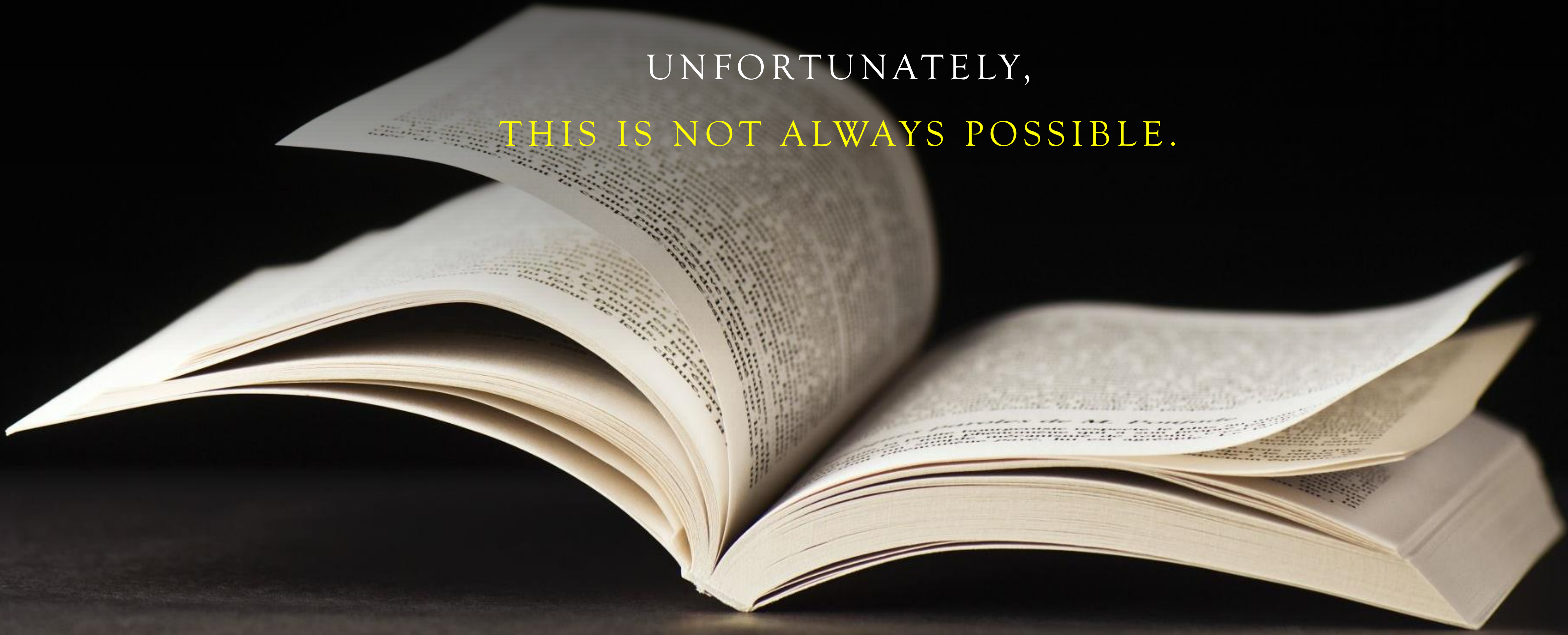
The plot of our course

FELIPE BUCHBINDER

WHEN WE BUILD A REGRESSION MODEL, WE WOULD LIKE IT TO CONTAIN ALL X'S THAT ARE IMPORTANT TO EXPLAIN OUR Y...

UNFORTUNATELY,

THIS IS NOT ALWAYS POSSIBLE.



Why not?

We might not have all the variables

Some variables may not be observable

Car insurance

$$\text{Expected cost} = \beta_0 + \beta_1 \cdot X + \epsilon$$

Observable variables:

- Age
- Gender
- Lives in a metropolitan area?
- (any other ideas?)

Car insurance

$$\text{Expected cost} = \beta_0 + \beta_1 \cdot X + \epsilon$$

Unobservable variables:

- Driving recklessly
- Goes to party often?
- Health issues (e.g. likelihood of syncope due to heart arrythmia)
- (any other ideas?)

Another example:

Impact of democracy on child mortality

Ross, M. (2006). Is democracy good for the poor?. *American Journal of Political Science*, *50*(4), 860-874.

$$\text{Child Mortality} = \beta_0 + \beta_1 \cdot \text{Democracy} + \beta_2 \cdot X + \epsilon$$

Observable variables

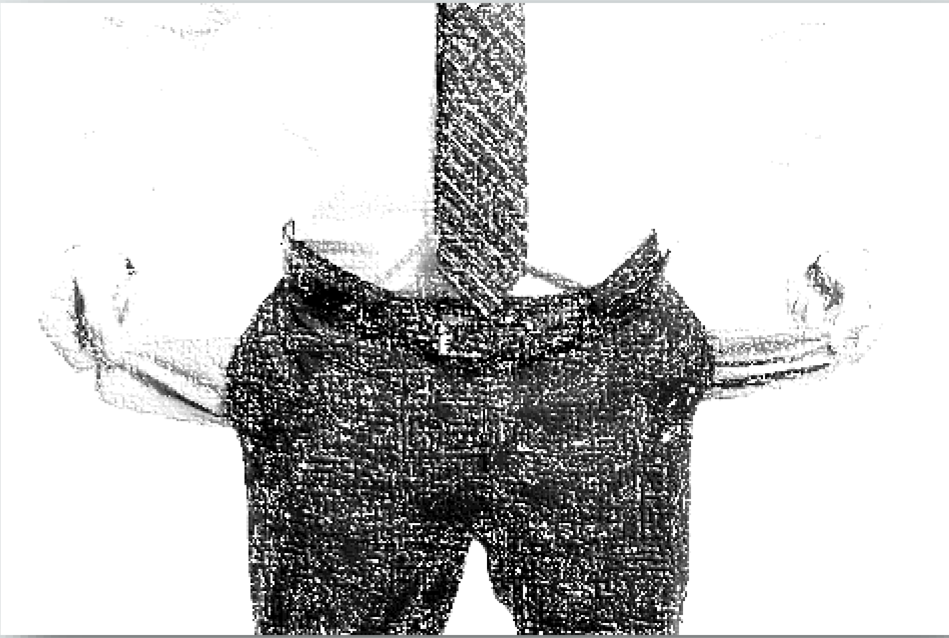
- GDP per capita
- Investment in health per capita
- Number of hospitals per capita
- (any other ideas?)

$$\text{Child Mortality} = \beta_0 + \beta_1 \cdot \text{Democracy} + \beta_2 \cdot X + \epsilon$$

Unobservable variables

- Health habits
- Dietary habits
- Cultural aspects regarding the caring and nourishing of children
- Existence of conflicts zones
- Etc.

Variables that differ between entities but are not observable are called **unobserved heterogeneities**.



