Data Science Looks at Discrimination

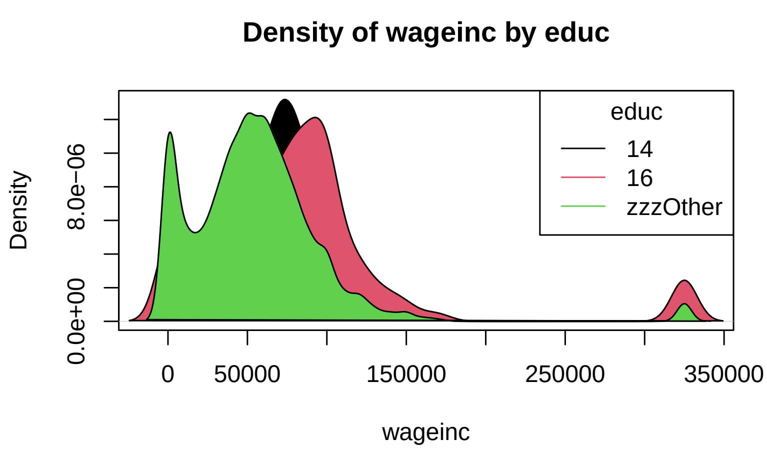
A toolkit for investigating bias in race, gender, age and so on

Taha Abdullah, Arjun Ashok
Brandon Estrada, Shubhada Martha
Norman Matloff, Aditya Mittal
Billy Ouattara, Jonathan Tran

2023-08-02

Table of contents

Overview



Discrimination is a key social issue in the US and in a number of other countries. There is lots of available data with which one might investigate possible discrimination. But how might such investigations be conducted?

Our **dsld** package provides both graphical and analytical tools for this purpose. We see it as widely applicable; here are just a few use cases:

- Quantitative analysis in instruction and research in social science.
- Corporate HR analysis and research.
- Litigation involving discrimination and related issues.
- Concerned citizenry.

This document provides a tutorial regarding applicable methodology, as well as introduction to use of the package.

Prerequisite background

In addition to having rudimentary skill in R, the user should have a very basic knowledge of statistical inference–mean, variance, confidence intervals and tests, and histograms. A "bare bones" refresher, with emphasis on intuition, is given in Appendix A.

Python wrappers are included for most functions.

The dsld package

The **dsld** package, which this tutorial uses for examples, has two aims:

- To enable exploratory analysis of possible discrimination effects through various graphical and tabular functions.
- To enable formal statistical analysis of such effects via addition of a number of group-comparison operations to general R functions such as **lm()** and **glm()**, thereby facilitating comparisons across races, genders and so on.

Introduction and Motivating Examples

To set the stage, consider the following:

UC Berkeley discrimination claims

UC Berkeley was accused of discriminating against female applicants for graduate school, and indeed the overall acceptance rate for women was lower than that for men. This seemed odd, given Berkeley's liberal reputation.

However, upon breaking the data down according to the program students were applying to, it was found that in every department, the female acceptance rate *within that department* was either higher than the male rate or of similar level. The problem: women were applying to more selective programs, causing their overall rate to below that of men.

This data is included in R, as the built-in dataset **UCBAdmissions**.

US Census data

The **svcensus** dataset is a subset of US census data from back in 2000, focusing on six engineering occupations. The question at hand is whether there is a gender pay gap. Again, the overall pay for men is higher, by about 25%. But what if we break things down by occupation? Though it does turn out that some occupations pay more than others, and that men and women are not distributed evenly among the occupations, there still is a gender pay gap, of about 16%.

Included here in the **dsld** package.

Commonality

In both examples, we have an outcome variable Y of interest–acceptance rate and wage income–and a sensitive variable S, which was gender in both examples. But in both cases, were concerned that merely comparing mean Y for each gender was an oversimplication, due to a possible *confounder*

C–department in the first example, occupation in the second. Failure to take confounders (there can be more than one, and usually are so) into account can lead to spurious "relations" between S and Y.

> **i** Confounder Adjustment Settings
>
> So, in general, we wish to investigate the impact of a sensitive variable S on an outcome variable Y, but *accounting for confounders* C. Let's call them "confounder adjustment" settings.

