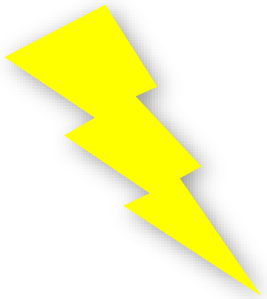
5. Regression and Causality

Reconsider the Bloom et al. (2012) studies that we saw in the introduction:

- Randomly drawn sample of 10,000 organizations.
- Authors define “best” management practices
- Firms adopting theses practices are more productive/profitable, grow faster and survive longer.
- Potential conclusion: Firms must “just” adopt these management practices and they will automatically be more productive/profitable, grow faster and survive longer



What is the problem with such a conclusion?

Recall:

- Regressions give us an approximation to Conditional Expectations
- Conditional Expectations *predict* the outcome of a variable on the basis of other variables
- If we know $E[Y|X]$ we can tell the following:
  - If you tell me a value of X (say x), what is the average value of Y we can expect when $X = x$?
  - *"Which job satisfaction can we expect in firms with performance pay as opposed to firms without?"*
- While this is a powerful property, it does not necessarily tell you:
  - If you change the value of X (say from x_1 to x_2) for objects in the population how is their average value of Y affected by this?
  - *"When we introduce performance pay, how would this change job satisfaction, on average?"*
- Typical reason: there are other variables affecting both X and Y

Counterfactuals and Causality

- The question whether a regression is causal boils down to the question whether the conditional expectation is causal
- If the CEF is causal we can estimate causal effects with a regression analysis
- To answer this question it is very useful to think about *potential outcomes* or *counterfactuals*
  "What would have happened, when a different decision had been made?"
- This seems hard to answer!
  (But it is often still a useful thought experiment in real life)
- But we sometimes can say something about the counterfactual using data
- When this is the case empirical research becomes very powerful!

5.1 Thinking about Potential Outcomes

- Suppose we want to investigate whether
  - a certain management practice
    (performance pay, wage increase, training,...)
  - causally affects some outcome variable Y_i
    (job satisfaction, performance,...)
- Let $C_i \in \{0,1\}$ be a dummy variable indicating whether the practice is implemented for person i
- What we would like to know is: what is the value of Y_i
  - if $C_i = 1$ ("person i is treated")
  - if $C_i = 0$ ("person i is not treated")
- Let this *potential outcome* be

$$Y_{C_i i} = \begin{cases} Y_{1i} & if \quad C_i = 1 \\ Y_{0i} & if \quad C_i = 0 \end{cases}$$

- The *causal effect* of C_i on Y_i is now $Y_{1i} - Y_{i0}$

The problem is:

- when we implement the practice we only observe Y_{1i}
- when we do not implement the practice we only observe Y_{0i}

In real life we do not observe the *counterfactual*

- What would have happened if we had decided differently?
- The *observed outcome* is Y_i where

$$Y_i = Y_{0i} + (Y_{1i} - Y_{0i}) \cdot C_i$$

- Running a simple regression (or comparing means) in a sample yields
  - $E[Y_i|C_i = 1]$ and
  - $E[Y_i|C_i = 0]$
- Here, one may be tempted to interpret

$$E[Y_i|C_i = 1] - E[Y_i|C_i = 0]$$

as the causal effect of C on Y

But note that

$$E[Y_i|C_i = 1] - E[Y_i|C_i = 0]$$
$$= E[Y_{1i}|C_i = 1] - E[Y_{0i}|C_i = 0]$$
$$= E[Y_{1i}|C_i = 1] - E[Y_{0i}|C_i = 1] + E[Y_{0i}|C_i = 1] - E[Y_{0i}|C_i = 0]$$
$$= E[Y_{1i} - Y_{0i}|C_i = 1] + E[Y_{0i}|C_i = 1] - E[Y_{0i}|C_i = 0]$$

- The causal effect of C on the group that is treated ($C = 1$) is

$$E[Y_{1i} - Y_{0i}|C_i = 1]$$

  - It is called the *average treatment effect on the treated (ATT)*
  - Very often this is what we want to know
  - *"Has job satisfaction increased in a group of employees because this group now receives performance pay?"*

- But: the regression coefficient may not estimate the ATT
  - It includes $E[Y_{0i}|C_i = 1] - E[Y_{0i}|C_i = 0]$
  - This is the *selection bias*

