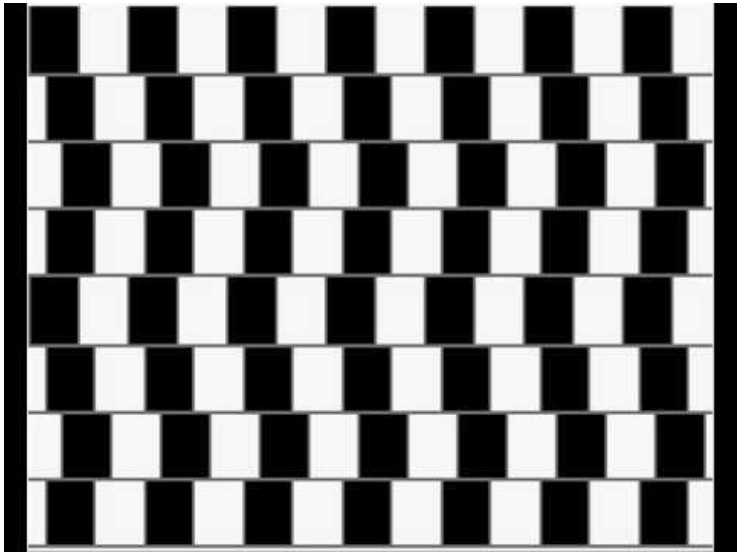
UNIVERSITY OF EDINBURGH
Business School

Data Illusions in Credit Risk Management: What Lies Behind Our (Potential) Mistakes



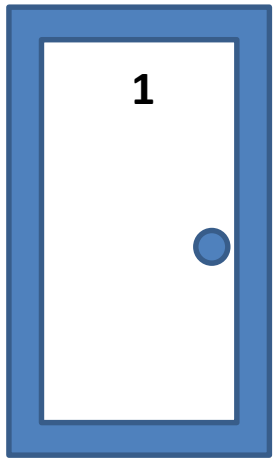
Fernando Moreira
University of Edinburgh Business School

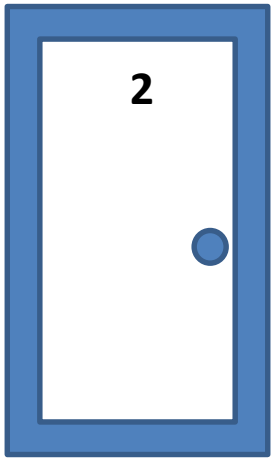
Credit Scoring and Credit Control XVI Conference
30 August 2019

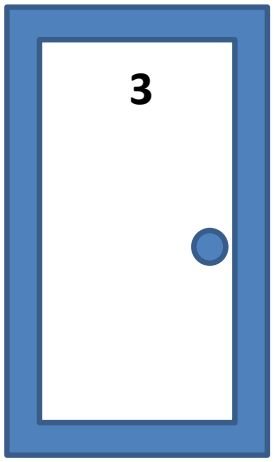
UNIVERSITY OF EDINBURGH
Business School

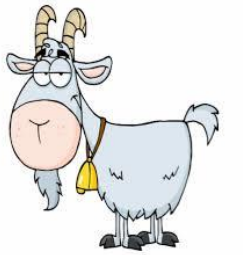
An initial puzzle…

- Let's Make a Deal (the Monty Hall problem)



