia.hoegden@uni-koeln.de, Frederik Aust frederik.aust@uni-koeln.de, eawtrey@uw.edu, SOMMERa@hec.fr, egidio.robusto@unipd.it, brysonrpope@gmail.com
Cc: Eric Luis Uhlmann eric.luis.uhlmann@gmail.com, Brian Nosek nosek@virginia.edu, Dan Martin dpmartin42@gmail.com

---

First, let me say that I find the position that one should not perform covariate analyses at all because causal conclusions with this data set are not possible to be completely baffling. Even if one were only interested in an overall effect size, this estimate could be wildly skewed by confounds in the data. If, say, defenders are more likely to receive red cards in general because they are more often in a position to commit hard fouls to prevent goals, and it just happens that in the base rates of the data set, darker-skinned players are more likely to be defenders, then an unbiased effect-size for skin tone cannot be properly estimated. So, even if one acknowledges that no causal conclusion is possible, it is still necessary to evaluate and control for possible confounds.

To be a confound, a variable must both (a) correlate with skin-tone ratings in the data set, and (b) predict the likelihood of receiving a red card. If either one is not true then this variable can not be obscuring or altering the "true" effect size of skin tone in the data set and does not need to be included. For condition (a), skin-tone ratings were significantly correlated with position (i.e. dark-skinned players are not equally distributed among the positions) and with game outcome (i.e., dark-skinned players were more likely to be on teams that lost or drew rather than won the game). For condition (b), position did significantly predict the likelihood of receiving a red-card and fewer red cards occurred in wins and draws as compared to losses. Therefore, position and game outcome are both potential confounds that could obscure and bias any estimate of the association between skin tone and red cards and should be included as covariates (see the Team 10 report for the details of these analyses). This is true regardless of whether one desires a "descriptive" estimate of the effect size or whether one is interested in potential causal mechanisms.

Now, if one further wanted to ask not just whether an association exists between skin tone and red cards (and how big this association is) but also how this association compares to all of the other factors that predict red cards, that would require a different approach. In this case one would want to build a model that explains the most variance by including all possible predictors of red cards and then eliminating those that are not significant to maximize some index like BIC. But, building an optimal model of red cards was not the primary purpose of the analysis featured in this project, and so this approach would not be optimal for estimating the basic effect size for skin color, as was the stated objective.

Regardless of whether one just wants to control for confounds or to build an optimally predictive model, It is also important to recognize that, from the list of covariates circulated and included with the data set, not all of them are appropriate to use in either of these analysis. The data set does not just involve a single season of data for the full set of players. Instead it takes all of the players from a certain year and then follows them backwards throughout their careers. So, for some players there may only be a single season, but for most there are many years of data across a wide span of time. However, despite this fact, much of the player-level data only has a single value that does not reflect this extended time span and results in misclassification errors for some data points when the data set is analyzed altogether.

Some of these misclassifications are likely to be uncommon and trivial - players do not often change positions throughout their career, nor is their height likely to vary (and to a lesser extent BMI, but it's possible that weight varies enough to make this problematic). In these cases, the variables can likely be safely treated as static values across the different years in the data set without introducing any nontrivial bias into the analysis.

However, other misclassifications are likely to be widespread. Because a large proportion of players exist in the data set across multiple years, player age (i.e., birthdate) cannot be treated as a static value to be extrapolated over all of the games featuring that player. In addition, because there is a high degree of movement of players between teams and even between leagues and countries (literally hundreds of players changing teams and leagues each year see http://www.soccernews.com/soccer-transfers/english-premier-league-transfers-2011-2012/; http://www.soccernews.com/soccer-transfers/spanish-la-liga-transfers-2011-2012/; http://www.soccernews.com/soccer-transfers/rest-of-europe-transfers-2011-2012/ for the single season before the one from which the values were drawn in the data set) it is impossible to extrapolate these values in the data set across multiple years with ANY degree of accuracy. That is, given the high number of players who are tracked across multiple years, there are thousands of cases in the data set in which the data collected concerning a particular game with a particular player-referee pair is assigned an incorrect value regarding the player's team and league. Therefore, these variables are invalid for analysis and should never be used as covariates for any application.