We can thus decompose:

$$\underbrace{E[Y_i|C_i = 1] - E[Y_i|C_i = 0]}_{\text{Observed difference in outcome}}$$

$$= \underbrace{E[Y_{1i} - Y_{0i}|C_i = 1]}_{\text{Average treatment effect on the treated}} + \underbrace{E[Y_{0i}|C_i = 1] - E[Y_{0i}|C_i = 0]}_{\text{Selection bias}}$$

- If $E[Y_{0i}|C_i = 1]$ differs from $E[Y_{0i}|C_i = 0]$
  - Treated and untreated individuals differ
  - $E[Y_{0i}|C_i = 0]$ is not the counterfactual outcome for the treated
- Then the regression estimates are biased estimates of the causal effect!

Example: Does a university education increase earnings?

- $E[Y_{0i}|C_i = 1]$ is the wage somebody who attended a university would earn when not having attended university
- It is very likely that $E[Y_{0i}|C_i = 1] > E[Y_{0i}|C_i = 0]$
- Hence, we would overestimate the true returns to a university education

Your Task **Simulated data set: Evaluation of a sales training**

- Write a script that generates a fictitious data set with 10000 observations

```
n=10000
df=pd.DataFrame(index=range(n))
```

- Generate a normally distributed random variable $ability$ with mean 100 and std. deviation 15: df['ability']=np.random.normal(100,15,n)
- Generate a dummy variable $training$:

  df['training']=(df.ability+np.random.normal(0,10,n)>=100)

  (Hence, more able people have a higher likelihood to be trained)
- Generate a variable sales:

```
df['sales']= 10000 + df.training*5000 + df.ability*100
                + np.random.normal(0,4000,n)
```

- This is the true causal relationship: the training increases sales by 5000
- But suppose we as researchers cannot observe $ability$
- Run a regression of sales on training & interpret the results (& save the notebook as SalesSim1)

Recall:

- A regression estimates the Conditional Expectation Function
- The CEF gives us $E[Y_i|C_i = 1] - E[Y_i|C_i = 0]$
- It identifies a causal effect only if $E[Y_{0i}|C_i = 1] - E[Y_{0i}|C_i = 0] = 0$

This is satisfied if C_i is *independent* of (Y_{0i}, Y_{1i})

- That is neither Y_{0i} nor Y_{1i} are systematically different for people with different realizations of C_i
- Let the symbol ⫫ indicate independence
- If the condition

$$(Y_{0i}, Y_{1i}) \perp\!\!\!\perp C_i$$

is satisfied we can use simple regressions (or here mean comparisons) to identify causal effects

5.2 Why are Experiments so Important?

- Suppose we have a randomized controlled experiment
  - That is C_i is randomly (that is *exogenously*) assigned to the individuals i
  - In turn, C_i is by construction independent of Y_{i0}
  - Hence, $E[Y_{0i}|C_i = 1] = E[Y_{0i}|C_i = 0]$
  - The selection bias is eliminated!
  - We obtain an unbiased estimator of the causal impact of C in the population
- In that case

$$E[Y_i|C_i = 1] - E[Y_i|C_i = 0] = E[Y_{i1} - Y_{i0}]$$

- A simple comparison between the averages of treatment and control yields an unbiased estimate of the causal effect
- The same holds for a regression on a treatment dummy

Your Task — Simulated data set: Evaluation of a sales training II

- Open your SalesSim1 notebook, save it as SalesSim2 to generate a different simulation, and run the whole notebook
- Now suppose that there is new training program which is *randomly assigned*
- Ass a cell at the end of the notebook to generate a dummy variable $training2$ which takes value 1 for 5% randomly chosen individuals

```
df['training2']=np.random.binomial(1, 0.05, n)
```

- Note: np.random. binomial(1,0.05,n) generates a vector of n binomial random variables with 1 trial each (taking value 1 with 5% probability)
- Assume that this new program also raises sales by 5000:

```
df['sales']= df.sales + df.training2*5000
```

- Run a regression of sales on training and training2
- Interpret the results & save the notebook