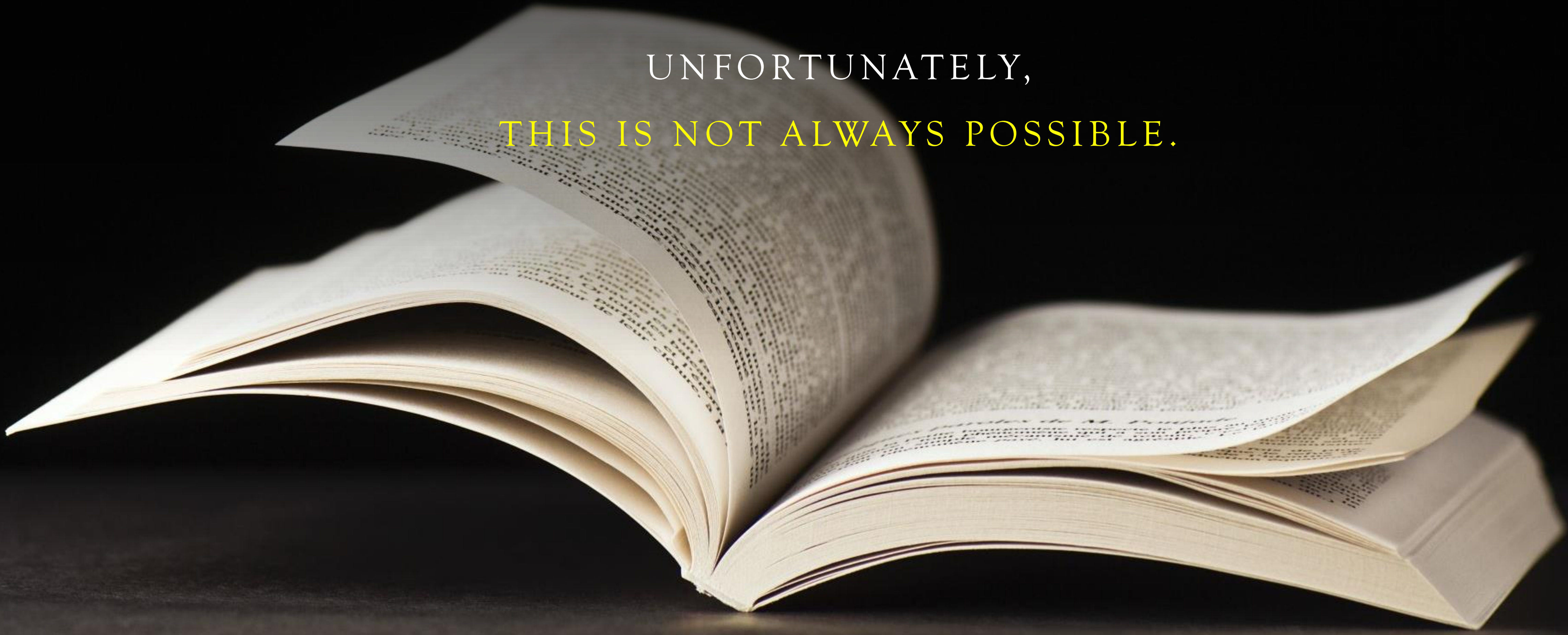
We might not have all the variables

Some variables may not be observable

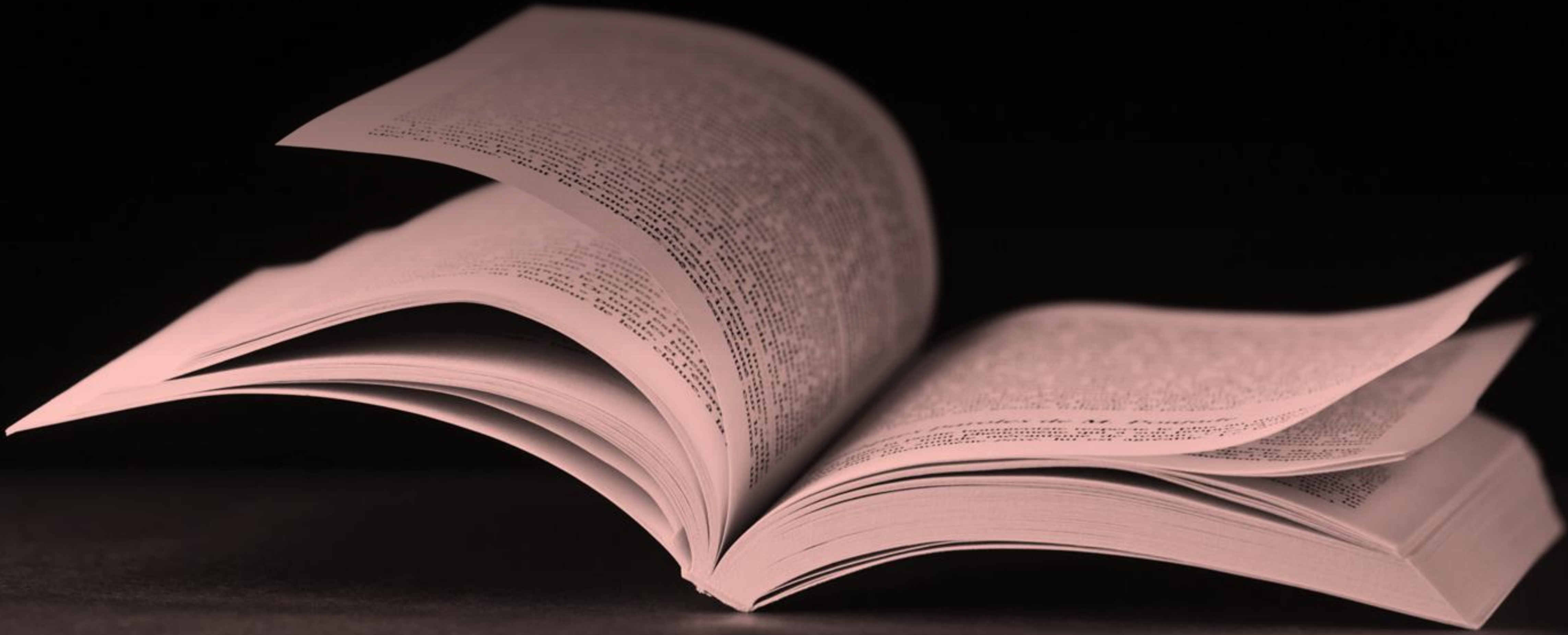
WHEN WE BUILD A REGRESSION MODEL,
WE WOULD LIKE IT TO CONTAIN ALL X'S
THAT ARE IMPORTANT TO EXPLAIN OUR Y…

UNFORTUNATELY,

THIS IS NOT ALWAYS POSSIBLE.



WHEN AN IMPORTANT X IS MISSING IN OUR MODEL, BAD THINGS CAN HAPPEN.



For example…?

Biased regression coefficients

Linear Regression refresher

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$
$$\boldsymbol{\epsilon} \sim \mathrm{N}(\mathbf{0}; \sigma^2 \mathbf{I})$$

$$\mathbf{b} = \left(\mathbf{X}^{\mathbf{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathbf{T}}\mathbf{y}$$

b is unbiased, meaning $\mathbb{E}(\mathbf{b}) = \boldsymbol{\beta}$

Now suppose there's a variable, **U**, that affects **Y** but we fail to put it in our model. We simply calculate **b** without it!

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{U} + \boldsymbol{\epsilon}$$
$$\boldsymbol{\epsilon} \sim \mathrm{N}(\mathbf{0}; \sigma^2\mathbf{I})$$

$$\mathbf{b} = \left(\mathbf{X}^{\mathbf{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathbf{T}}\mathbf{y}$$

Now suppose there's a variable, **U**, that affects **Y** but we fail to put it in our model. We simply calculate **b** without it!

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{U} + \boldsymbol{\epsilon}$$
$$\boldsymbol{\epsilon} \sim \mathrm{N}(\mathbf{0}; \sigma^2\mathbf{I})$$

$$\mathbf{b} = \left(\mathbf{X}^{\mathbf{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathbf{T}}\mathbf{y}$$

In this case, **b** is no longer an unbiased estimate of $\boldsymbol{\beta}$!

$$\mathbf{b} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\boldsymbol{y}$$

$$= \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}(\mathbf{X}\boldsymbol{\beta} + \mathbf{U} + \boldsymbol{\epsilon})$$

$$= \underbrace{\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{X}}_{\mathbf{I}}\boldsymbol{\beta} + \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{U} + \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\boldsymbol{\epsilon}$$

$$= \boldsymbol{\beta} + \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{U} + \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\boldsymbol{\epsilon}$$

Taking the expected value...

$$\mathbb{E}(\mathbf{b}) = \boldsymbol{\beta} + \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{U} + \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\underbrace{\mathbb{E}\left(\mathbf{X}^{\mathrm{T}}\boldsymbol{\epsilon}\right)}_{0}$$

$$\therefore$$

$$\mathbb{E}(\mathbf{b}) = \boldsymbol{\beta} + \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{U}$$

Thus, in general, $\mathbb{E}(\mathbf{b}) \neq \boldsymbol{\beta}$:

$\mathbf{b}$ is a biased estimate of $\boldsymbol{\beta}$

I propose a name for this Theorem…

$$\mathbf{b} = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\boldsymbol{y}$$

$$= (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}(\mathbf{X}\boldsymbol{\beta} + \mathbf{U} + \epsilon)$$

$$= \underbrace{(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{X}}_{\mathrm{I}}\boldsymbol{\beta} + (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{U} + (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\epsilon$$

$$= \boldsymbol{\beta} + (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{U} + (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\epsilon$$

