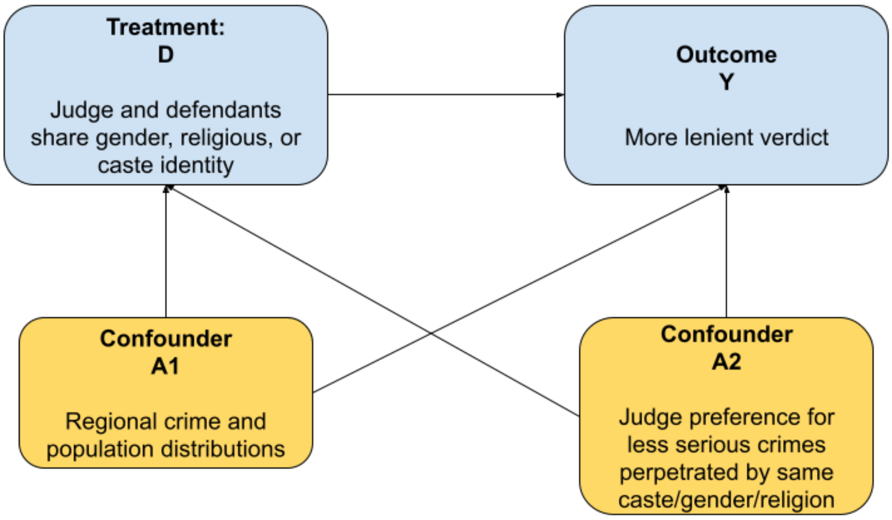
Kevin Blin 17-802-562 keblin@student.ethz.ch

Take home exam - Cohort 1

Building a Robot Judge

Question 1

a) The relationship analyzed is whether same gender/religion/caste between judges and defendants result in a more lenient verdict. If this revels true after taking into account confounders this would be a sign of in-group bias in the Indian judiciary system. Confounders that can be ruled out thanks to the random assignment of judges to defendants are for example the regional crime and demographic distributions, or potential preferences of judges for particular association of crime types and defendant profile.


Treatment:
D
Judge and defendants share gender, religious, or caste identity
Outcome
Y
More lenient verdict
Confounder
A1
Regional crime and population distributions
Confounder
A2
Judge preference for less serious crimes perpetrated by same caste/gender/religion

b) The balance tests described in equations 3-4 are done to verify the assumption of random judge-defendant assignments. The terms in the equations are:

   - judgeFemale, judgeMuslim, defFemale and defMuslim are binary variables set to true if the judge or defendant match the demographic attribute their named after
   - ϕct(i) is a court-month or court-year fixed effect
   - ζs(i) is an act and section fixed effect
   - Xi includes controls for demographics attributes

The two fixed effects are introduced to ensures that the comparison is done between defendants in the same court at the same time, and charged with similar crimes. If the assumption holds, by taking into account the fixed effects and other control factors, the equation should not show any correlation with all weights close to zero with minimum variance., in particular β1 and γ1 as this would introduce an in-group preference.

In table2 the row "Female defendants" and "Muslim defendant" respectively correspond to the obtained coefficient for β1 and γ1. The four specification in Table2 describe differs in the following:

- Column 1-2 reports the likelihood of being assigned to a female judge (equation 3), while column 2-3 report the likelihood of being assigned to a muslim judge (equation 4)
- Column 1 and 3 control for court-month fixed effects while column 2-4 control for court-year fixed effect.

c) Criminal charge fixed effects are not a collider as they are not influenced by the outcome (by the simple fact that the crime happened earlier in time than the verdict/outcome).

d) If cases were not randomly assigned, and given the not necessarily linear high dimensional case covariate in the dataset, we should use double ML.

Question 2

The huge mistake discussed in “Rookie Data Science Mistake Invalidates a Dozen Medical Studies” consist in balancing the dataset using oversampling BEFORE splitting the dataset in training and testing data. This should instead be done only AFTER splitting on the training set, otherwise the duplicates created by oversampling could leak into the test data, making the test results unrepresentative of the generalization capabilities of the model with previously unseen data.

```python
## WRONG WORKFLOW

import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.utils import resample

# Load a dataeset (e.g: data from homework 5)
df = pd.read_csv('data/HW05.csv')

#identify minority class
df_0 = df[df.iloc[:, -1] == 0]
df_1 = df[df.iloc[:, -1] == 1]
count0 = len(df_0)
count1 = len(df_1)

# upsample and concat to get a balanced dataset
if count0 < count1:
    # 0 is the minority class
    df_0_upsampled = resample(df_0,random_state=42,n_samples=int(np.ceil(count1/count0)),replace=True)
    balanced_df = pd.concat([df_0_upsampled, df_1])
else:
    # 1 is the minority class, upsample
    print(np.ceil(count0/count1))
    df_1_upsampled = resample(df_1,random_state=42,n_samples=int(np.ceil(count0/count1)),replace=True)
    balanced_df = pd.concat([df_1_upsampled, df_0])

# Train data
X = balanced_df.iloc[:, :-1]
# Target:
y = balanced_df.iloc[:, -1]
# Split the data into 80% training and 20% test
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