5.3 Control Variables & Omitted Variable Bias

- But what if we do not have an experiment?
- In multiple regression we "control for" other covariates X_i
- (When) does this help us to identify causal effects?
- We can write $E[Y_i|X_i, C_i = 1] - E[Y_i|X_i, C_i = 0]$

$$= E[Y_{1i} - Y_{0i}|X_i, C_i = 1] + E[Y_{0i}|X_i, C_i = 1] - E[Y_{0i}|X_i, C_i = 0]$$

The Conditional Independence Assumption (CIA)
If the *conditional independence assumption holds,* i.e.

$$Y_{ci} \perp\!\!\!\perp C_i \mid X_i \text{ for all values of c,}$$

(conditional on X the treatment C_i is independent of potential outcomes), then

$$E[Y_i|X_i, C_i = 1] - E[Y_i|X_i, C_i = 0] = E[Y_{1i} - Y_{0i}|X_i, C_i = 1],$$

i.e. the difference in conditional expectations has a causal interpretation.

Note:

- This is a weaker property than the independence assumption $(Y_{0i}, Y_{1i}) \perp\!\!\!\perp C_i$ above
- We do not need that C_i is independent from potential values
- But it needs to be independent for people who have the same values for a set of observable co-variates

The *Conditional Independence Assumption* is crucial in many applications

- Useful question: is C_i as good as randomly assigned conditional on X_i?
- Or, in other words: are the variables in X_i the only reason why (Y_{0i}, Y_{1i}) are correlated with C_i?
- This is also called the "*selection on observables*" assumption: i.e. selection into the treatment only depends on observable variables X_i; beyond that it is random
- In that case a regression which controls for X_i (in a proper manner) has a causal interpretation

Analogously: Continuous "treatment" variable

- Think in terms of a causal model $Y_{si} \equiv f_i(s)$
  - $f_i(s)$ describes how an object i (person, firm, ...) responds to changes in some variable s
  - or: determines the outcome for all *potential* realizations of s
- Now let $f_i(s) \equiv f(s, X_i)$
- Distinction between CEF $E\left[Y_i \middle| S_i, X_i\right]$ (or regression as its approximation) and causal model $f(s, X_i)$
  - The CEF describes the mean of Y when I draw objects with the same values of (S_i, X_i) from the population (and regressions approximate these conditional expectations)
  - The causal model $f(s, X_i)$ describes how Y changes when I change s
- Regressions approximate the causal model when the CIA holds

A Note on Terminology: *Identifying Assumptions*

- When we use *observational data* (that is data that we observe but which has not been generated by an experiment), we can never be entirely sure that our regression captures the causal effect
- But still for many questions it is hard to design an appropriate field experiment
- We can (and should) still try to say something about causality
- In order to do so, we typically state so called *identifying assumptions*
  - That is: we make clear under what conditions our empirical approach would capture a causal effect
- The conditional independence assumption is an example for such an identifying assumption

Omitted Variable Bias

- Assume that the causal relationship between Y_i and C_i is determined by

$$Y_i = \alpha + \rho \cdot C_i + \gamma \cdot X_i + v_i$$

  where that v_i is uncorrelated with all regressors

- When the CIA holds, then ρ is equal to the coefficient in the linear regression of Y_i on C_i and X_i
- But assume that we cannot (or do not) include X_i and estimate

$$Y_i = \widetilde{\alpha} + \widetilde{\rho} \cdot C_i + \eta_i$$

- The short regression yields (use the true causal relationship)

$$\widetilde{\rho} = \frac{Cov[C_i, Y_i]}{V[C_i]} = \frac{Cov[C_i, \alpha + \rho \cdot C_i + \gamma \cdot X_i + v_i]}{V[C_i]}$$

$$= \rho + \frac{Cov[C_i, \gamma \cdot X_i + v_i]}{V[C_i]}$$

$$= \rho + \gamma \cdot \frac{Cov[C_i, X_i]}{V[C_i]}$$

- If $Cov[C_i, X_i] \neq 0$ the coefficient is biased ("*omitted variable bias*")

$$\tilde{\rho} = \rho + \gamma \cdot \frac{Cov[C_i, X_i]}{V[C_i]}$$

- But $\frac{Cov[C_i,X_i]}{V[C_i]}$ is the coefficient in a regression

$$\underbrace{X_i}_{\text{Omitted variable}} = \delta_0 + \delta_c * \underbrace{C_i}_{\substack{\text{Included} \\ \text{"endogenous"} \\ \text{variable}}} + v_i$$