In summary, the following covariates can legitimately be considered: (a) all of those measured at the level of the individual game (goals, victories, draws, defeats, referee, referee country, number of games the player and referee were paired), (b) those that are not perfectly static across the multiple years in the data set, but are virtually unchanging (height, position; BMI or weight might be a matter for legitimate discussion). However, the remaining covariates (Club, League Country) are too variable across the multiple years and therefore can not be legitimately considered in the analysis because they are misspecified for a high proportion of the data points and would introduce bias and error into the analyses.

-DM

On 2/11/16 5:20 AM, Silberzahn, Raphael wrote:

> Dear all,
>
> I am writing you regarding our Crowdsourcing project. You indicated that you'd be willing to help in an assessment of additional aspects

From: **Johannes Ullrich** j.ullrich@psychologie.uzh.ch
Subject: Re: CS1: Covariate Discussion
Date: 25 Feb 2016 14:14
To: Daniel C. Molden molden@northwestern.edu
Cc: Silberzahn, Raphael RSilberzahn@iese.edu, ssyoon@temple.edu, Rickard Carlsson rickard.carlsson@lnu.se, Christoph Spörlein ch.spoerlein@gmail.com, Erikson Kaszubowski erikson84@yahoo.com.br, Alicia Hofelich Mohr hofelich@umn.edu, Garret S. Christensen garret@berkeley.edu, mathew.evans@manchester.ac.uk, b.a.nijstad@rug.nl, Tom Stafford t.stafford@sheffield.ac.uk, Michelangelo Vianello michelangelo.vianello@unipd.it, Kent Hui huingan@msu.edu, dkennedy@uwb.edu, feng.bai10@rotman.utoronto.ca, fabia.hoegden@uni-koeln.de, Frederik Aust frederik.aust@uni-koeln.de, eawtrey@uw.edu, SOMMERa@hec.fr, egidio.robusto@unipd.it, brysonrpope@gmail.com, Eric Luis Uhlmann eric.luis.uhlmann@gmail.com, Brian Nosek nosek@virginia.edu, Dan Martin dpmartin42@gmail.com

---

I agree with Daniel's classification of covariates that do not make any sense, but I disagree that it is necessary to include position or game outcome. Consider the case of the insurance company learning that a possible new customer takes antibiotics, which predicts problems for the company (see quote below for context). That statistical relationship would be a biased estimate of the true causal effect, but that's not what the insurance company is interested in at the moment. They're happy with predicting what the customer would cost them based on the information about antibiotics. It's not even clear what „biased" would mean in this predictive context.

Best,
Johannes

"The difference between correlation and causation
is the difference between prediction and
control. Both are useful concepts, but they lead
to different uses and they can appear to be opposites.
For example, a correlational survey would
find that people who received antibiotics last
year are more likely to be dead this year than
people who received no antibiotics last year. So
we can use antibiotics to predict death.However,
by means of an experiment, we can randomly
assign people with infections to two groups, one
that receives antibiotics and the other receives a
placebo. The results of such studies show that
antibiotics cause a reduction in the death rate.
So, we find that receiving antibiotics is positively
correlated with death in surveys and receiving
antibiotics is negatively correlated with death in
an experiment. Although paradoxical, there is
no contradiction.
Both correlational and causal relations are
interesting and useful, even when they seem to
say the opposite things. Suppose you have a life
insurance company; you sell insurance that pays
out when a person dies. Before you sell someone
insurance, you could ask if they have been
taking antibiotics. If yes, you do not want to sell
them insurance because they are likely to die.
However, if you already sold a policy to a client
and that person becomes sick, you would like
them to take antibiotics because it causes a reduction
in the death rate."

(Birnbaum, 2007, Designing online experiments)

--
University of Zurich
Prof. Dr. Johannes Ullrich
Department of Psychology
Social Psychology
Binzmühlestrasse 14 / 15
CH-8050 Zürich
Switzerland
Tel. +41 44 635 72 76
Fax +41 44 635 72 79
Website: t.uzh.ch/10

> Am 23.02.2016 um 21:45 schrieb Daniel C. Molden <molden@northwestern.edu>:
>
> First, let me say that I find the position that one should not perform covariate analyses at all because causal conclusions with this data