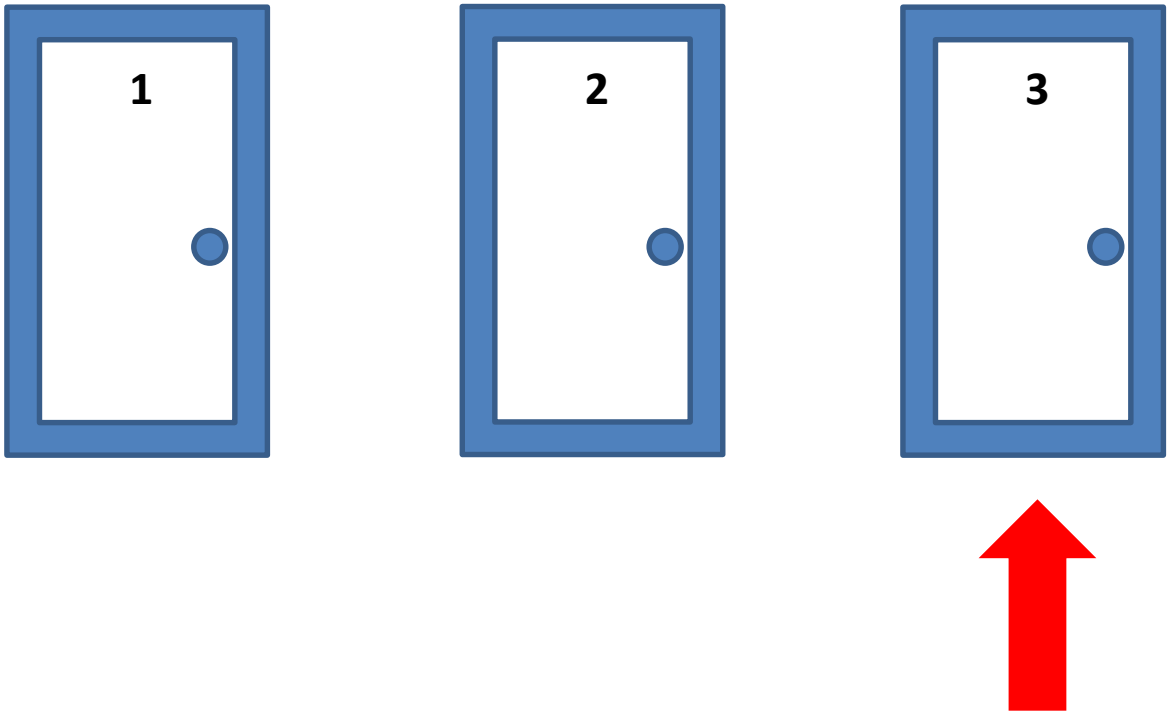






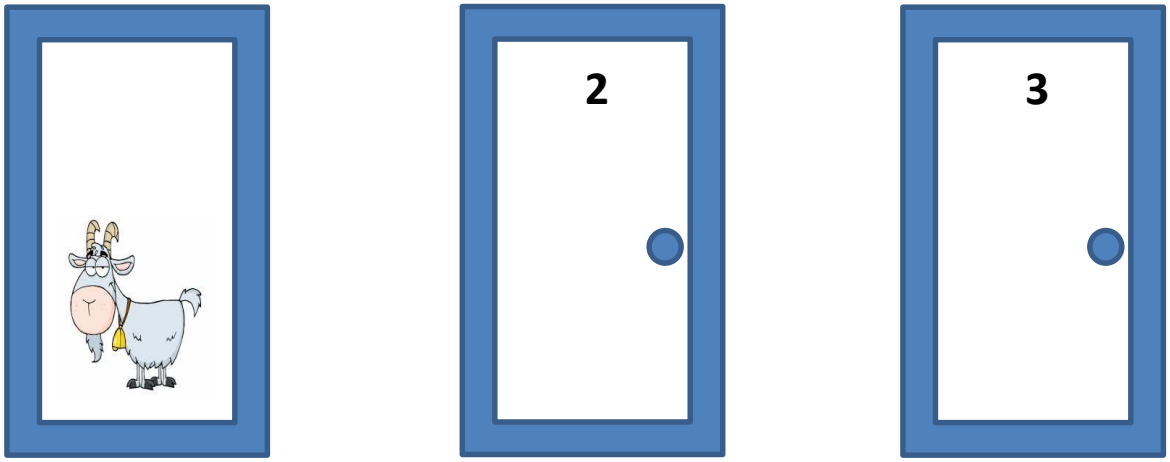
An initial puzzle…

- Let's Make a Deal (the Monty Hall problem)



An initial puzzle…

- Let's Make a Deal (the Monty Hall problem)



The show presenter opens one of the other two doors where (he knows) there is a goat (e.g. 1)

Then you are given the opportunity to change the door you had chosen (i.e. in this example from 3 to 2). Should you do it? Would this increase your probability of winning the car?

An example

- "Unlikeliness to pay" (UP) is one of the criteria used to define default
  - Regulation EU, No 575/2013, Article 178, paragraph 1(a)
- Borrowers are asked to report UP events if they happen
  - this is only one of the ways to identify UP; other alternatives are possible
- Such events could refer to job loss, illness, divorce, family bereavement, etc.
- Nevertheless, classifying a borrower as defaulter just because they have reported a particular event (e.g. divorce) may be misleading

UNIVERSITY OF EDINBURGH
Business School

An example

- The occurrence of UP events doesn't necessarily result in default and, even if default happens, it may be due to other reasons
- This could lead to an overidentification of defaults, which would result in overestimated regulatory capital and excessive provisions
- A better understanding of the causes of default would also contribute to improvements in collection procedures

An analysis

- (Hypothetical) Random sample: 800 borrowers in a portfolio
  - 400 have reported 'unlikeliness to pay' events (*UP*)
  - 400 have not reported 'unlikeliness to pay' events (*non-UP*)
- After some time, we check who defaulted:

	Default (a)	Non-default (b)	Total (c) = (a) + (b)	% default (a)/(c)
UP	200	200	400	50%
Non-UP	160	240	400	40%

That is, 50% of the UP borrowers defaulted while 40% of the non-UP borrowers defaulted

A deeper analysis …

- Assume we suspect that the region where borrowers live affects the occurrence of default (e.g. due to behavioural/ cultural or economic reasons)
- We then split our sample (800 borrowers) into two groups (Regions A and B)