```
## RIGHT WORKFLOW
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.utils import resample

# Load a dataeset (e.g: data from homework 5)
df = pd.read_csv('data/HW05.csv')

# Train data
X = df.iloc[:, :-1]
# Target:
y = df.iloc[:, -1]
# Split the data into 80% training and 20% test
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Recompose dataset with only training data
df_train = pd.DataFrame(X_train)
df_train["target"] = y_train
#identify minority class
df_train_0 = df[df.iloc[:, -1] == 0]
df_train_1 = df[df.iloc[:, -1] == 1]
count0 = len(df_0)
count1 = len(df_1)

# upsample and concat to get a balanced dataset
if count0 < count1:
    # 0 is the minority class
    df_0_train_upsampled = resample(df_0,random_state=42,n_samples=int(np.ceil(count1/
count0)),replace=True)
    balanced_train_df = pd.concat([df_0_upsampled, df_1])
else:
    # 1 is the minority class, upsample
    df_train_1_upsampled = resample(df_1,random_state=42,n_samples=int(np.ceil(count0/
count1)),replace=True)
    balanced_train_df = pd.concat([df_1_upsampled, df_0])
```

Question 3

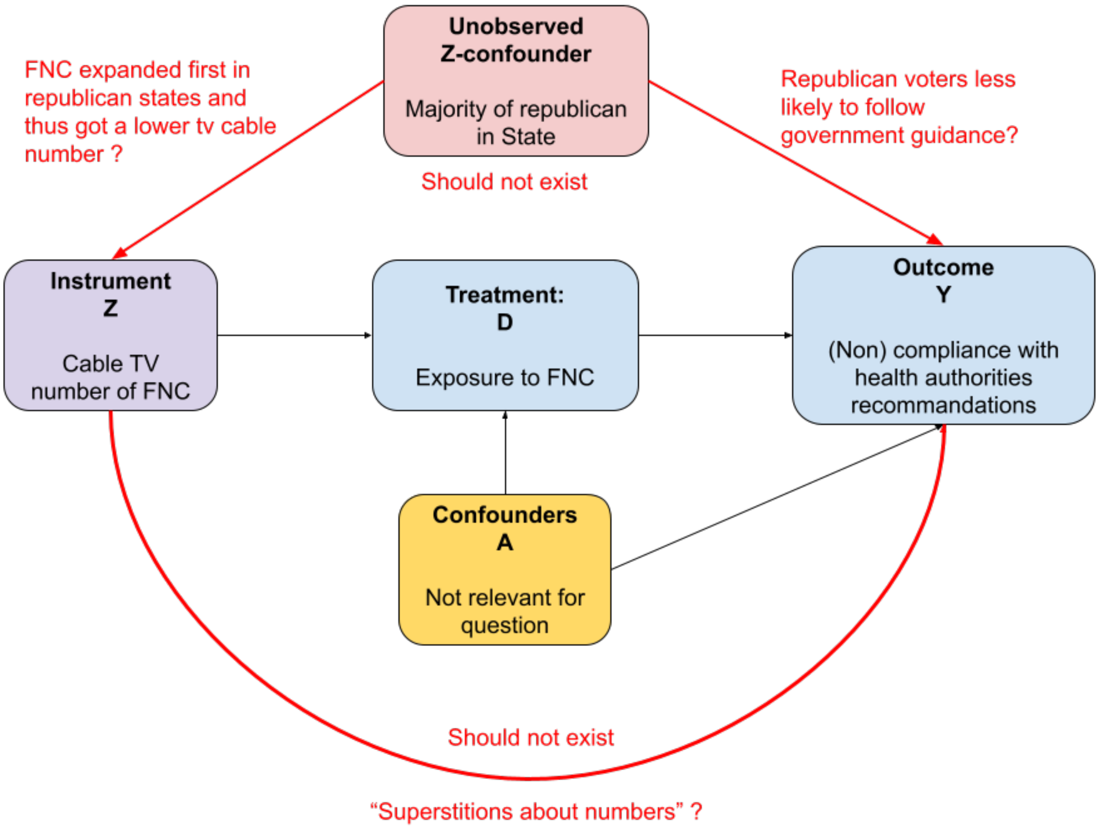
a) In this paper the outcome correspond to the compliance with health authorities recommendations, the treatment is the amount of exposure to FNC, and the instrument variable used is the cable TV channel number of FNC.

b) The three assumptions needed to verify the validity of an IV are relevance, exogeneity and exclusion.

- Relevance: the IV must be correlated with the treatment. The assumption in the paper is that viewers spend more time on channels who are assigned a lower number. The authors runs an experiment (reported in Figure 2) to show the first stage relationship between the channel position in the cable system and the associated viewership. The assumption would be violated if there was no correlation between the position of the channel and the likelihood of viewership.
- Exogeneity: the requirement for exogeneity is that no unobserved factors affect both the outcome and the instrument. In this case the pseudo-random assignment of tv channel numbers makes this assumption very likely. This assumption is explained in further detail in the supplementary section S1, where the authors refer to several other studies establishing this link. This assumption could be violated if the underlying mechanism to assign the channel numbers could also somewhat directly influence the non respect of sanitary recommendations.

- Exclusion: an IV respect exclusion if the instrument only affects outcome through treatment variable (no direct causal link from IV to outcome). In this context this is rather clear that the semi-randomly assigned number is not causing any change in behaviour but is rather the time spent watching FNC (independently of the channel number) that influences the outcome. This assumption could be violated if for some mysterious reasons a pseudo-random number could directly affect the actions of people.

c) As already explained in point b) the chosen instrument for FNC viewership seems to well satisfy all the three required assumptions for being a good IV. I would however raise the possibility of not complete exogeneity. The existence of an historically strong republican majority in a given locality could potentially affect both the individual compliance to health recommendations (republicans seemed in general more unlikely to follow health guidances) as well as an earlier expansion of FNC in that territory which would have resulted in a lower cable TV number (assuming that these numbers were assigned in increasing order for each new channel).


Unobserved
Z-confounder
Majority of republican in State
FNC expanded first in republican states and thus got a lower tv cable number ?
Republican voters less likely to follow government guidance?
Should not exist
Instrument
Z
Cable TV number of FNC
Treatment:
D
Exposure to FNC
Outcome
Y
(Non) compliance with health authorities recommandations
Confounders
A
Not relevant for question
Should not exist
“Superstitions about numbers” ?

Question 4

The four conditions for ML predictions to be sufficient for decision-making are:

1. Payoff of the decision does not depend on other factors besides predictions Y-hat (benefits are are mostly homogeneous).