From: **Rickard Carlsson** rickard.carlsson@lnu.se
Subject: Re: CS1: Covariate Discussion
Date: 25 Feb 2016 14:52
To: Johannes Ullrich j.ullrich@psychologie.uzh.ch
Cc: Daniel C. Molden molden@northwestern.edu, Silberzahn, Raphael RSilberzahn@iese.edu, ssyoon@temple.edu, Christoph Spörlein ch.spoerlein@gmail.com, Erikson Kaszubowski erikson84@yahoo.com.br, Alicia Hofelich Mohr hofelich@umn.edu, Garret S. Christensen garret@berkeley.edu, mathew.evans@manchester.ac.uk, b.a.nijstad@rug.nl, Tom Stafford t.stafford@sheffield.ac.uk, Michelangelo Vianello michelangelo.vianello@unipd.it, Kent Hui huingan@msu.edu, dkennedy@uwb.edu, feng.bai10@rotman.utoronto.ca, fabia.hoegden@uni-koeln.de, Frederik Aust frederik.aust@uni-koeln.de, eawtrey@uw.edu, SOMMERa@hec.fr, egidio.robusto@unipd.it, brysonrpope@gmail.com, Eric Luis Uhlmann eric.luis.uhlmann@gmail.com, Brian Nosek nosek@virginia.edu, Dan Martin dpmartin42@gmail.com

---

Hi,

Interesting discussion so far. I think some of the different approaches might reflect different conceptualization of discrimination, bias etc. If so then it's not a statistical disagreement but rather a theoretical one.

Some (like me) focus on direcru(disparate treatment) discrimination and then it becomes very important to establish causality. Even when it's not possible, one would try as hard as possible to equate players of different skin tones. If differences remain, one would take this as evidence of discrimination. The strength of that evidence hinges upon whether the difference could be explained by something else except discrimination.

However, others are more focused on the societal level impact and if a group is in a relatively worse position then this is a finding by itself. As long as the sample is unbiased, things such as player position etc., would not be confounds, but rather mediators.

Take the labor market as an example. If skin tone is associated with less pay, but this effect disappears when controlling for education etc., then we have no evidence of direct discrimination, but we have a main result of inequality that warrants further investigation.

The question is then if the red cards has inherent value (like salaries) so that a main result of inequality is interesting. If so a model free of covariates might be interesting. However, even if a model without covariates has some merit from this perspective , I strongly believe that it has to be accompanied by a full model. The "confounds" works in both ways and we might have evidence of direct discrimination even without a difference without covariates.

Best,
Rickard Carlsson

25 feb. 2016 kl. 14:14 skrev Johannes Ullrich <j.ullrich@psychologie.uzh.ch>:

> I agree with Daniel's classification of covariates that do not make any sense, but I disagree that it is necessary to include position or game outcome. Consider the case of the insurance company learning that a possible new customer takes antibiotics, which predicts problems for the company (see quote below for context). That statistical relationship would be a biased estimate of the true causal effect, but that's not what the insurance company is interested in at the moment. They're happy with predicting what the customer would cost them based on the information about antibiotics. It's not even clear what „biased" would mean in this predictive context.
>
> Best,
> Johannes
>
> "The difference between correlation and causation
> is the difference between prediction and
> control. Both are useful concepts, but they lead
> to different uses and they can appear to be opposites.
> For example, a correlational survey would
> find that people who received antibiotics last
> year are more likely to be dead this year than
> people who received no antibiotics last year. So
> we can use antibiotics to predict death.However,
> by means of an experiment, we can randomly
> assign people with infections to two groups, one
> that receives antibiotics and the other receives a
> placebo. The results of such studies show that
> antibiotics cause a reduction in the death rate.
> So, we find that receiving antibiotics is positively
> correlated with death in surveys and receiving
> antibiotics is negatively correlated with death in
> an experiment. Although paradoxical, there is
> no contradiction.
> Both correlational and causal relations are
> interesting and useful, even when they seem to
> say the opposite things. Suppose you have a life
> insurance company; you sell insurance that pays
> out when a person dies. Before you sell someone
> insurance, you could ask if they have been
> taking antibiotics. If yes, you do not want to sell
> them insurance because they are likely to die.
> However, if you already sold a policy to a client