Region A				
	Default (a)	Non-default (b)	Total (c) = (a) + (b)	% default (a)/(c)
UP	180	120	300	60%
Non-UP	70	30	100	70%

Region B				
	Default (a)	Non-default (b)	Total (c) = (a) + (b)	% default (a)/(c)
UP	20	80	100	20%
Non-UP	90	210	300	30%

UNIVERSITY OF EDINBURGH
Business School

Now, compare the results

Whole sample				
	Default (a)	Non-default (b)	Total (c) = (a) + (b)	% default (a)/(c)
UP	200	200	400	50%
Non-UP	160	240	400	40%

Region A				
	Default (a)	Non-default (b)	Total (c) = (a) + (b)	% default (a)/(c)
UP	180	120	300	60%
Non-UP	70	30	100	70%

Region B				
	Default (a)	Non-default (b)	Total (c) = (a) + (b)	% default (a)/(c)
UP	20	80	100	20%
Non-UP	90	210	300	30%

Conflicting results

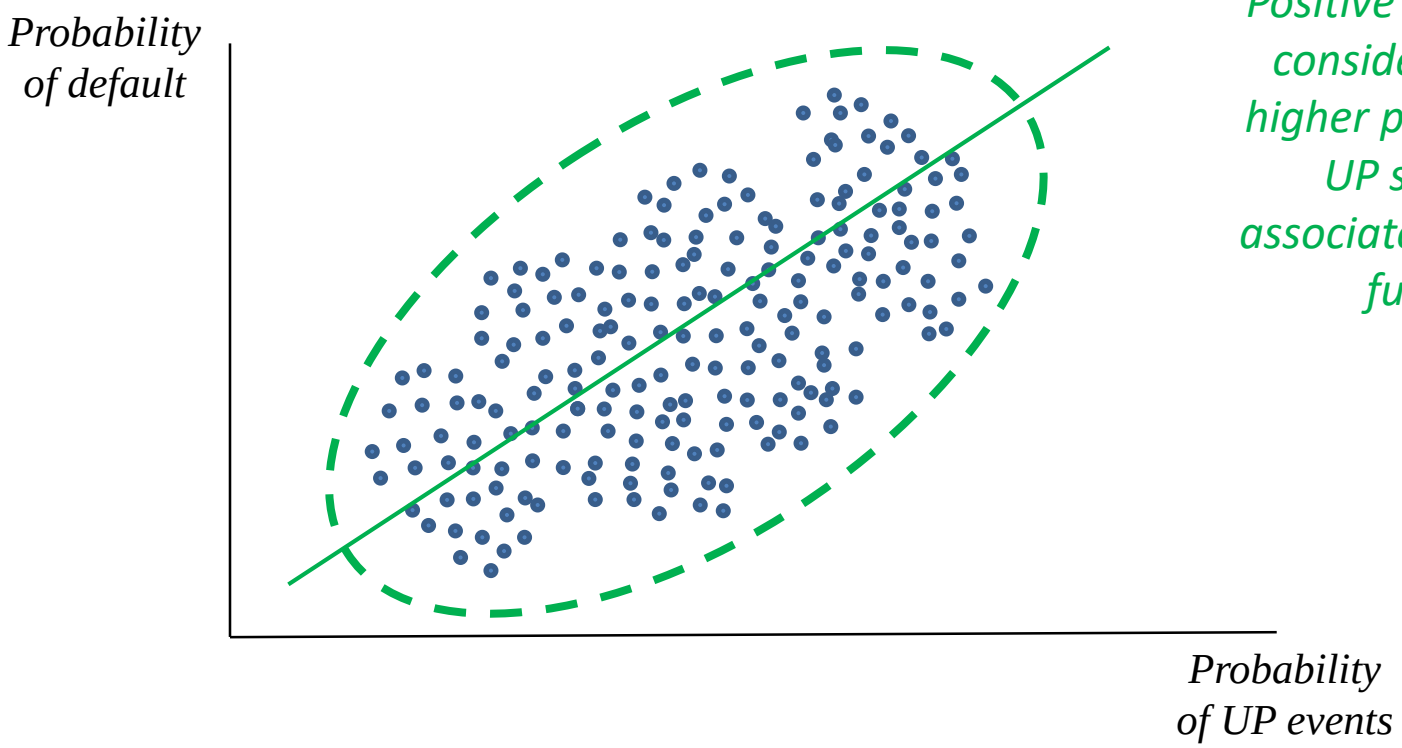
- Which result (whole sample or two groups) should be used?
- If the whole sample is used, UP events seem to be related to more default cases
  - then we should use those events to classify the respective borrowers as defaulters?
- If the split sample is used, both groups (Regions A and B) suggest that UP events are not a good predictor of default
- Would this mean that we should see UP events as good indicators of default when region is unknown and the opposite when it is known?