Taking the expected value…

$$\mathbb{E}(\mathbf{b}) = \boldsymbol{\beta} + (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{U} + (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\underbrace{\mathbb{E}(\mathbf{X}^\mathsf{T}\epsilon)}_{0}$$

$$\therefore$$

$$\mathbb{E}(\mathbf{b}) = \boldsymbol{\beta} + (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{U}$$

Thus, in general, $\mathbb{E}(\mathbf{b}) \neq \boldsymbol{\beta}$:

b is a biased estimate of $\boldsymbol{\beta}$

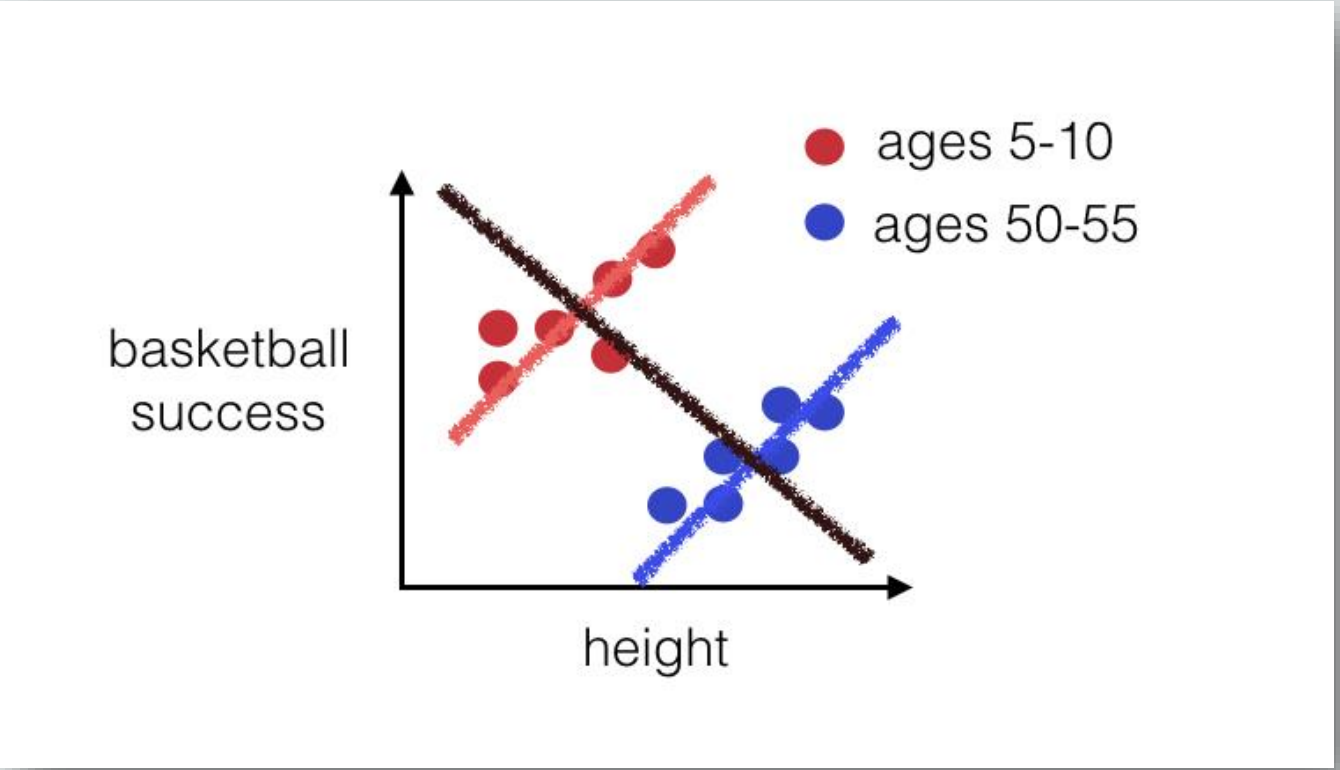
Maleficent's Theorem

A variable that wasn't invited to a regression will curse the coefficients of those that were, making them biased

An extreme scenario:

Simpson’s Paradox

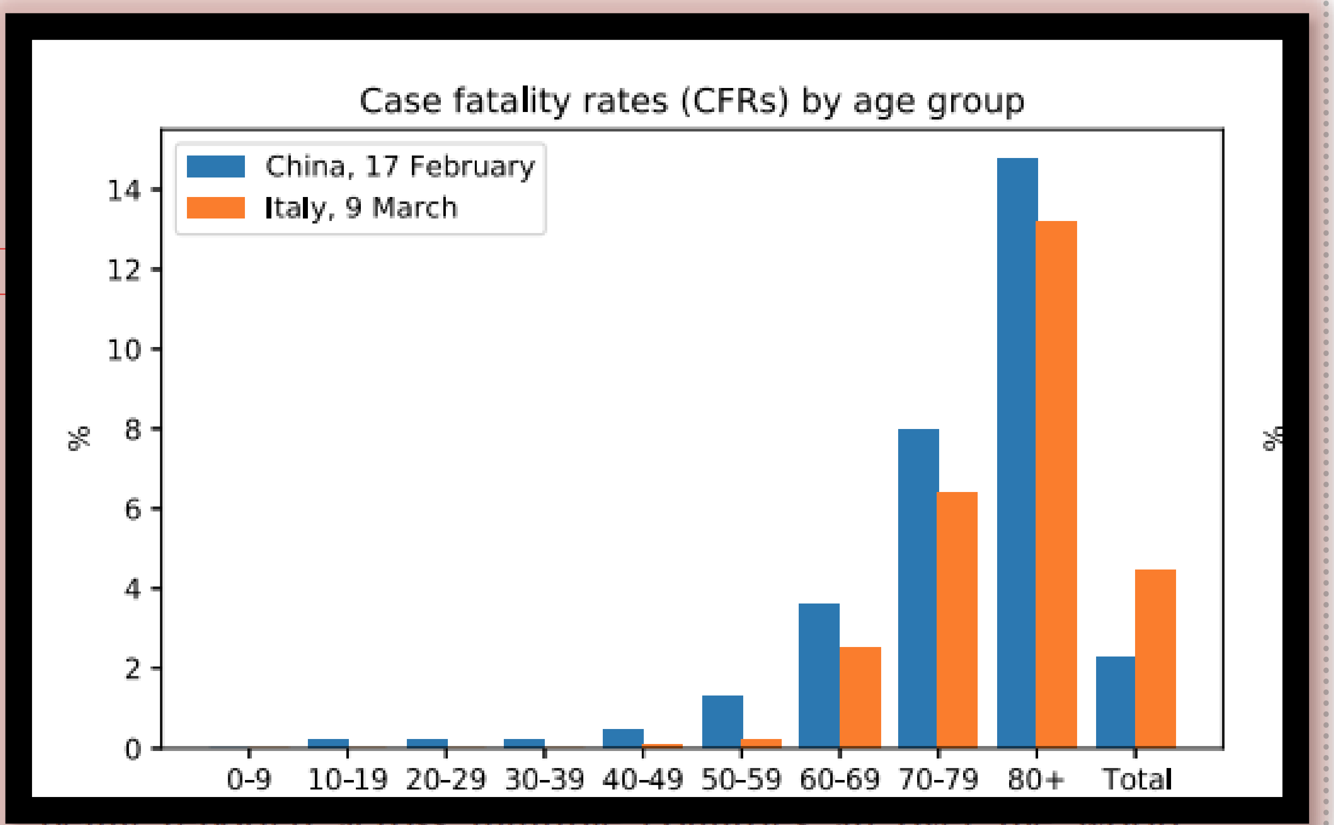
Wanna be good at Basketball? Be short!