From: **Daniel C. Molden** molden@northwestern.edu
Subject: Re: CS1: Covariate Discussion
Date: 3 Mar 2016 18:21
To: Rickard Carlsson rickard.carlsson@lnu.se, Johannes Ullrich j.ullrich@psychologie.uzh.ch
Cc: Silberzahn, Raphael RSilberzahn@iese.edu, ssyoon@temple.edu, Christoph Spörlein ch.spoerlein@gmail.com, Erikson Kaszubowski erikson84@yahoo.com.br, Alicia Hofelich Mohr hofelich@umn.edu, Garret S. Christensen garret@berkeley.edu, mathew.evans@manchester.ac.uk, b.a.nijstad@rug.nl, Tom Stafford t.stafford@sheffield.ac.uk, Michelangelo Vianello michelangelo.vianello@unipd.it, Kent Hui huingan@msu.edu, dkennedy@uwb.edu, feng.bai10@rotman.utoronto.ca, fabia.hoegden@uni-koeln.de, Frederik Aust frederik.aust@uni-koeln.de, eawtrey@uw.edu, SOMMERa@hec.fr, egidio.robusto@unipd.it, brysonrpope@gmail.com, Eric Luis Uhlmann eric.luis.uhlmann@gmail.com, Brian Nosek nosek@virginia.edu, Dan Martin dpmartin42@gmail.com

---

I appreciate Johannes's point, and I think Rickard has clearly stated the difference between the no-covariate vs. controlling for confounds perspectives. So, this discussion then boils down to whether the present research question on skin tone fits more of the actuarial question of whether any relationship exists at all or the causal question of whether darker skin tone invites discrimination. If you read the original proposal document and the literature cited, it seems like the latter causal hypothesis is what was intended in this case (even without the further analysis of implicit and explicit bias). I would further argue that the former actuarial approach does not make sense in this context. Because red cards are not an outcome of any particular inherent value for an owner, or manager, or observer to predict (as the data shows they are rare occurrences and so their overall impact on team performance will be dwarfed by other outcomes and attributes such as goals scored or passes completed) knowing that they are predicted by skin tone without any further context that would shed some light on why that connection exists does not provide valuable information. So, while I do not disagree with Birnbaum (2007), I also do not think that analysis can apply to all cases and that the red card/skin tone association under examination here is not analogous to the insurance company situation outline below.

-DM

On 2/25/16 7:52 AM, Rickard Carlsson wrote:

> Hi,
>
> Interesting discussion so far. I think some of the different approaches might reflect different conceptualization of discrimination, bias etc. If so then it's not a statistical disagreement but rather a theoretical one.
>
> Some (like me) focus on direcru(disparate treatment) discrimination and then it becomes very important to establish causality. Even when it's not possible, one would try as hard as possible to equate players of different skin tones. If differences remain, one would take this as evidence of discrimination. The strength of that evidence hinges upon whether the difference could be explained by something else except discrimination.
>
> However, others are more focused on the societal level impact and if a group is in a relatively worse position then this is a finding by itself. As long as the sample is unbiased, things such as player position etc., would not be confounds, but rather mediators.
>
> Take the labor market as an example. If skin tone is associated with less pay, but this effect disappears when controlling for education etc., then we have no evidence of direct discrimination, but we have a main result of inequality that warrants further investigation.
>
> The question is then if the red cards has inherent value (like salaries) so that a main result of inequality is interesting. If so a model free of covariates might be interesting. However, even if a model without covariates has some merit from this perspective , I strongly believe that it has to be accompanied by a full model. The "confounds" works in both ways and we might have evidence of direct discrimination even without a difference without covariates.
>
> Best,
> Rickard Carlsson
>
> 25 feb. 2016 kl. 14:14 skrev Johannes Ullrich <j.ullrich@psychologie.uzh.ch>:
>
>> I agree with Daniel's classification of covariates that do not make any sense, but I disagree that it is necessary to include position or game outcome. Consider the case of the insurance company learning that a possible new customer takes antibiotics, which predicts problems for the company (see quote below for context). That statistical relationship would be a biased estimate of the true causal effect, but that's not what the insurance company is interested in at the moment. They're happy with predicting what the customer would cost them based on the information about antibiotics. It's not even clear what „biased" would mean in this predictive context.
>>
>> Best,
>> Johannes
>>
>> "The difference between correlation and causation
>> is the difference between prediction and
>> control. Both are useful concepts, but they lead
>> to different uses and they can appear to be opposites.
>> For example, a correlational survey would
>> find that people who received antibiotics last
>> year are more likely to be dead this year than
>> people who received no antibiotics last year. So
>> we can use antibiotics to predict death.However,
>> by means of an experiment, we can randomly
>> assign people with infections to two groups, one
>> that receives antibiotics and the other receives a
>> placebo. The results of such studies show that