Translation into regression language

- In regression language:
  - groups or subsamples = “control for” variables (i.e. to include them in the regressions)
- Therefore, the problem above is equivalent to the question of what (control) variables we should include in our regressions
  - adding inappropriate variables or refraining from include pertinent variables would lead to wrong conclusions

Graphical illustration

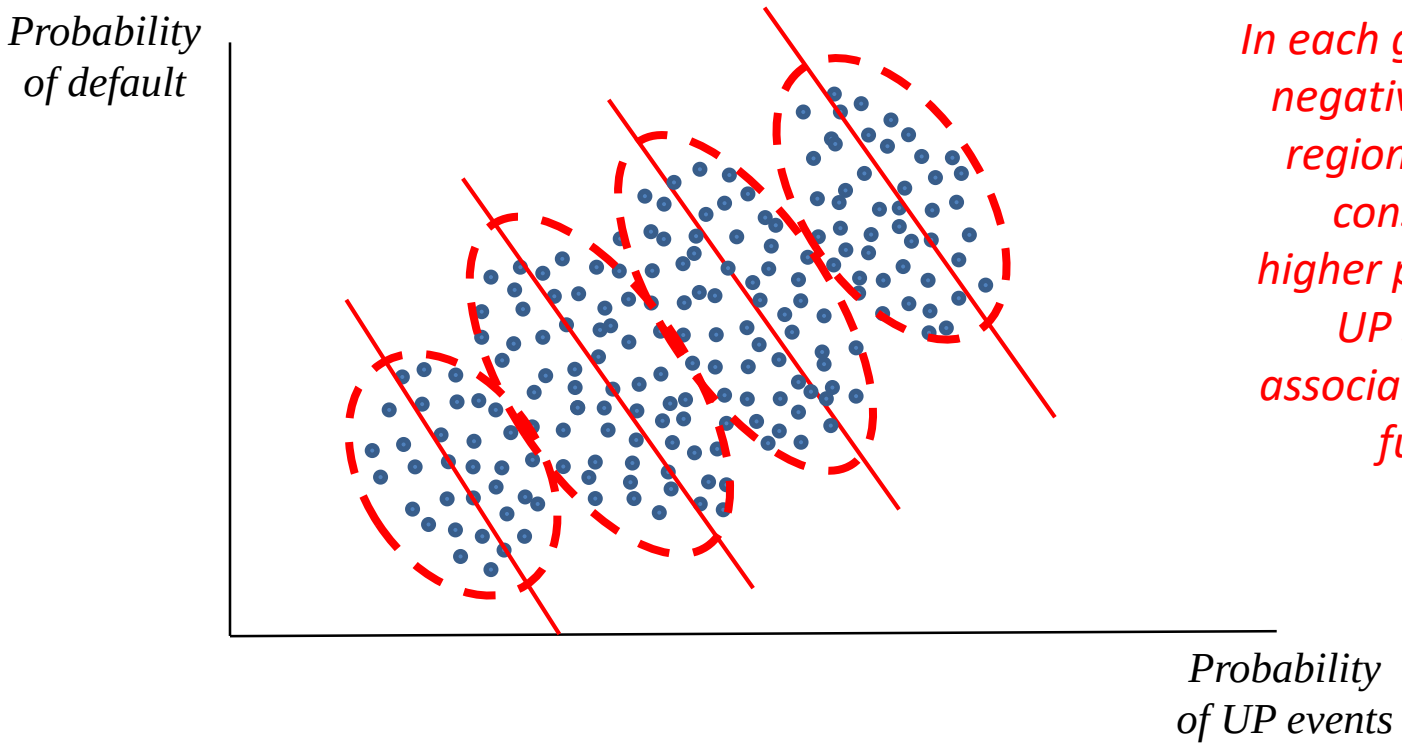
A slightly different example for the purpose of visual illustration: relationship between the probability of UP events (if we could estimate it) and the future probability of default (PD) for the respective borrowers



Positive slope without considering region: higher probabilities of UP shocks are associated with higher future PD