Simpson's Paradox in COVID-19 Case Fatality Rates: A Mediation Analysis of Age-Related Causal Effects

Julius von Kügelgen, Luigi Gresele, and Bernhard Schölkopf

Abstract—We point out an instantiation of Simpson's paradox in COVID-19 case fatality rates (CFRs): comparing a large-scale study from China (February 17) with early reports from Italy (March 9), we find that CFRs are lower in Italy for every age group, but higher overall. This phenomenon is explained by a stark difference in case demographic between the two countries. Using this as a motivating example, we introduce basic concepts from mediation analysis and show how these can be used to quantify different direct and indirect effects when assuming a coarse-grained causal graph involving country, age, and case fatality. We curate an age-stratified CFR dataset with >750 k cases and conduct a case study, investigating total, direct, and indirect (age-mediated) causal effects between different countries and at different points in time. This allows us to separate age-related effects from others unrelated to age and facilitates a more transparent comparison of CFRs across countries at different stages of the COVID-19 pandemic. Using longitudinal data from Italy, we discover a sign reversal of the direct causal effect in mid-March, which temporally aligns with the reported collapse of the healthcare system in parts of the country. Moreover, we find that direct and indirect effects across 132 pairs of countries are only weakly correlated, suggesting that a country's policy and case

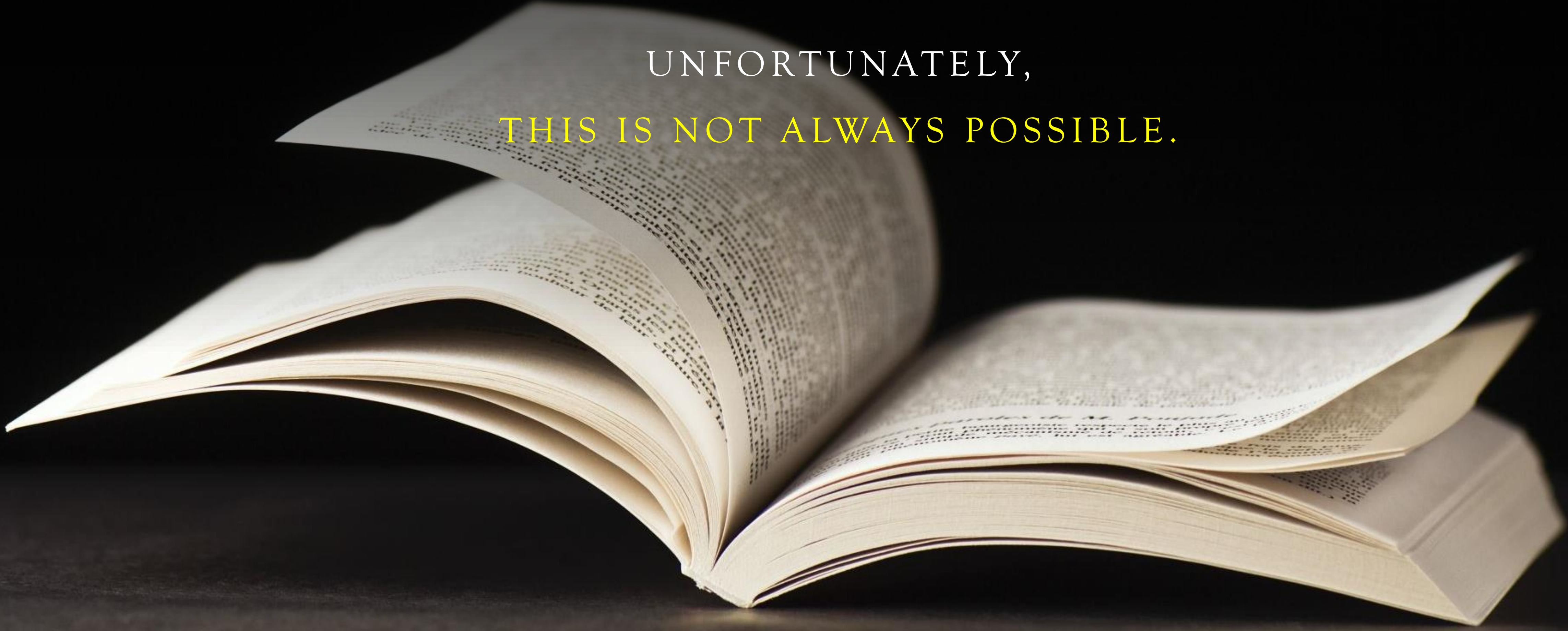


being reported across multiple countries all over the world, ultimately leading to the World Health Organization declaring it a pandemic on March 11, 2020 [1]. As of September 28, 2020, the pandemic led to more than 33 million confirmed cases and

WHEN WE BUILD A REGRESSION MODEL, WE WOULD LIKE IT TO CONTAIN ALL X'S THAT ARE IMPORTANT TO EXPLAIN OUR Y...

UNFORTUNATELY,

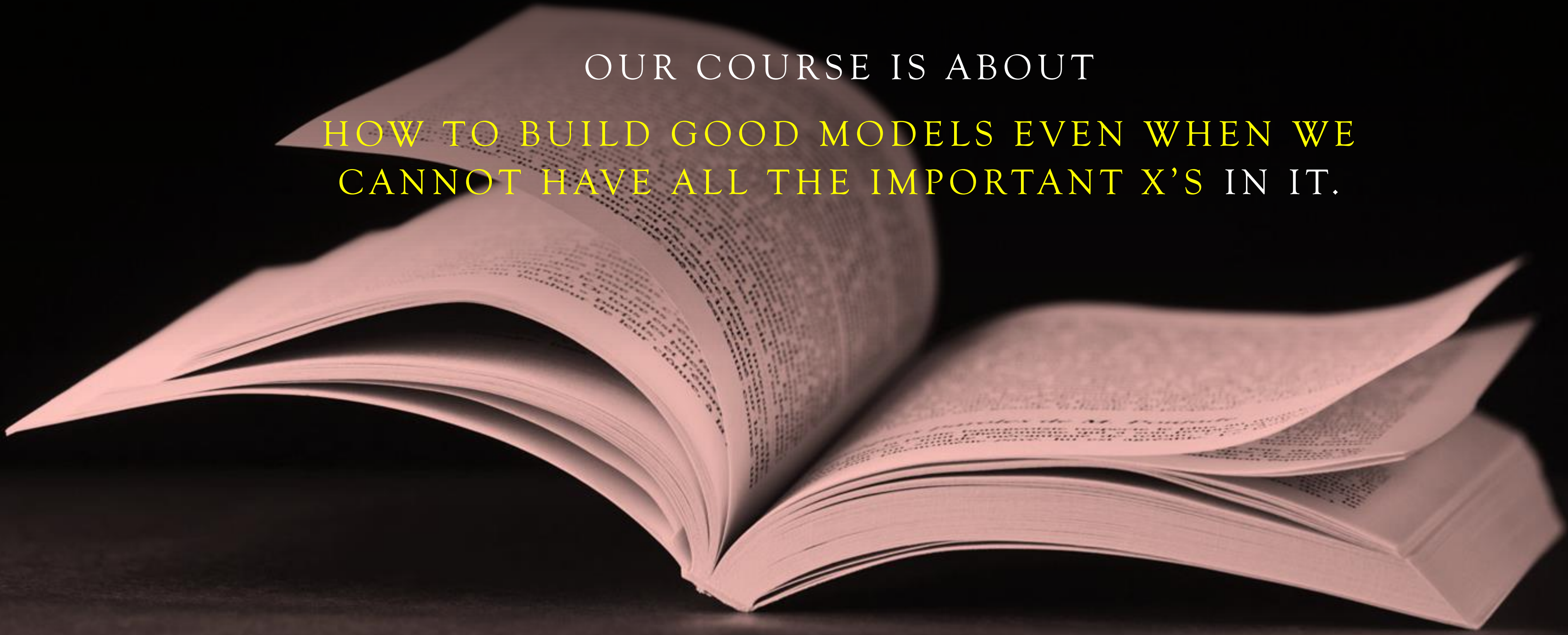
THIS IS NOT ALWAYS POSSIBLE.



WHEN AN IMPORTANT X IS MISSING IN OUR MODEL, BAD THINGS CAN HAPPEN.

OUR COURSE IS ABOUT

HOW TO BUILD GOOD MODELS EVEN WHEN WE CANNOT HAVE ALL THE IMPORTANT X'S IN IT.



Our strategy will be to
analyze things over time,
so we can have a feeling of how things
usually are.
If an X matters, it's effect should be
made visible by observing something
over time and comparing it with
others.