- Then

$$\tilde{\rho} = \frac{Cov[C_i, Y_i]}{V[C_i]} = \rho + \gamma \cdot \delta_c$$

Hence: If C_i is *endogenously* determined by X_i and we cannot observe X_i

- then the regression will yield a biased estimate of the causal effect
- the size of this *omitted variable bias* is $\gamma \cdot \delta_c$

Study

Education and Wages

- Consider association between wages and education in the NLSY97
- We find that CEF of wages is strongly increasing in education
  - But is this a causal effect?
  - It seems quite likely that there is omitted variable bias
- In 1997 and early 1998, the NLSY97 respondents were given the *Armed Services Vocational Aptitude Battery* (ASVAB) wich comprises 10 tests that measure knowledge and skills in a number of areas
- First: Regress wages in 2012 on dummy variables for educational degrees
- In a second step:
  - Control for a standardized ASVAB score (Mathematical Knowledge, Arithmetic Reasoning, Word Knowledge, and Paragraph Comprehension)

Your Task

OVB (Simulated Sales Training Evaluation III)

- Open the SalesSim1 notebook
- Again regress
  - $Sales$ on $training$
  - $Sales$ on $training$ and $ability$
- Regress $ability$ on the "endogenous" variable $training$

  How do you interpret the coefficient of $training$ in the last regression?

  (Note this is not causal! but think of CEF interpretation of regression)
- Compute the OVB using this coefficient
- Interpret the size of the OVB

6. Panel Data and Fixed Effects

- When we have longitudinal data we can potentially tackle OVB when the unobserved omitted factors are *stable over time*
- Setting:
  - We can measure the outcome variable for a set of objects (people, firms, ...) at several point in time
  - The key variable of interest (the „treatment“) changes over time
  - We study the association between the change in the treatment variable and the change in the outcome variable
- Here: Consider *Fixed Effects Models* as one important approach

6.1 Fixed Effects

- Consider again the potential outcome framework (time index $t = 1, \dots, T$)

$$Y_{C_{it}it} = \begin{cases} Y_{1it} & if \quad C_{it} = 1 \\ Y_{0it} & if \quad C_{it} = 0 \end{cases}$$

- Assume now that

$$E[Y_{0it}|A_i, X_{it}, t, C_{it}] = E[Y_{0it}|A_i, X_{it}, t]$$

  where

  - X_{it} is a vector of observed (time varying) covariates and
  - A_i is is a vector of *unobservable* factors that are fixed over time (no time index t! For instance, a person‘s ability or personality)

- The assumption states that C_{it} is as good as randomly assigned conditional on A_i and X_{it}
- This is a sensible identifying assumption whenever any unobserved determinants of the treatment (that also may affect the outcomes beyond the treatment) are time constant

- Consider now the following linear model

$$E[Y_{0it}|A_i, X_{it}, t] = \alpha + X'_{it}\beta + A'_i\gamma + \lambda_t$$

- And assume that the causal effect is a constant ρ

$$E[Y_{1it}|A_i, X_{it}, t] - E[Y_{0it}|A_i, X_{it}, t] = \rho$$

- Hence, we can write

$$Y_{it} = \alpha_i + \lambda_t + \rho C_{it} + X'_{it}\beta + \epsilon_{it}$$

where $\epsilon_{it} = Y_{0it} - E[Y_{0it}|A_i, X_{it}, t]$ and $\alpha_i = \alpha + A'_i\gamma$

- When we impose these assumptions, running a regression will estimate the causal effect ρ of C on Y

- This is a fixed effects model:
  - The α_i are parameters to be estimated (estimating a dummy for every person)
  - The γ_i are time effects that are also estimated (estimating a dummy for every period)

Study

Lazear's (2000) study on Performance Pay at Safelite

- Safelite is a large auto glass company in the US
- Business: replace broken windshields.
- New compensation scheme in January 1994: Piece rate scheme (PPP) replaced hourly-wage scheme in 1994
- The piece rate scheme was phased in over 19 months, starting from the headquarter town.
- The gradual implementation of piece rate allows for within-worker variation identifying the incentive effect of piece rate on effort.
- But: also high turnover rates; many workers also hired after the introduction of the PPP
- In the following:
  - Unit of observation = Worker in a given month;
  - Productivity measure: Average windshields installed by the worker on a given day.