Now contrast the above examples with a different kind:

COMPAS recidivism data

COMPAS is a commercial machine learning software tool for aiding judges to predict recidivism by those convicted of crimes. A 2016 **Pro Publica** article investigated, finding the tool to be racially biased; African-American defendants tended to be given harsher ratings–i.e. higher estimated probabilities of recidivism–than similarly situated white defendants.

Northpointe, the firm that developed COMPAS, disagrees with the *Pro Publica* analysis, and we are not supporting either side here. But if the COMPAS tool were in fact biased, how could the analysis be fixed?

A key point is that any remedy must not only avoid using race directly, but must also minimize the impact of variables O that are separate from race but still correlated with it, known as *proxies*. If, say, educational attainment is correlated with race, its inclusion in our analysis will mean that race is still playing a role in our analysis after all.

> **i** Fair ML Settings
>
> Thus our goal is to predict the outcome variable Y, without using the sensitive variable S, while making only limited use of the proxy variables O.

Summary: the two kinds of discrimination analysis covered here

This COMPAS example falls in the category of *fairness in machine learning ML*.

Note the difference between accounting for confounders on the one hand, and fair ML on the other. Here is a side-by-side comparison:

aspect	confounder adjustment	fair ML
goal	estimate an effect	predict an outcome
harm	comes from society	comes from an algorithm
side info	adjust for confounders	limit impact of proxies

Summary of symbols

We'll use X to denote the rest of the variables, i.e. those that are related to Y but are not S, C or O. The general terminology is that Y is variously termed the *outcome variable*, *target variable* or *dependent variable*; the X, C, S and O variables are known collectively as *covariates*, *features* or *independent variables*.

example	Y	C	S	O
UCB admits	acceptance	program	gender	-
Census	wage	e.g. occupation	gender	-
COMPAS	sentence	-	race	e.g. education

Format of this tutorial

We treat the topics in this order:

- adjusting for confounders
- fair ML

Within each of the above topics, we cover:

- graphical and tabular exploration

- formal quantitative analysis

In each case, we present explanations of the relevant concepts, so that this is a general tutorial on methodology for analysis of discrimination, and show the details of using our **dsld** package to make use of that methodology.

So, let's get started.

Part I: Adjustment for Confounders

How do we adjust for confounders? The most common approach involves linear models, with which we express the mean Y for given values of the X, C and S variables in a linear form.

There will also be the question of *which* possible confounders to use.

Example: a simple gender wage gap analysis

Consider the **svcensus** data example above, investigating a possible gender pay gap. So Y is wage and S is gender. We might treat age as a confounder C, reasoning as follows. Older workers tend to have more experience and thus higher wages, and if there is an age differential in our data, say with female workers tending to be older, this may mask a gender pay gap.

So, let's take the set of confounders C to consist of age, and for simplicity in this introductory example, not include any other confounders, such as occupation, and let's not include any other variables X.

Initial analysis

Our linear model would thus be

$$\text{mean W} = b_0 + b_1 \text{ A} + b_2 \text{ M}$$

where W is wage, A is age and M is an indicator variable, with M = 1 for men and M = 0 for women. The parameters b_i are estimated by fitting the model to the data:

The column svcensus$gender is an R factor. Our function **dsldLinear** calls R's **lm**, which replaces that column by a dummy variable **gendermale**, our M above.

```r
svcensus1 <-
   svcensus[,c(1,4,6)]  # age, wage, gender
z <- dsldLinear(svcensus1,'wageinc','gender')
coef(z)  # print the estimated coefficients b_i
```

```
$gender
(Intercept)         age  gendermale
 31079.9174    489.5728  13098.2091
```

Interpretation of b_2

Lots here to discuss, which we will gradually cover below. For now, note that the *estimated* b_2 turns out to be about $13,000, which is the wage gap, if any. Here's why:

Always keep in mind that statistical quantities are only estimated, since we work only with sample data from some population, real or conceptual. Hence the need for standard errors, confidence intervals and so on.

Under the model, the mean wage for, say, 36-year-old men is

$$b_0 + 36\ b_1 + 1\ b_2$$

while for women of that wage it is

$$b_0 + 36\ b_1$$

The difference is b_2. But if we look at, for instance, people of age 43, the mean wages for men and women are

$$b_0 + 43\ b_1 + 1\ b_2$$

and

$$b_0 + 43\ b_1$$

and the difference *is still* b_2. So we can speak of b_2 as *the* gender wage gap, at any age. According to the model, younger men earn an estimated $13,000 more than younger women, with the same-sized gap between older men and older women.