Why do we date before we get engaged?

$$\begin{pmatrix}\text{Will}\\ \text{I be happy?}\end{pmatrix} = \beta_0 + \beta_1 \begin{pmatrix}\text{Is s/he}\\ \text{fun?}\end{pmatrix} + \beta_2 \begin{pmatrix}\text{Do we}\\ \text{have chemistry?}\end{pmatrix} + \beta_3(\text{Does s/he love me?}) + \epsilon$$

$$\begin{pmatrix}\text{Will}\\ \text{I be happy?}\end{pmatrix} = \beta_0 + \beta_1\begin{pmatrix}\text{Is s/he}\\ \text{fun?}\end{pmatrix} + \beta_2\begin{pmatrix}\text{Do we}\\ \text{have chemistry?}\end{pmatrix} + \beta_3\underbrace{\left(\text{Does s/he love me?}\right)}^{\text{Unobservable}} + \epsilon$$

$$\begin{pmatrix}\text{Will} \\ \text{I be happy?}\end{pmatrix} = \beta_0 + \beta_1 \begin{pmatrix}\text{Is s/he} \\ \text{fun?}\end{pmatrix} + \beta_2 \begin{pmatrix}\text{Do we} \\ \text{have chemistry?}\end{pmatrix} + \beta_3 \overbrace{(\text{Does s/he love me?})}^{\text{Unobservable}} + \epsilon$$

Over time, you'll see how s/he usually acts towards you.

Unobservable

$$\begin{pmatrix}\text{Will}\\ \text{I be happy?}\end{pmatrix} = \beta_0 + \beta_1 \begin{pmatrix}\text{Is s/he}\\ \text{fun?}\end{pmatrix} + \beta_2 \begin{pmatrix}\text{Do we}\\ \text{have chemistry?}\end{pmatrix} + \beta_3 (\text{Does s/he love me?}) + \epsilon$$

Over time, you'll see how s/he usually acts towards you.

By dating other people
(not at once! ☺)
you'll be able to see how other people usually treat you and assess if s/he is special

$$\begin{pmatrix}\text{Will}\\ \text{I be happy?}\end{pmatrix} = \beta_0 + \beta_1\begin{pmatrix}\text{Is s/he}\\ \text{fun?}\end{pmatrix} + \beta_2\begin{pmatrix}\text{Do we}\\ \text{have chemistry?}\end{pmatrix} + \beta_3\overset{\text{Unobservable}}{(\text{Does s/he love me?})} + \epsilon$$

Over time, you'll see how s/he usually acts towards you.

By dating other people
(not at once! ☺)
you'll be able to see how other people usually treat you and assess if s/he is special

A less silly example:

Impact of democracy on child mortality

Ross, M. (2006). Is democracy good for the poor?. *American Journal of Political Science*, *50*(4), 860-874.

What control variables would you use in this model?

$$\text{Child Mortality} = \beta_0 + \beta_1 \cdot \text{Democracy} + \beta_2 \cdot X + \epsilon$$

What could X be?

- GDP per capita
- Investment in health per capita
- Number of hospitals per capita
- (any other ideas?)

Some things that affect child mortality are unobserved

$$\text{Child Mortality} = \beta_0 + \beta_1 \cdot \text{Democracy} + \beta_2 X + \boldsymbol{U} + \epsilon$$

What could $\boldsymbol{U}$ be?

- Health habits
- Dietary habits
- Cultural aspects regarding the caring and nourishing of children
- Existence of conflicts zones
- Etc.

Some things that affect child mortality are unobserved

$$\text{Child Mortality} = \beta_0 + \beta_1 \cdot \text{Democracy} + \beta_2 X + U + \epsilon$$

What could U be?

- Health habits
- Dietary habits
- Cultural aspects regarding the caring and nourishing of children
- Existence of conflicts zones
- Etc.

We can get a sense of these
by observing many countries over time

"**LOTS** OF THINGS ARE INVISIBLE, BUT WE DON'T KNOW HOW MANY BECAUSE WE CAN'T **SEE** THEM."

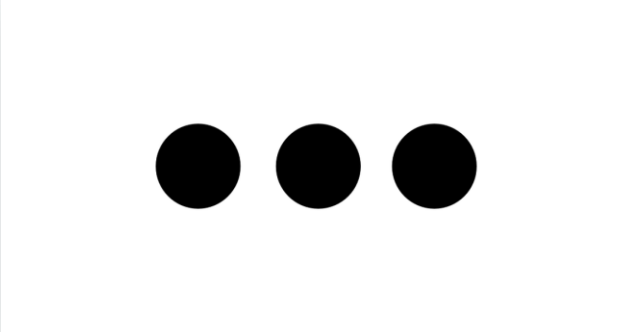

L'essentiel est invisible
pour les yeux

To deal with unobserved heterogeneities, we'll observe many entities over time

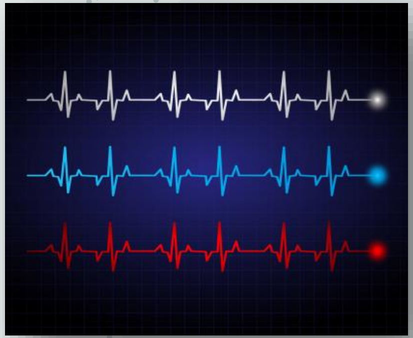
4 kinds of data

Cross-Sectional

Time Series



Pooled



Panel

The general panel data model

a.k.a. the most important equation in our course

The general panel data model

a.k.a. the most important equation in our course

$$y_{it} = \underbrace{\beta_0}_{\text{Intercept}} + \underbrace{\beta_1 X_{1it} + \cdots + \beta_p X_{pit}}_{\substack{\text{X's are called covariates} \\ \text{and are observed}}} + \underbrace{U_i}_{\text{Unobserved}} + \underbrace{\epsilon_{it}}_{\substack{\text{Idyossincratic} \\ \text{error}}}$$

or, in matrix notation,

$$\mathbf{Y_{it} = X_{it}\beta + U_i + \epsilon_{it}}$$

The big question is how to estimate $\boldsymbol{\beta}$ well even though we do not know $\mathbf{U}$, since it is unobserved.

The general panel data model

a.k.a. the most important equation in our course

$$y_{it} = \underbrace{\beta_0}_{\text{Intercept}} + \underbrace{\beta_1 X_{1it} + \cdots + \beta_p X_{pit}}_{\text{X's are called covariates and are observed}} + \underbrace{U_i}_{\text{Unobserved}} + \underbrace{\epsilon_{it}}_{\text{Idyossincratic error}}$$

or, in matrix notation,

$$\mathbf{Y}_{\mathbf{it}} = \mathbf{X}_{\mathbf{it}}\boldsymbol{\beta} + \mathbf{U}_{\mathbf{i}} + \boldsymbol{\epsilon}_{\mathbf{it}}$$

The big question is how to estimate $\boldsymbol{\beta}$ well even though we do not know $\mathbf{U}$, since it is unobserved.

Data is for each entity at each time → Panel Data

The general panel data model

a.k.a. the most important equation in our course

$$y_{it} = \underbrace{\beta_0}_{\text{Intercept}} + \underbrace{\beta_1 X_{1it} + \cdots + \beta_p X_{pit}}_{\text{X's are called covariates and are observed}} + \underbrace{U_i}_{\text{Unobserved}} + \underbrace{\epsilon_{it}}_{\text{Idyossincratic error}}$$

or, in matrix notation,

$$\mathbf{Y_{it}} = \mathbf{X_{it}}\boldsymbol{\beta} + \mathbf{U_i} + \boldsymbol{\epsilon}_{\mathbf{it}}$$