Graphical illustration

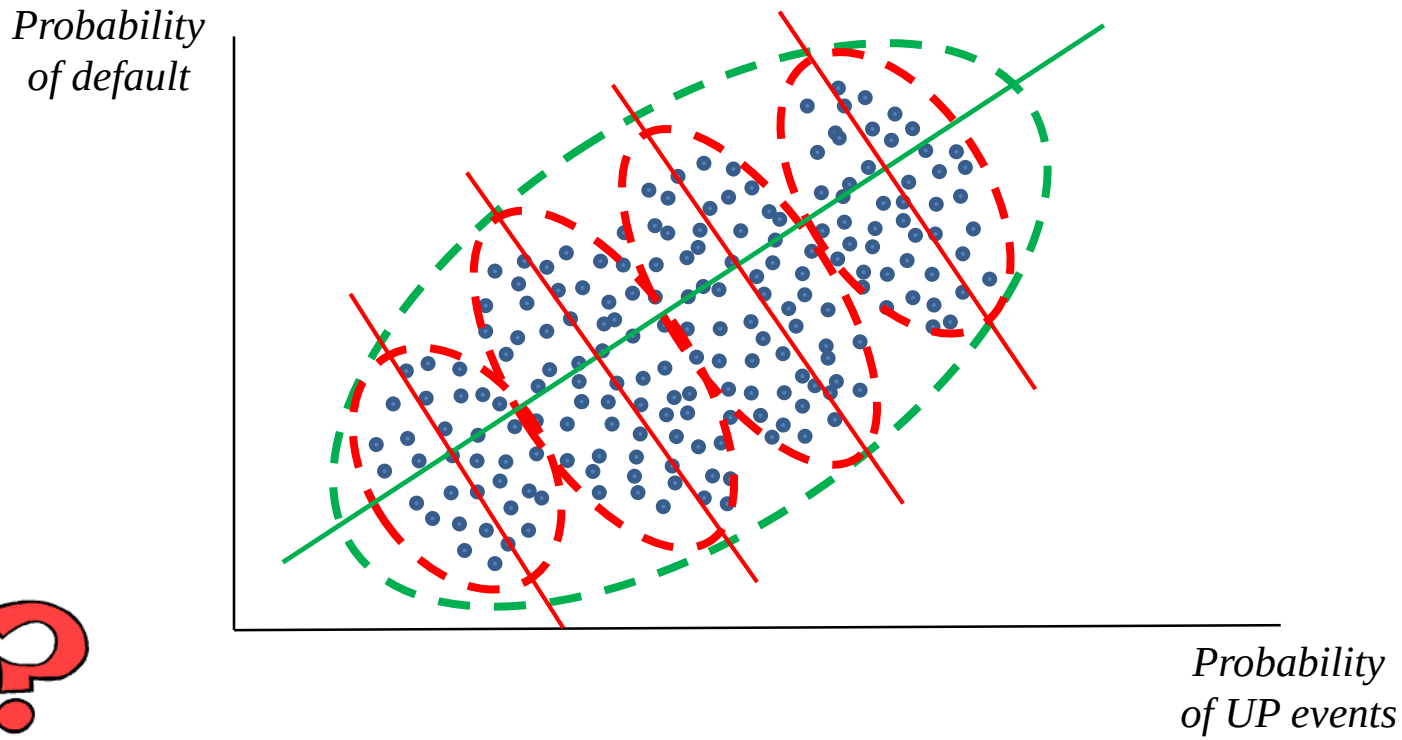
Assume we split the sample into four groups representing four regions



In each group (regions), negative slope when region is taken into consideration: higher probabilities of UP shocks are associated with lower future PD

Graphical illustration

In sum…



So, which slope gives the right answer?