The above approach to dealing with confounders is a very common one. But it raises questions, such as:

- What are the assumptions underlying that model? And how might we check whether they are (approximately) valid?

In addition, the data here are, as is commonly the case, *observational*, as opposed to being the result of a *randomized clinical trial*; there may be serious issues, due to unobserved confounders. Such problems might be solvable via an advanced (and rather controversial) methodology known as *causal inference*. Unfortunately, details are beyond our scope in this tutorial, but we will explain some basic concepts in

- We chose only one C variable here, age. We might also include occupation, as noted earlier. In some datasets, might have dozens of possible confounders. How do we choose which ones to use in our model? And for that matter, why not use them all?

- The above model, in which the gender wage gap was uniform across all wages, may not be adequate. How can we determine this, and what alternative models might we use?

Statistical inference

The full output of **dsldLinear()** goes to the heart of discrimination analysis, enabling statistical inferences on differences in levels of the sensitive variable S. Let's take a look:

```r
summary(z)
```

```
$`Summary Coefficients`
    Covariate   Estimate Standard.Error        PValue
1 (Intercept) 31079.9174     1378.08158 3.012511e-111
2         age   489.5728       30.26461  1.733205e-58
3  gendermale 13098.2091      790.44515  2.897184e-61

$`Sensitive Factor Level Comparisons`
         Factors Compared Estimates Standard Errors      P-Value
Estimate    male - female  13098.21        790.4451 2.897184e-61
```

The first half of this output is from **lm()**, which is called by **dsldLinear()**. The second half is the "value added" material from **dsld**.

So, an approximate 95% confidence interval for the gender wage gap is

$$13098.2091 \pm 1.96 \text{ x } 790.44515$$

or (11548.94,14647.48).

Since the gender gap here is simply b_2, the CI could of course have also been obtained directly from the **lm** half of the output. But with an S having more than two levels, e.g. race, the **dsld** enhancement is quite valuable.

With-interactions model

As discussed above, in our model

$$\text{mean W} = b_0 + b_1 \text{ A} + b_2 \text{ M}$$

we identified b_2 as *the* difference in mean wage between men and women, regardless of age, so that for instance:

> According to the model, younger men earn about \$13,000 more than younger women, with the same-sized gap between older men and older women.

But that may not be true. On the contrary, gender discrimination and age discrimination may interact. It may be, for instance, that the gender gap is small at younger ages but much larger for older people.

Interaction between two types of discrimination is called *intersectionality* by some analysts.

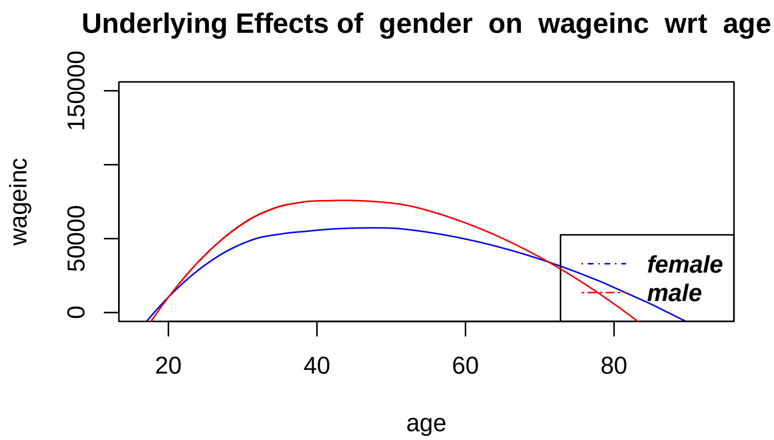
Assessing linearity

As noted, linear models are ubiquitous in observational data analysis. Open any professional journal in medicine, sociology, economics and so on, and you'll see many applications of this methodology. But how would one check that most basic assumption, the linearity of the mean Y for given X, C and S values?

> **Assumptions–not just a formality**
>
> Assumptions *matter*. They are never perfectly satisfied, but failure to be even approximately valid can mean deciding that there is no discrimination when it actually is there, or vice versa. It can mean bad medication being declared by the government as good, or vice versa. In litigation, if a key expert witness is exposed by opposing counsel as not having checked the assumptions in his/her analysis, the side for which the witness was testifying will likely lose the case on the spot.