2. Environment factors (i.e. decision subjects) do not change behaviour in response to the algorithm.
3. Decision-makers respond predictably to the algorithm. (e.g each decision-maker *j* follows the algorithm threshold rule).
4. Decision-maker gets continuous feedback on model accuracy

When applied to the video ad campaign these conditions would be:

1. Showing the political ad to a given person increases the likelihood for voting the candidate in the ad in an homogenous way. This is likely not the case as, even among viewers on which the ad would have the desired effect, some might be more receptive and it would be better to focus on them. Also some viewers might be more "influential" than others and convincing them to swing on the side of the desired candidate could have a much bigger impact than persuading a person with few social connections.
2. Video ad viewers should not change their behaviour in response of the algorithm. This might not happen if the viewer realise he's being targeted by a political ad and then decides to stop watching further ads or, even worse, vote for another candidate as a retaliation.
3. The political campaign organiser must follow the algorithm recommendations. This could not happen if there are some factors not taken into account by the algorithm that the campaigners must respect (e.g. some local legislation regulating the type of political ads that are allowed).
4. The political campaign organiser must received constant feedback on how their algorithm perform. While this could be easily done online thanks to engagement metrics, tv or public spaces ads would make it harder to measure the effectiveness. On top of that engagement is not a prefect predictor of the final voting decision, thus the optimal feedback is very sparse and only applicable to future campaigns and not the current one.

Question 5

1. Admitted student should be similarly likely to complete their studies and don't waste school resources
2. Applicants should not change their applications essays in response to the algorithm
3. The school should follow the algorithm decision. That's the point we'll discuss below
4. The school should observe if the admitted candidates perform well, and the drop-out rate decreases after the introduction of the algorithm screening. However it is not possible to observe if rejected applicants would have performed well, making a continuous feedback incomplete.

Schools (especially top ones) can receive way more applications than the available places (e.g. MIT, Stanford, Harvard all have an acceptance rate of about only 4%), a partially automated system is thus advisable to save screening resources. Now the question lies in whether the screening should be completely automated (e.g. point 3 completely respected) or if we should leave some space to the school to make exceptions. As discussed in class, if all relevant characteristics to take the decision can be captured in the provided data (in this case the admission essay) further human supervision might be useless. However current language analysis algorithm (often based on LLMs) tend to lack a causal understanding of the text and purely rely on statistical information that can be heavily biased depending on the chosen training corpus. This would strongly speak against the use of these automated tools, especially if they also lack explanation abilities to justify their choice.

On the other hand human can also be (unconsciously) biased and take sub-optimal decisions, leaving lot of space to unsatisfying arbitrary decisions. A combination of automated review and human intervention currently looks like the best solution, with the automated tools extracting potentially relevant information from the text (e.g. common red flags like gap years or important elements like previous school or GPA) that can speed up the human workflow while letting the human in control of the final decision.

Question 6

a) This article measures fairness as **calibration / sufficiency**, fairness would be obtained by equalizing precision across the groups two groups (white and black sounding names or persons). This can be inferred by the following excerpts:

- "Their résumés were about as similar as two résumés can be. One person was named Jamal, the other Brendan."