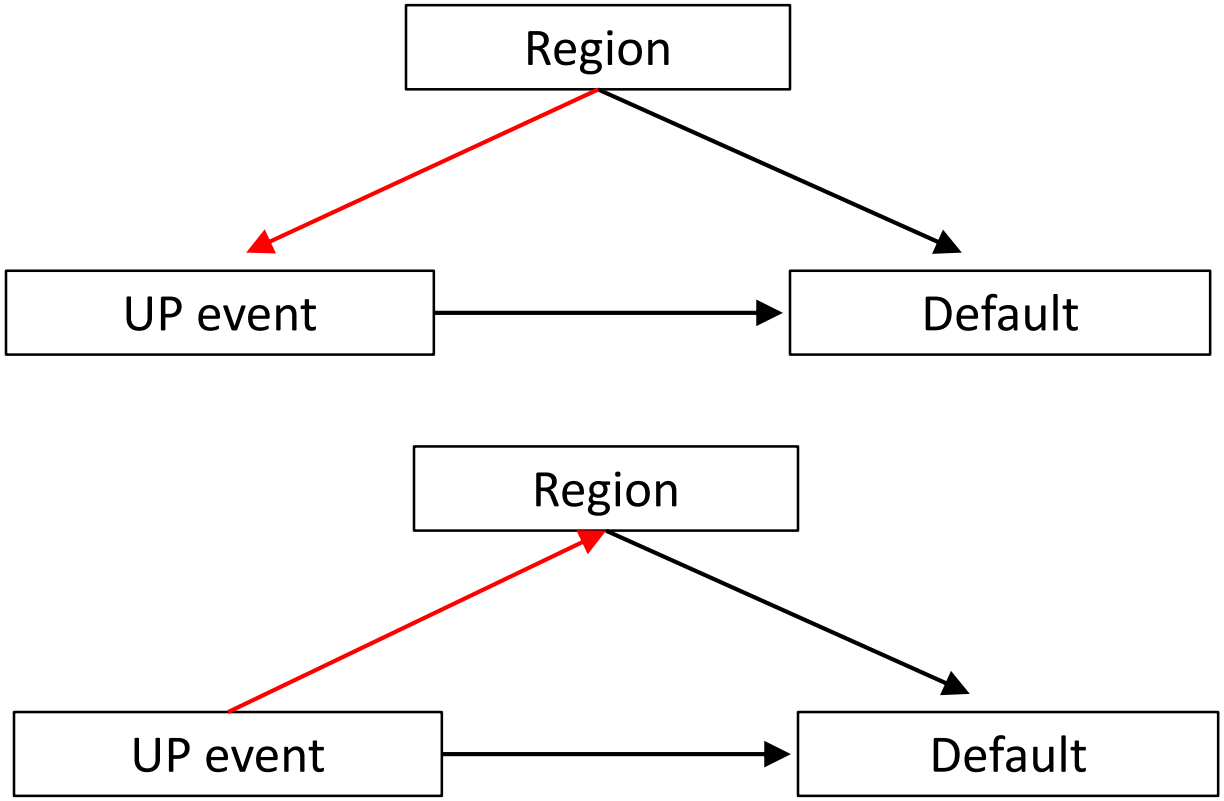
The solution

- It depends on…
- … the ***data generating process***
- That is, on how the variables are "created" ("generated")
- However, we may not know this. Data sets look all the same!

Borrower ID	UP event (1 = Yes, 0 = No)	Region	Default (1 = Yes, 0 = No)
AA2459	1	A	1
CX3010	0	B	1
FL7652	0	A	0
…	…	…	…
ZD0417	1	B	0

The solution

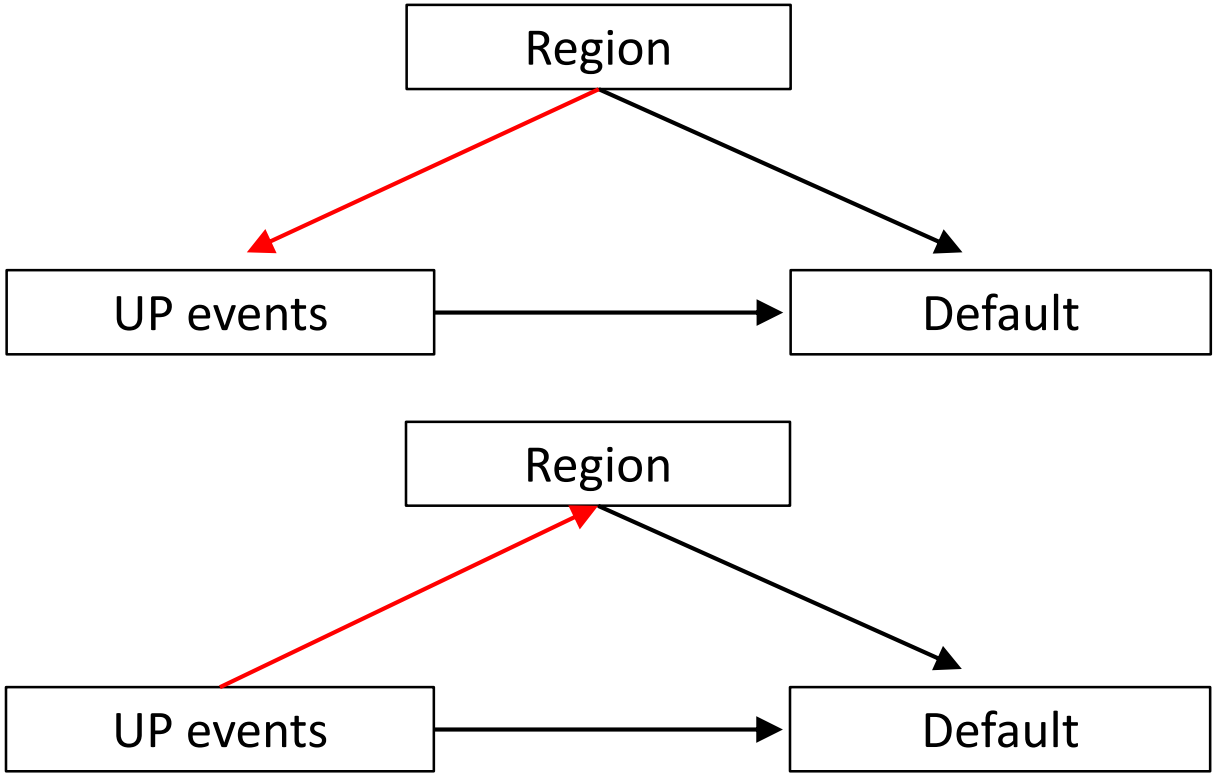
- In our original example, would the previous data set be represented by which of the following?



This may remind you of the classical question: which came first?