From: **Frederik Aust** frederik.aust@uni-koeln.de
Subject: Re: CS1: Covariate Discussion
Date: 9 Mar 2016 16:51
To: Daniel C. Molden molden@northwestern.edu, Rickard Carlsson rickard.carlsson@lnu.se, Johannes Ullrich j.ullrich@psychologie.uzh.ch
Cc: Silberzahn, Raphael RSilberzahn@iese.edu, ssyoon@temple.edu, Christoph Spörlein ch.spoerlein@gmail.com, Erikson Kaszubowski erikson84@yahoo.com.br, Alicia Hofelich Mohr hofelich@umn.edu, Garret S. Christensen garret@berkeley.edu, mathew.evans@manchester.ac.uk, b.a.nijstad@rug.nl, Tom Stafford t.stafford@sheffield.ac.uk, Michelangelo Vianello michelangelo.vianello@unipd.it, Kent Hui huingan@msu.edu, dkennedy@uwb.edu, feng.bai10@rotman.utoronto.ca, fabia.hoegden@uni-koeln.de, eawtrey@uw.edu, SOMMERa@hec.fr, egidio.robusto@unipd.it, brysonrpope@gmail.com, Eric Luis Uhlmann eric.luis.uhlmann@gmail.com, Brian Nosek nosek@virginia.edu, Dan Martin dpmartin42@gmail.com

---

Hi,

from a purely theoretical perspective, I think Rickard Carlsson has pinpointed the distinction between the two general approaches and I agree with him and Daniel Molden that the inclusion of covariates is merited. I also think that there are probably good arguments to be made for the inclusion of any available covariate.

From a practical perspective, I generally agree with Daniel Molden's classification of the aptness of the covariates in our dataset, but I think it is important to note that the degree of noise in the covariates is unknown. The discussion about justifications to use covariates appears to be based on varying combinations of intuition, informed guessing, and data; this subjectivity is, at least in part, reflected by the selection of covariates in the submitted analyses.

As a case in point, Daniel Molden argued player position is a legitimate covariate because it is virtually unchanged throughout a player's career. While I agree (our team used this covariate), I think it could just as well be argued that player position is a problematic covariate. Firstly, there are several prominent examples of professional soccer players that have changed positions throughout their career. Some of these changes may be minor but frequent (e.g., from left to right midfielder) or major and less frequent (e.g., defender to midfielder, midfielder to attacker).* Secondly, position labels appear to be only loosely related to player behaviour and tactics on the field (http://www.americansocceranalysis.com/home/2015/5/4/everything-you-think-you-know-about-player-positions-is-wrong).

In other words, I fear that the question about aptness of the covariates in our dataset will be difficult to resolve through rhetoric rather than a review of the literature on, or a formal analyses of, the prevalence of position changes, league countries, weight gains and losses, etc. among soccer players.

It appears this discussion is motivated by a concern that either including or not including (specific) covariates may bias effect size estimates. Another approach to address this concern could then be to compare estimates of both models.

In response to the first covariate discussion, we reanalysed the data without nesting players in league countries but we also ran a model without any covariates (we did keep crossed random intercepts for players and referees). The estimate changed but not much:

Round 2 model: 1.382 [1.120, 1.705]
Without league country: 1.405 [1.141, 1.731]
No covariates: 1.346 [1.082, 1.674]

I realize that these results do not generalize to other approaches but, in the spirit of "actively sharing thoughts", I wanted to add this bit of our analyses to the discussion.

Best regards,
Frederik Aust

*We did include the player position covariate in our analyses and tried to mitigate the noise in the player position covariate by collapsing the levels to more general positions.

Am 03.03.2016 um 18:21 schrieb Daniel C. Molden:

> I appreciate Johannes's point, and I think Rickard has clearly stated the difference between the no-covariate vs. controlling for confounds perspectives. So, this discussion then boils down to whether the present research question on skin tone fits more of the actuarial question of whether any relationship exists at all or the causal question of whether darker skin tone invites discrimination. If you read the original proposal document and the literature cited, it seems like the latter causal hypothesis is what was intended in this case (even without the further analysis of implicit and explicit bias). I would further argue that the former actuarial approach does not make sense in this context. Because red cards are not an outcome of any particular inherent value for an owner, or manager, or observer to predict (as the data shows they are rare occurrences and so their overall impact on team performance will be dwarfed by other outcomes and attributes such as goals scored or passes completed) knowing that they are predicted by skin tone without any further context that would shed some light on why that connection exists does not provide valuable