Safelite: Regression analysis

TABLE 3—REGRESSION RESULTS

Regression number	Dummy for PPP person-month observation	Tenure	Time since PPP	New regime	R^2	Description
1	0.368 (0.013)				0.04	Dummies for month and year included
2	0.197 (0.009)				0.73	Dummies for month and year; worker-specific dummies included (2,755 individual workers)
3	0.313 (0.014)	0.343 (0.017)	0.107 (0.024)		0.05	Dummies for month and year included
4	0.202 (0.009)	0.224 (0.058)	0.273 (0.018)		0.76	Dummies for month and year; worker-specific dummies included (2,755 individual workers)
5	0.309 (0.014)	0.424 (0.019)	0.130 (0.024)	0.243 (0.025)	0.06	Dummies for month and year included

Notes: Standard errors are reported in parentheses below the coefficients.
Dependent variable: ln output-per-worker-per-day.
Number of observations: 29,837.

Safelite (continued): What do the worker fixed effects do here?

- Regression without worker fixed effects (row 1)
  - this gives us an estimate of the causal effect of the treatment on the *average performance* of all workers working at a given point in time
  - (when believe that the treatment is as good as randomly assigned conditional on the time period which seems very plausible here)

- However: if we are interested in the causal effect of the treatment on the performance of an *average given worker* this is a „biased“ estimate
  - This is the case when the ability of workers depends on the treatment
  - For instance, when the PPP allows to hire better workers
  - Then $E[Y_{0it}|t, C_{it} = 1] > E[Y_{0it}|t, C_{it} = 0]$

    i.e. workers hired under the PPP would be better even without the PPP
  - In this respect the conditional independence assumption is violated
  - There is a classical selection bias and the PPP dummy should give a too high estimate for the causal effect of the PPP on a *given* worker

What do the worker fixed effects do here?

- The worker fixed effects model (row 2) takes this problem into account
- It imposes the weaker assumption that

$$E[Y_{0it}|A_i, t, C_{it}] = E[Y_{0it}|A_i, t]$$

- When A_i captures the workers unobserved ability this assumption states that for workers *of the same ability* the counterfactual performance is independent of the treatment
- The fixed effects model in a sense estimates the unoberved abilities of the workers (using that a worker‘s performance is observed over many months)
- It thus estimates the causal effect of the PPP conditional on worker‘s abilities
- Note: The model without fixed effects is here not wrong, it estimates something different
  - Without worker fixed effects it estimates the total effect on performance which includes a *selection* and an *incentive effect*
  - With worker fixed effects it estimates the pure incentive effect

Estimating Fixed Effects Models

- Estimating the coefficients of individual dummy variables seems demanding in large panels (1000 employees = 1000 fixed effects)
- However, if we are not interested in knowing the specific values of the individual fixed effects, we can estimate the model in a simpler manner
- Consider

$$Y_{it} = \alpha_i + \lambda_t + \rho C_{it} + X'_{it}\beta + \epsilon_{it}$$

- Now take the average across all time periods $\overline{Y_i} = \frac{1}{T}\sum_{t=1}^{T} Y_{it}$

$$\overline{Y}_i = \alpha_i + \overline{\lambda} + \rho\overline{C}_i + \overline{X}'_i\beta + \overline{\epsilon}_i$$

and substract this from Y_{it}

$$Y_{it} - \overline{Y}_i = \lambda_t - \overline{\lambda} + \rho\left(C_{it} - \overline{C}_i\right) + (X'_{it} - \overline{X}'_i)\beta + \epsilon_{it} - \overline{\epsilon}_i$$

→ The α_i are eliminated!

$$Y_{it} - \overline{Y}_i = \lambda_t - \overline{\lambda} + \rho\left(C_{it} - \overline{C}_i\right) + (X'_{it} - \overline{X}'_i)\beta + \epsilon_{it} - \overline{\epsilon}_i$$

- Hence,
  - replace the outcome variable by its deviation from the mean over time
  - replace the explanatory variables by their deviations from their means over time
  - Regress the „de-meaned“ outcome on the „de-meaned“ explanatory variables