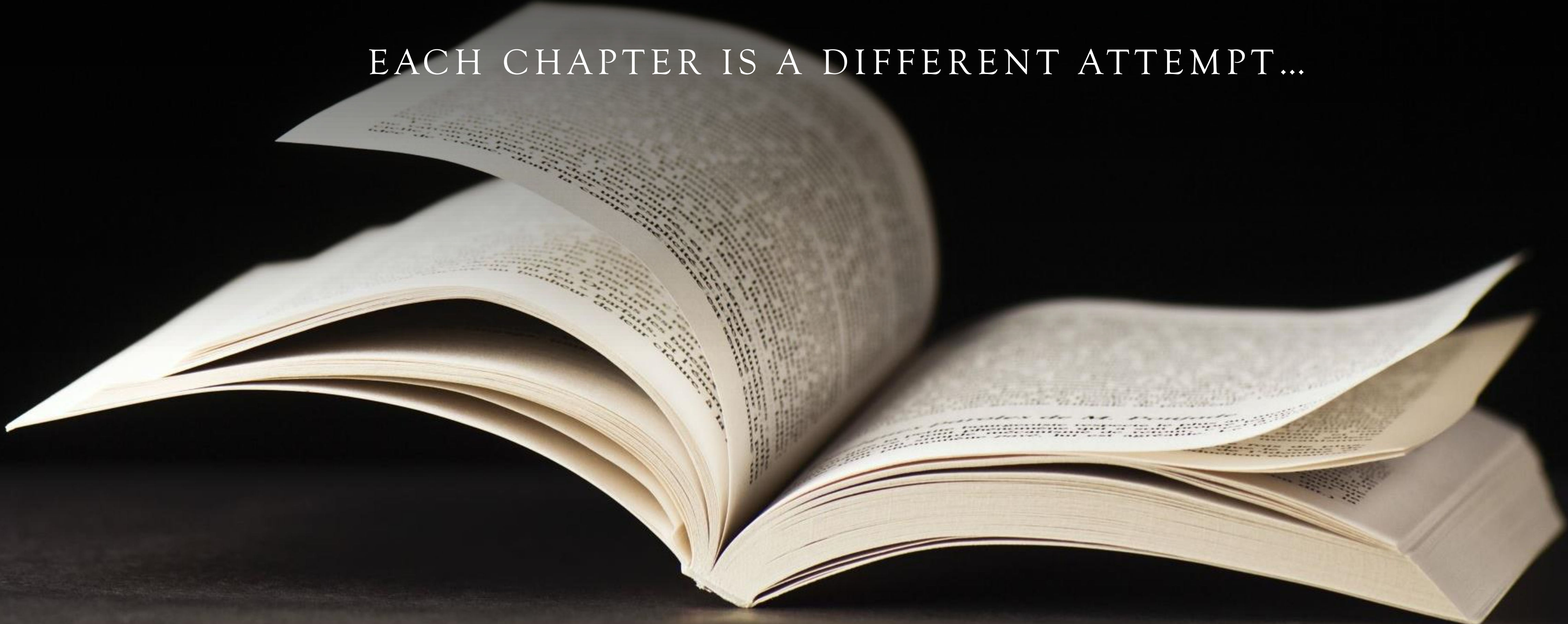
The big question is how to estimate $\boldsymbol{\beta}$ well even though we do not know $\mathbf{U}$, since it is unobserved.

Data is for each entity
at each time → Panel Data

Unobserved heterogeneity
is assumed to be a characteristic of the entity
that does not to vary through time
(entity-fixed effect)

OUR COURSE IS THE STORY OF HOW HUMANITY TACKLED THE CHALLENGE OF MAKING A REGRESSION MODEL WITHOUT KNOWING ALL THE X'S…

EACH CHAPTER IS A DIFFERENT ATTEMPT…

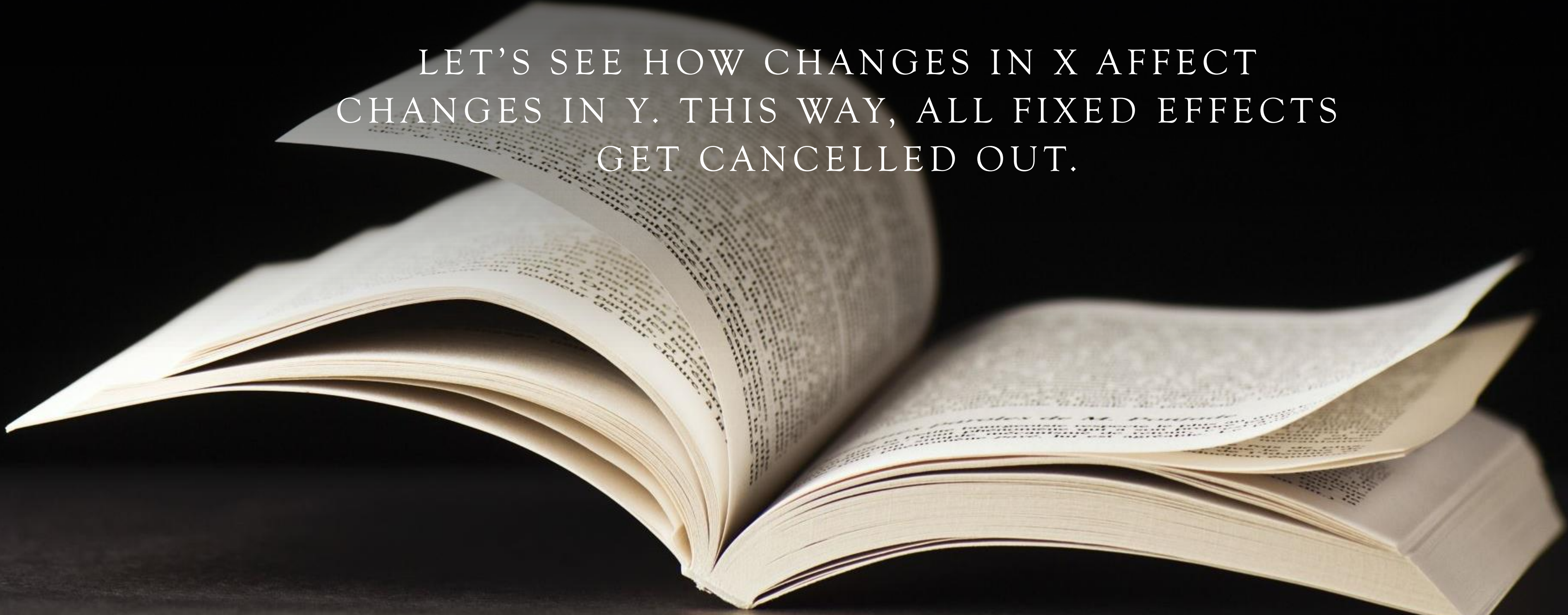


CHAPTER 1: POOLED REGRESSION

WHEN I FORGOT I HAD PANEL DATA AND SIMPLY RAN A TRADITIONAL REGRESSION ON EVERYTHING!

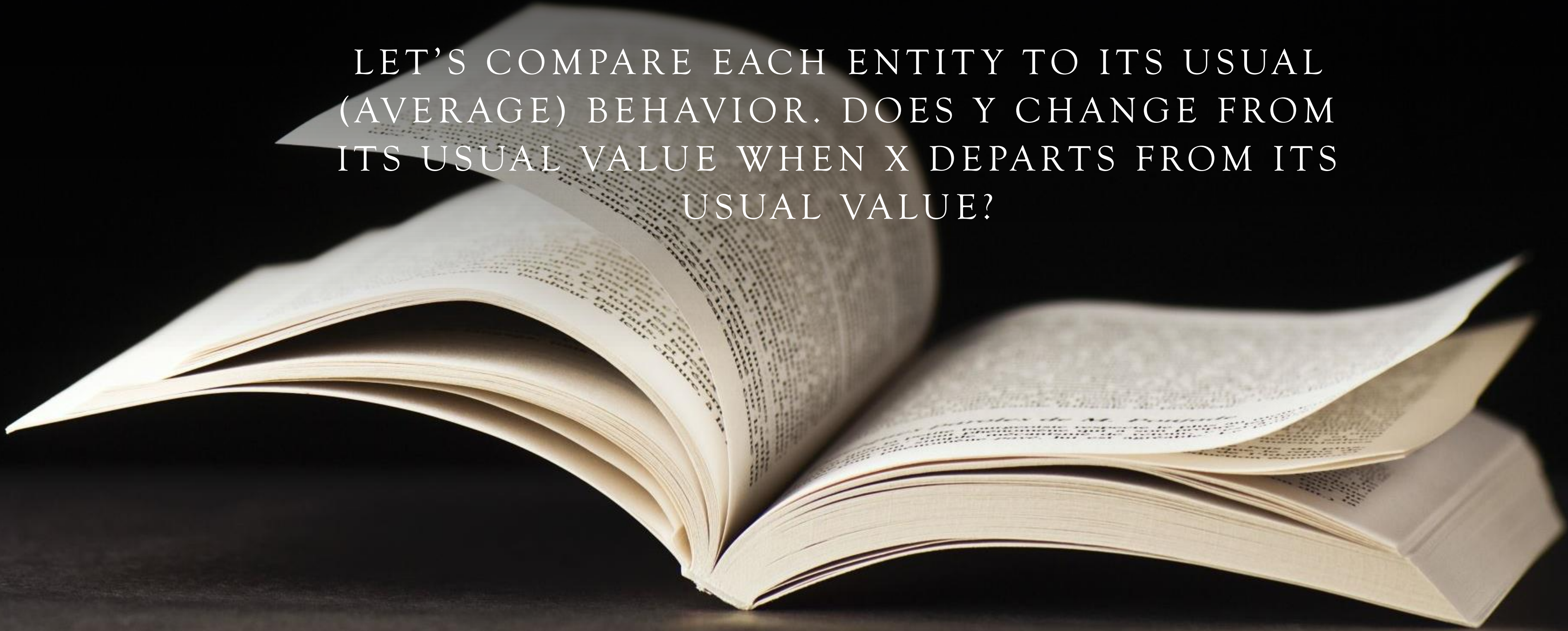
CHAPTER 2: FIRST DIFFERENCES

LET’S SEE HOW CHANGES IN X AFFECT CHANGES IN Y. THIS WAY, ALL FIXED EFFECTS GET CANCELLED OUT.