The solution

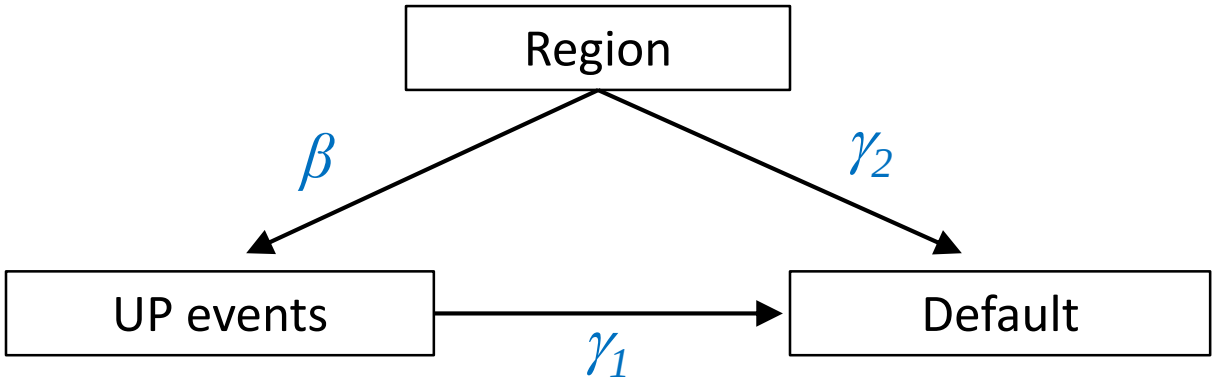
- If represented by the first, we should rely on the analysis based on the subgroups (because region affects UP events, we need to "control for" it)
- If the second is better, we would *not* split the sample



How to find out the most suitable 'process' (path)?

- A suggestion: **Structural Equation Modelling (SEM)**
- In short, propose possible processes (based on judgement or prior information) and test their feasibility regarding the data available
- Broadly speaking, two steps (after suggesting possible models)
  - check whether models are feasible (i.e. cannot be statistically rejected)
  - among the ones that pass the feasibility test, choose the most suitable (best-fit) one (e.g., based on Information Criteria tests)

Structural Equation Models



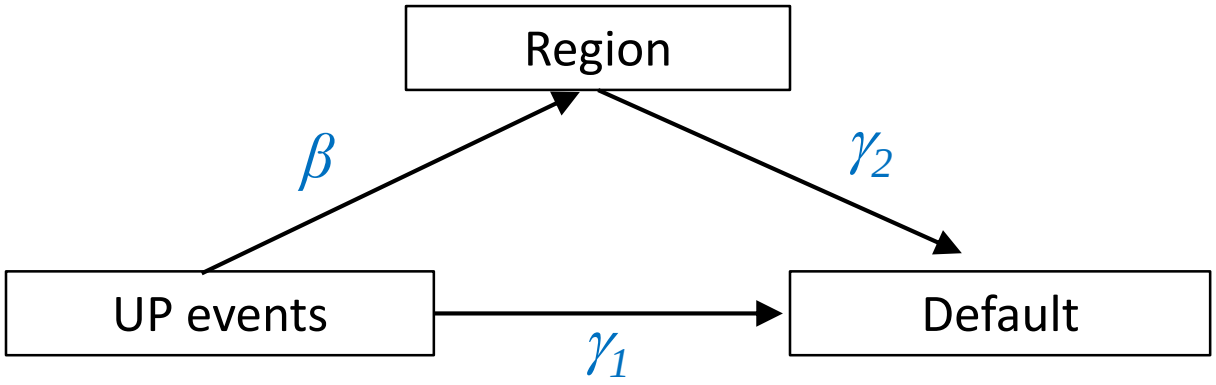
- The diagrams are 'converted' into equations (regressions). In our first assumed model (above):

$$UP = \alpha + \beta * R + \varepsilon$$

$$D = \mu + \gamma_1 * UP + \gamma_2 * R + \upsilon$$

- where: R = region; UP = UP events; D = default; ε and υ are error terms
- Here, we *should* control for region
  - Default depends on region

Structural Equation Models



- In the second model (above), the equations are:

$$R = \alpha + \beta * UP + \varepsilon$$

$$D = \mu + \gamma_1 * UP + \gamma_2 * R + \upsilon$$

- where: R = region; UP = UP events; D = default; ε and υ are error terms
- Here, we *should <u>not</u>* control for region
  - *controlling for region would 'block' one of the paths through which UP events may affect default*
  - *plugging the first equation into the second one makes R disappears*

Is the model suitable?