- "Both were grappling with diabetes and high blood pressure. One patient was black, the other was white."
- "Both studies documented racial injustice: In the first, the applicant with a black-sounding name got fewer job interviews. In the second, the black patient received worse care.

All this indicates that the the applicants are identical but the protected attribute (race) and that should thus receive the same outcome: conditioning on the other attributes, both groups get the same healthcare treatment or call job interview.

b) The two others concepts of fairness are **statistical parity/independence** and **separation**.

- Separation correspond to equalizing recall across groups. For the first study (binary decision) this can be verified by checking that both true positive rates (recall for positive class) and false positive rates are the same across groups. For the continuous case of the second study, the authors should check for independence of the predictions with respect to the protected attribute A given the true label Y: $\hat{y} \perp A \mid Y$.
- Statistical parity / independence require the average predicted outcome to be the same for the two groups. This would be satisfied if independently of the health status of CV both black and white individuals would receive the same treatment or have the same chance of being called for an interview.

Question 7

a) A global explanation like feature importance is referred to the whole dataset and not a single datapoint. It explains which are the most relevant metrics on average that the model uses to make predictions and in the case of corruption auditing it is useful to understand which indicators prosecutors should pay more attention to. On the opposite, a local explanation like shapely values (or counterfactual explanations) would be applied to a specific municipality to determine which features contributed the most to determine the model prediction (or closest the set of features that would change the prediction outcome) for this municipality. A local explanation could be more useful if the auditors want to build a strong case against a given municipality using irregularities specific to that municipality.

b) The baseline model from section 3, is designed to use the municipal budget information and predict from that the probability that a given municipality is fiscally corrupt. The other tasks are analysis of the effect of revenue shocks or auditing on corruption, and finally to use the baseline model predictions to guide policymakers. Just as LLMs are trained for a baseline task (language modelling) and can later on be adapted to other tasks just by chaining some weights, it seems reasonable that the baseline corruption prediction model learns a set of features that then can be used or not in some downstream tasks. Actually since the task are different we would expect a different subset of the features to be retained, for example a more restricted and interpretable set seems more adapted to optimize the task of guiding policymakers.

c) The concept of statistical fairness used by the authors is the one of "statistical parity / independence". This is a direct consequence of the chosen structure for the decision making process. First the prediction algorithm is applied as it is, but then the results are filtered by party and "Within each party, we audit the same share of municipalities. Then by construction, the incidence of audits is equal across parties". The results is thus an equality in outcomes, all groups (political parties) having the same chance of being targeted by an audit. The study shows that this approach doesn't penalise excessively the performance of the targeting algorithm but if the True Corruption Rate (base rates) were much more different from one party to another this approach to fairness could result in a severe reduction in efficiency. In that case it would be more appropriate to equalize recall across parties: the algorithm would still be able to take advantage of the existing base rates and target more parties that are more prone to corruption, and fairness would be guaranteed in the sense that, given the party, the model would achieve the same accuracy in predicting corruption.