CHAPTER 3: FIXED EFFECTS

LET'S COMPARE EACH ENTITY TO ITS USUAL (AVERAGE) BEHAVIOR. DOES Y CHANGE FROM ITS USUAL VALUE WHEN X DEPARTS FROM ITS USUAL VALUE?



CHAPTER 4: RANDOM EFFECTS

A VERY COMPLICATED SOLUTION FOR A VERY SPECIFIC CASE. WHY DO WE STILL TEACH THIS STUFF?

CHAPTER 5: MODELS FOR RESIDUALS CORRELATED AS AN AR(1)

WHEN THINGS THAT HAPPEN IN VEGAS DON'T JUST STAY IN VEGAS…

CHAPTER 5: MODELS FOR RESIDUALS CORRELATED AS AN AR(1)

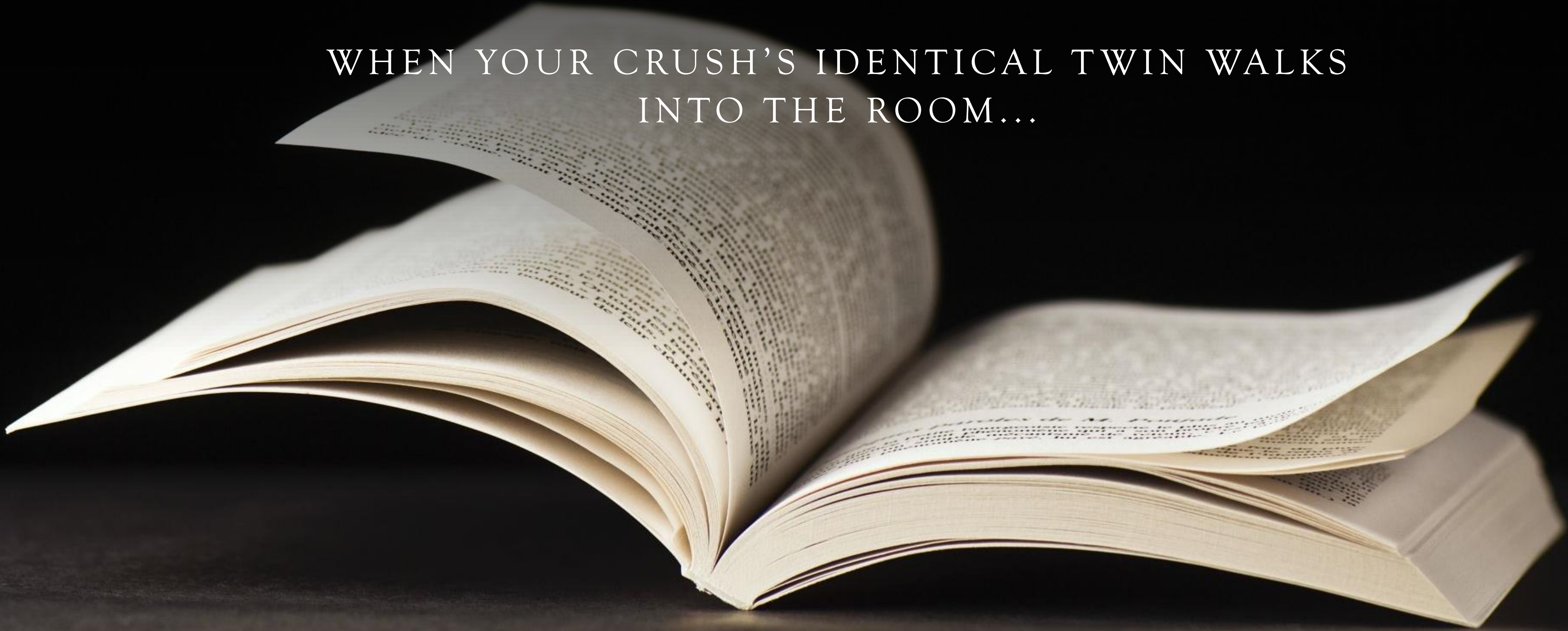
WHEN THINGS THAT HAPPEN IN VEGAS DON'T JUST STAY IN VEGAS…

CHAPTER 6: CHOOSING THE BEST MODEL

WHEN ALL THE CHARACTERS COME TOGETHER AND YOU GET TO PICK YOUR FAVORITE

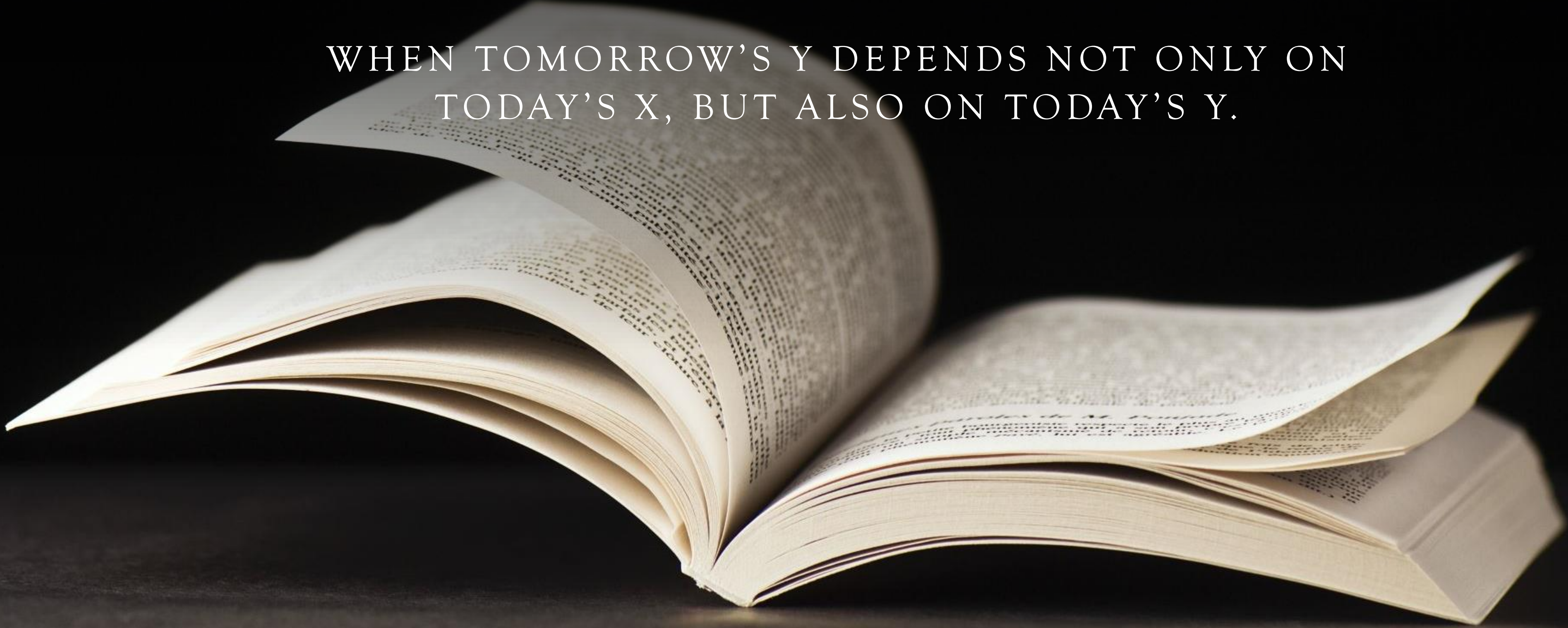
CHAPTER 7: INSTRUMENTAL VARIABLES

WHEN YOUR CRUSH'S IDENTICAL TWIN WALKS INTO THE ROOM…



CHAPTER 8: DYNAMIC PANELS

WHEN TOMORROW’S Y DEPENDS NOT ONLY ON TODAY’S X, BUT ALSO ON TODAY’S Y.