Typically, linearity is checked graphically. A common approach involves plotting the *residuals*, which are the differences between the fitted line and the Y values. Here, though, we use another graphical approach, via a **dsld** function that may be more informative.

```
dsldConditDisparity(svcensus1,'wageinc','gender',
   'age','age > 0',yLim=c(0,150000))
```



The function plots a smoothed graph of Y against a user-specified C variable, once for each level of S. So, the call here says, “Plot wage income against age, for each gender.”

The function has a ‘conditions’ argument; we have none here, so we just used a trival one, ‘age > 0’

The model has mean Y being a linear of function of age, so we should expect to see approximate straight lines here. Yet the relation certainly looks nonlinear, possibly reflecting age discrimination against both very young and very old workers. We are already investigating one kind of discrimination here, gender, so again for simplicity let’s keep age as a confounder.

Updated model

But we must do something about the substantial nonlinearity we’ve discovered, and one possible remedy is to add an age^2 term be added to the equation:

Adding a squared term does not make our model nonlinear, as it is still linear in the b_i; if we, say, double each of those, the entire expression is doubled, the definition of linearity. The model is nonlinear in age but linear in the b_i.

mean $W = b_0 + b_1\ A + b_2\ A^2 + b_3\ M$

```r
svcensus1$age2 <- svcensus1$age^2
z <- dsldLinear(svcensus1,'wageinc','gender')
coef(z)  # print the estimated coefficients b_i
```

```
$gender
  (Intercept)           age    gendermale          age2
-104196.65579    7251.30962   15270.56685     -79.16059
```

So we see that the original wage gap figure of about $13,000 was incorrect, underestimating it by about 15%.

We see in this example that misspecifying a linear model can have a major impact on its accuracy. Further issues of model assumptions are beyond the scope of this book, but the interested reader is referred one of the most popular applied linear models books, *Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis*, by Prof. Frank Harrell, Jr. of the Vanderbilt University School of Medicine.

Other assumptions

Other than linearity, the standard errors reported by **lm()** also assume that variance of wage income is approximately constant across ages and genders. Lack of this property has some effect on the accuracy of reported standard errors, but this can be adjusted via the so-called *sandwich* operation, an option in **dsldLinear()**.

It is also assumed that wage income has a normal/gaussian distribution at each level, but the Central Limit Theorem's implications for the sums created by **lm()** are in fact approximately normal.

Part II: Fairness in Machine Learning

Motivation

why the need for fairml...

Example

introduce example + context...

FairML Methodology

introduce solutions they offer, what the results would be on the example, etc.

EDF Fair Methodology

introduce solutions they offer, what the results would be on the example, etc.

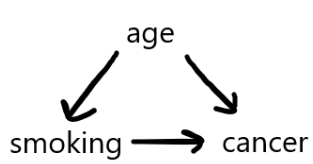
Appendix A: Fast Lane to Statistics

Appendix B: A Note on Causal Inference

One of the most common ways to do causal inference is to show the effects of your variables in a directed acyclic graph (DAG). Suppose you think there is a relationship between smoking and developing cancer. Your graph would look like this.

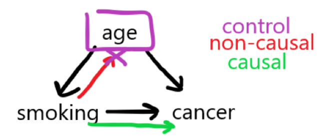
With an arrow from a smoking node to a cancer node. Because these nodes are connected, even ignoring the arrows, these two variables would have a statistical effect on each other. The arrow pointing from the smoking node to the cancer node indicates that smoking causes cancer.

Now, suppose you suspect that older people are more likely to smoke and are more likely to develop cancer. A 3rd variable, age, is added to the DAG like this.

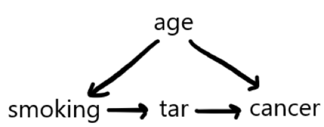


There is a new path in between smoking and cancer. But because the arrows are pointing the wrong ways, this *backdoor* *path* doesn't depict a causal relation between smoking and cancer. However, the existence of this additional path means there is an additional statistical effect between smoking and cancer. This is how confounding is represented a DAG.

But, by controlling for this age variable, you block this backdoor path from altering the true statistical effect observed between smoking and cancer.