- Usual criteria to guide our decision (normally provided by statistical packages designed for SEM estimation such as LISREL, Amos, Stata, Tetrad or R) :

Criteria		Acceptable level	Interpretation
Chi-square	χ^2	> 0.10 (for 10% confidence level)	For values greater than 0.10, the null hypothesis in favour of fit cannot be rejected
Root Mean Square Error of Approximation	RMSEA	0.05 to 0.08	Values in this range indicate close fit
Comparative Fit Index	CFI	0 (no fit) to 1 (perfect fit)	At least 0.90 indicates good fit
Tucker-Lewis Index	TLI	0 (no fit) to 1 (perfect fit)	At least 0.90 indicates good fit

Best-fit model

- If more than one model pass the initial test (i.e. they cannot be rejected), use Alkaike Information Criteria (AIC) and the Bayesian Information Criteria (BIC) measures
  - the model with the smallest AIC and BIC values should be chosen
  - these statistics are also provided by specialised software such as those mentioned before

Conclusions

- Data generating process is our main guide to decide what variables should be taken into consideration in our (econometrics) models
- Usually, we do not know the "process" that generated the data we have in hand
  - or worse … we believe we know the process but we may be wrong
  - SEM can help us to find the most likely process among the ones we can think of
  - expert judgment is extremely helpful (processes/models should make sense)
- This issue is important in several fields

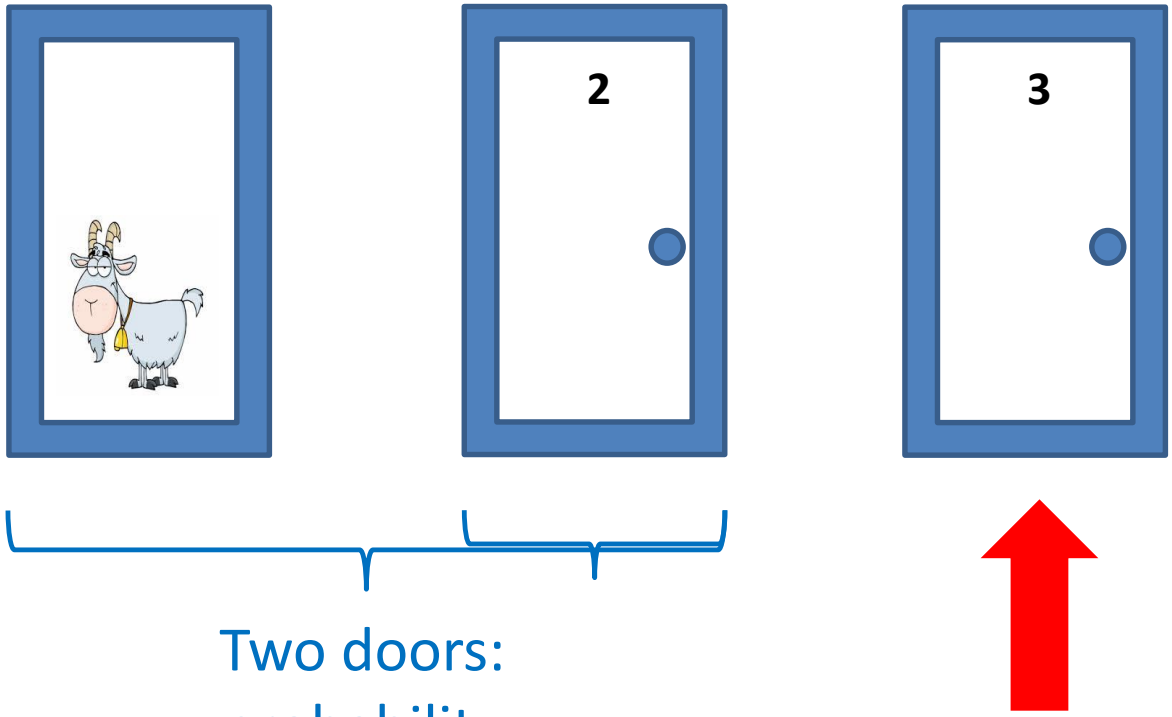
Basic references

- **Data generating process**
  - Pearl, J. (2009). *Causality: Models, Reasoning and Inference*. Cambridge: Cambridge University Press, 2nd edition (in particular, Chapter 6)

- **Structural Equation Model (SEM)**
  - Schumacker, R., Lomax, R. (2016). *A Beginner's Guide to Structural Equation Modeling*. New York: Routledge, 4th edition (in particular, Chapter 7)

Back to your initial puzzle

- Let's Make a Deal (the Monty Hall problem)



Back to our initial puzzle

- The Monty Hall problem – solution

(reminder: you chose door 3)

In green oval: what the presenter is allowed to show

Door 1	Door 2	Door 3	Outcome if you change	Outcome if you stay
Car	Goat	Goat	Win	Lose
Goat	**Car**	Goat	Win	Lose
Goat	Goat	**Car**	Lose	Win

- Hence, if you switch doors, your probability of winning is 2/3, rows 1 and 2 (while 1/3 if you keep your original choice, door 3)
- This happens because the presenter is forced to open a door that neither has the car nor was chosen by you (this is the "*data generating process*")