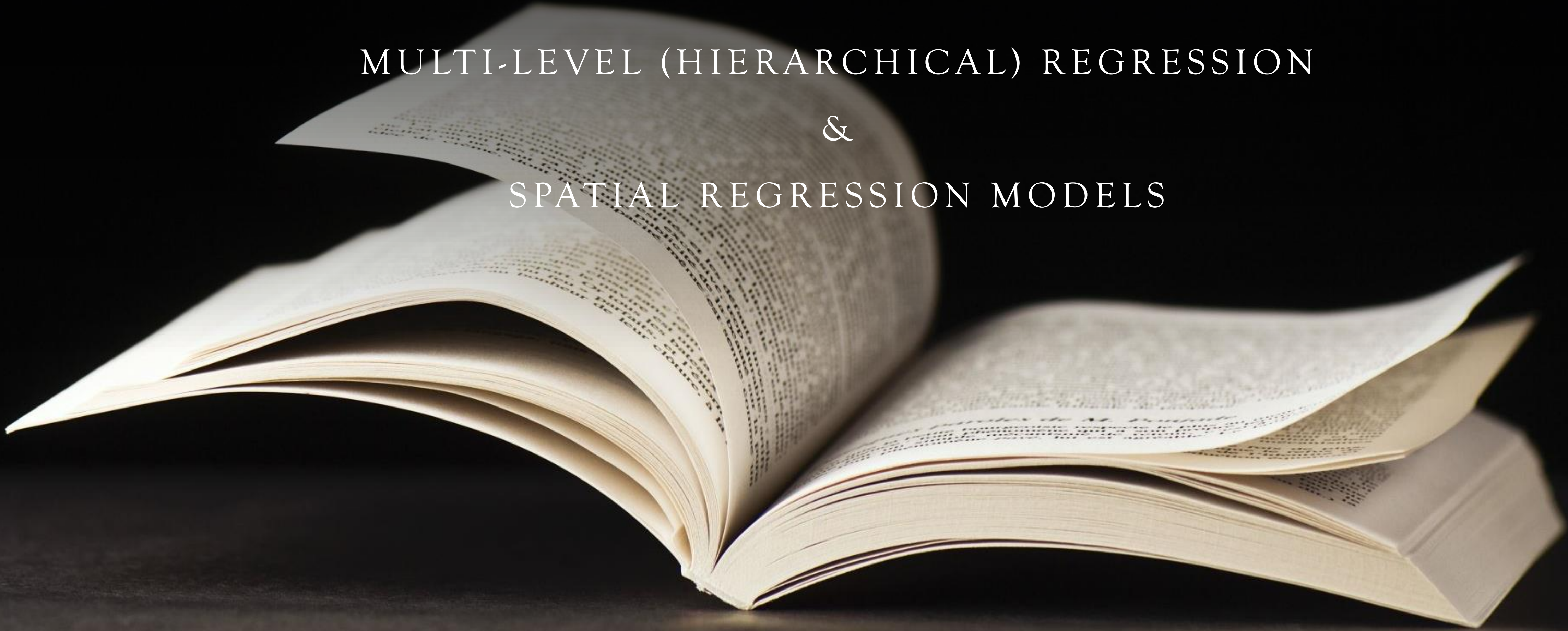


CHAPTER 9: GENERALIZED TIME MODELS

MULTI-LEVEL (HIERARCHICAL) REGRESSION

&

SPATIAL REGRESSION MODELS

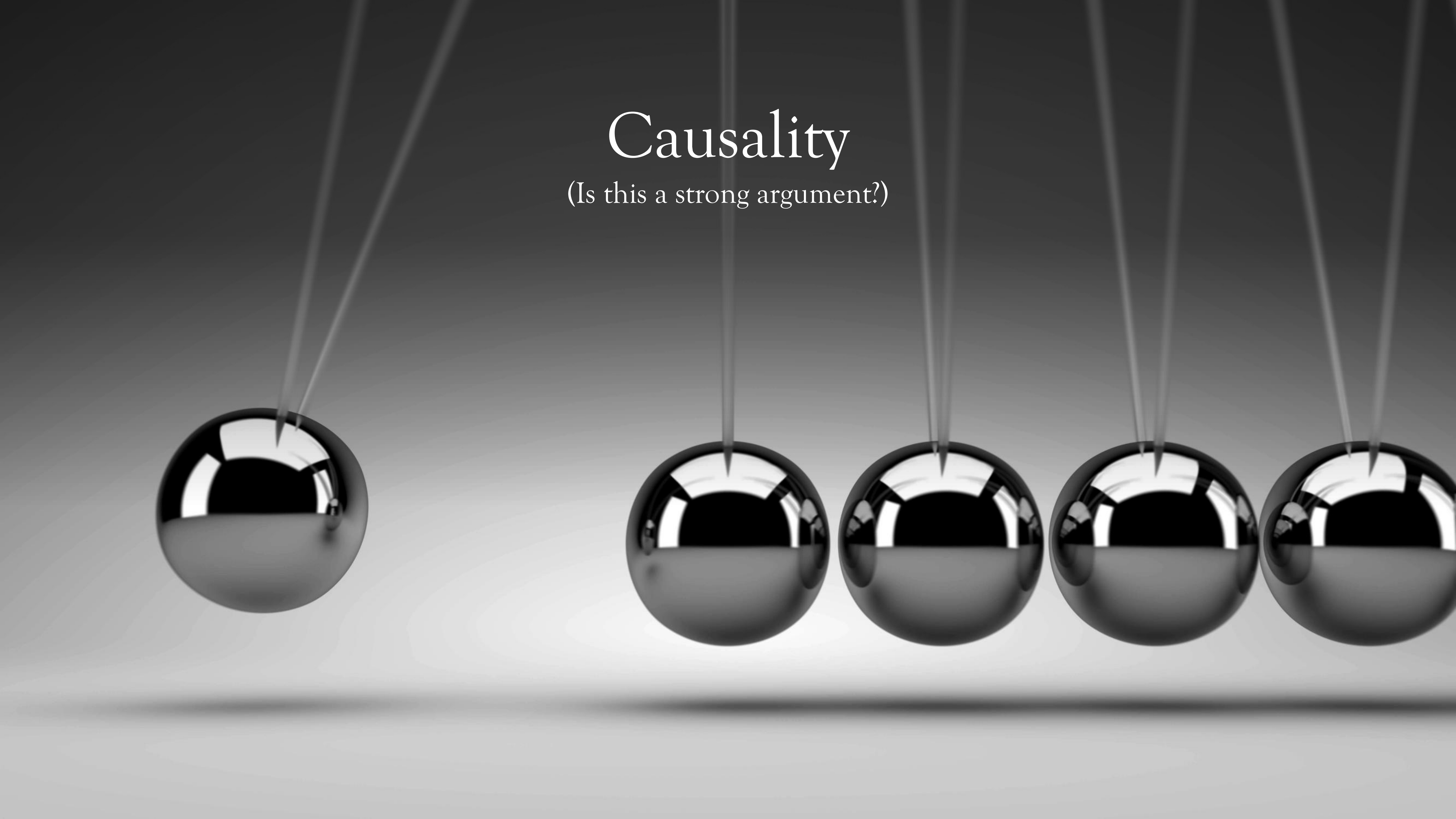


Some final comments before you go...

There’s another reason to use Panel Data…

Causality

(Is this a strong argument?)



Scattered comments about panel data

N entities overved over T time periods

- Typically $N \gg T$ (why?)
- Balanced vs. Imbalanced panel data
- Entities have characteristics that do not change through time (entity-fixed effects)
- Time periods may have characteristics that affect all entities equally (time-fixed effects)

Could you give examples of entity–fixed and time-fixed effects?

How would this change the "general" Panel Data model?

- Entity-fixed effects are much more common, and time-fixed effects are often neglected. Indeed, we often speak of "fixed effects" (with no qualification) to refer to "entity-fixed effects".