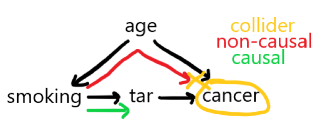


Suppose you add a new variable, the presence tar in lungs, that acts as a median between smoking and cancer.



The arrows all pointing in the same direction still indicates that smoking causes cancer, just through an intermediate step. But what about the relationship between smoking and tar?

There is the direct path between smoking and tar, indicating that smoking causes tar in lungs. But what of the backdoor path? Interestingly, this path is blocked, even if you didn't control age.

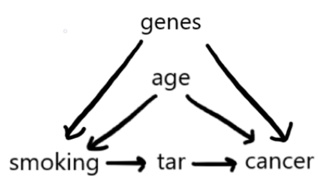


Because both tar and age both affect cancer, cancer is a a *collider* between tar and age. A collider indicates that there is no statistical effect between tar and age in the path that goes through cancer. But perhaps counter-intuitively, if you were to control for cancer, like if you were only to sample those with cancer, you would find an inverse statistical relationship between age and tar.

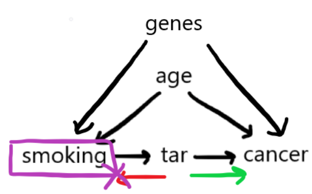
There is no connection between having COVID and having appendicitis, at least that we know of. Having appendicitis does not increase your chance of having COVID, and vice versa. However, having either of these attributes increases the likelihood of ending up in the hospital. Two variables have the same effect. That is, hospitalization is a collider in between COVID and appendicitis. An interesting thing happens when you control for that effect, ending up in the hospital. If you took a survey of patients at a hospital, you would find that those with appendicitis are less likely to have COVID. Think about it this way: If someone was in the hospital, and they didn't have appendicitis, it becomes likely that their hospitalization is because of COVID. So by controlling for a collider will create a statistical association between two variables that are otherwise unrelated, which is something you want to avoid.

With this information, there is something very powerful you can do with casual inference: control for a confounder you can't measure.

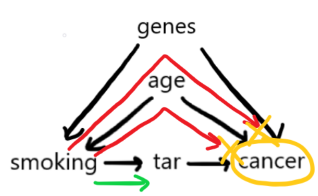
Lets say that one's genetics can both increase one's likelihood of becoming a smoker and developing cancer. Because you can't measure one's genetics, you cant control for it. But you still want to get the true causal effect between smoking and cancer.



We can solve this by getting the causal effect of smoking and tar, and combining that with the causal effect of tar and cancer. By controlling for the smoking variable, you can block the backdoor path between tar and cancer.



And, because cancer is a collider for genes, age, and tar, there the backdoor path is blocked between smoking and tar. Now, you have isolated the causal effects of smoking and tar, and tar and cancer.



Combining the two, usually by multiplying the linear regression coefficients, gives us the causal effect of smoking and cancer. And we did that by not even having to control for age or genes directly!

A Catch

All of this sounds very appealing, but there is a huge assumption we are making with these graphs: That these graphs are even accurate to our data. Data isn't always simple enough that we can represent the relationships between variables in just a few nodes and arrows. And as our data becomes more complex, it becomes harder justify certain decisions, like whether a variable X directly effects another Y, or are there some intermediate steps Z in between X and Y, as in our smoking-tar-cancer example.

Even in a simple example there seems to be no single correct answer in how a graph should be constructed. Say you were trying to model the DAG described in the UC Berkeley discrimination claims. However, there seems to be two ways to model it. Either one's gender influences the type of department they apply to, or the type of department influences the majority gender of the people that apply to it.

Both constructions seem valid to the scenario, but have two wildly different interpretations. If you were to control for the department, under the first graph, you may be underestimating the relationship of gender and acceptance rate. Under the second graph, you are controlling for the a confounding variable, isolating the true relationship of gender and acceptance rate.

All this is to say that constructing an accurate DAG requires extreme depth of knowledge in the field you are analyzing. Two different experts in the same field can come up with two entirely different graphs for a given scenario. And if different DAGS can produce two different sets of variables that should be controlled, your conclusions may be entirely different depending on how you construct your graph.

Because there is no way to know the correct way to construct a graph, the credibility for this kind of analysis is questionable at best.

Appendix C: Standard Errors via the Bootstrap