  $\rightarrow$ This gives us an estimate of ρ

  $\rightarrow$ We can estimate ρ and β without having to estimate the α_i

- This model is sometimes also called the *within-estimator:*

  It estimates the effect of ρ on Y from the within person variation in C

Python

Fixed Effects Regressions in Python

- Panel regressions in Python can be done with library linearmodels
- Install by !pip install linearmodels
- Import by from linearmodels import PanelOLS
- In order to run a panel regression use a MultiIndex DataFrame that is a DataFrame that uses two indices
  - one index for the entity variable (the omitted time constant variable)
  - one index for the time variable

```
df=df.set_index(['entity', 'year'])
```

- Then fit the model by

```
reg = PanelOLS.from_formula('y ~ x + EntityEffects + TimeEffects', data=df).fit()
```

- Then print the output with print(reg)
  (Note the different notation to statsmodels: can directly print the results)

Your Task — Fixed Effects

- Open the notebook in which you estimated the association between Management Practices and ROCE
- For a part of the observations the data set contains panel data
- The paper by Bloom et al. (2012) contains the following table, where the third colums shows the result of a fixed effects regression
- Please replicate this regression using PanelOLS
- Note:
  - The variable account_id contains an identifier for each firm
  - The variable emp contains the number of employees and ppent the capital (fixed assets)
  - You can generate logs by using np.log(x) directly in the formula

	(1)	(2)	(3)
Sector	Manufact.	Manufact.	Manufact.
Dependent variable	Log (Sales)	Log (Sales)	Log (Sales)
Management	0.523*** (0.030)	0.233*** (0.024)	0.048** (0.022)
Ln(Employees)	0.915*** (0.019)	0.659*** (0.026)	0.364*** (0.109)
Ln(Capital)		0.289*** (0.020)	0.244*** (0.087)
Country controls	No	Yes	NA
Industry controls	No	Yes	NA
General controls	No	Yes	NA
Firm fixed effects	No	No	Yes
Organizations	2,927	2,927	1,453
Observations	7,094	7,094	5,561

Your Task — Fixed Effects (Simulated Sales Training Evaluation VII)

Generate the following notebook

```
n=2000
df1=pd.DataFrame(index=range(n))
df1['ability']=np.random.normal(100,15,n)
df1['year']=1
df1['persnr']=df1.index
df1['training']=0
## Now copy the DataFrame (i.e. generate observations for second year)
df2=df1.copy()
df2['year']=2
## Training only in year 2:
df2['training']=((df2['ability']+np.random.normal(0,10,n)>=100))
## Generate DataFrame that spans both years by appending the two data frames
df=pd.concat([df1,df2], sort=False)
df['sales']= 10000 + df.training*5000 + df.ability*100 + df.year*2000
                + np.random.normal(0,4000,2*n)
```

Note:

- The script generated a data frame simulating two years of data in which
  - Sales of each subject are observed in each year
  - training is affected by ability
  - subjects are only trained in year 2

Now analyze the generated data:

- Run an OLS regression of sales on training and year
- Define the time and entity indices
- Run a fixed effects regression

But note important caveats:

1. When you want to interpret the results of a Fixed Effects regression causally, a key underlying assumption is the so-called *common trend assumption*
   - That is „treatment“ and „control“ units follow the same underlying time trend
   - This is a key identifying assumption
2. When the treatment C_{it} hardly varies over time it is hard to evaluate the causal effect effect ρ
   - In the extreme when C_{it} is completely stable then $C_{it} = \overline{C}_{it}$
   - Not identifying a significant effect in the data then does not necessarily imply that there is no such effect
3. Fixed effects can only eliminate *time constant* omitted variables
   - If the treatment is correlated with time varying unobserved variables omitted variable issues remain

Conclusion

- Two important tasks that we want to solve when analyzing data:
  - Prediction: When we observe X_i, how can be predict Y_i?
  - Causality: When we change X_i, how does this change Y_i?
- When predicting a variable we estimate the conditional expectation $E[Y|X]$
  - We can approximating $E[Y|X]$ through regressions which yields simple & interpretable functions
  - When X is highly multidimensional and $E[Y|X]$ highly non-linear other machine learning algorithms may be preferable
- When we want to identify causal effects of active changes in X, knowing $E[Y|X]$ is not enough
  - We can cleanly estimate causal effects if X is exogenously assigned such as in an experiment
  - A comprehensive set of control variables or panel data and within-unit variation of our variable of interest X can help to overcome biases