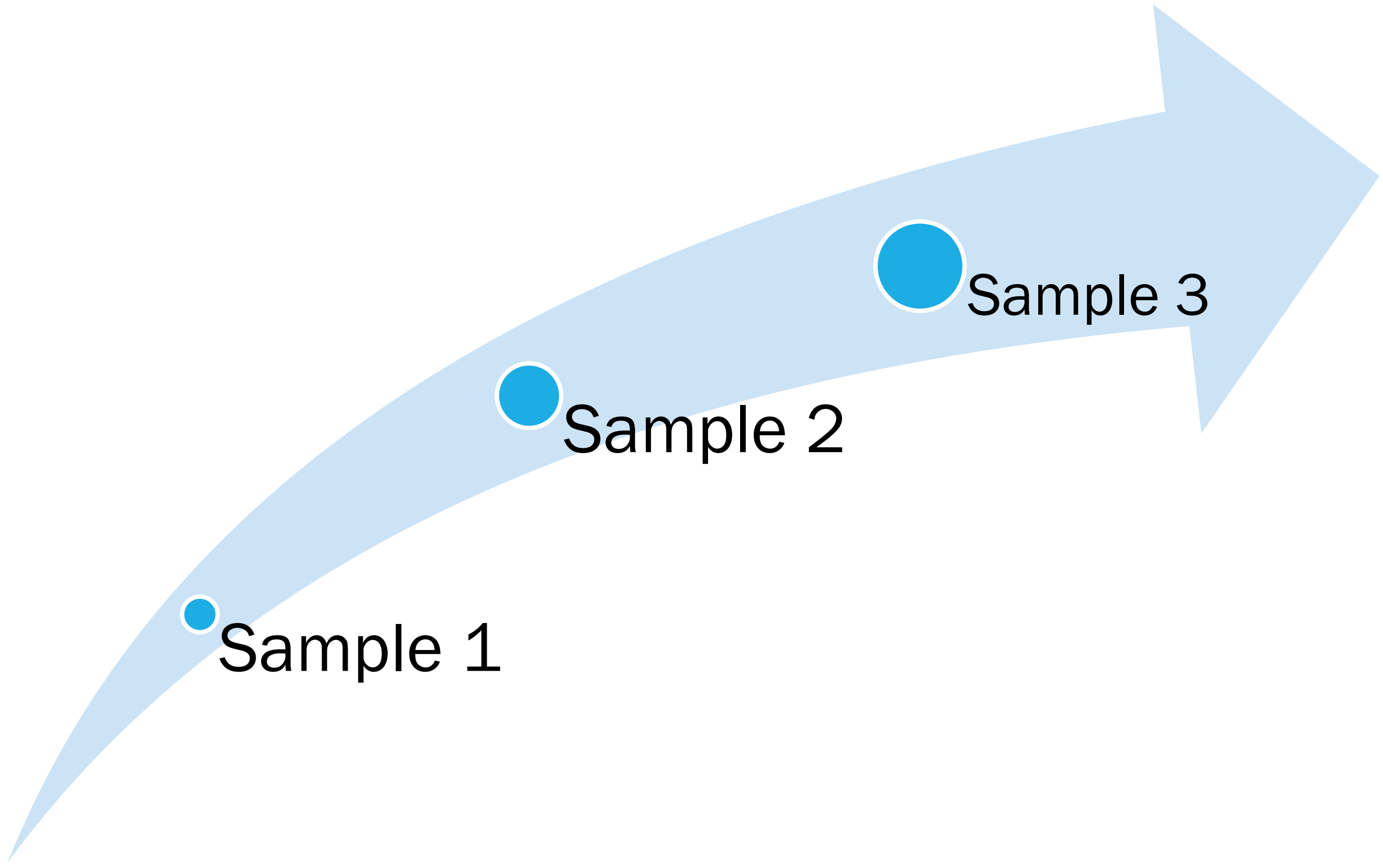
POOLED REGRESSION

FELIPE BUCHBINDER

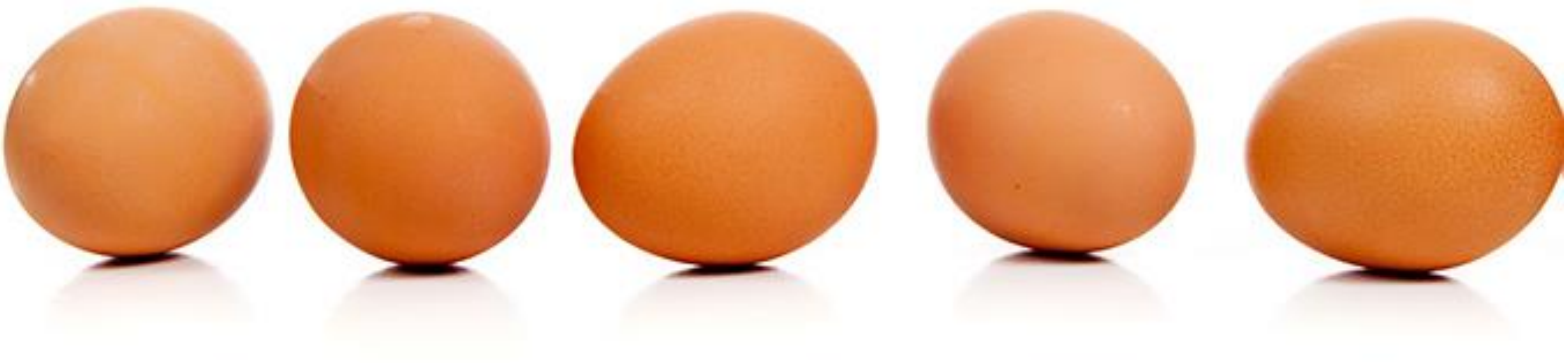
A POOLED REGRESSION TREATS PANEL DATA *AS IF IT WERE* POOLED DATA

SERIOUSLY, WHAT DOES THIS MEAN?

POOLED DATA :
A DIFFERENT SAMPLE OF ENTITIES AT EACH TIME PERIOD

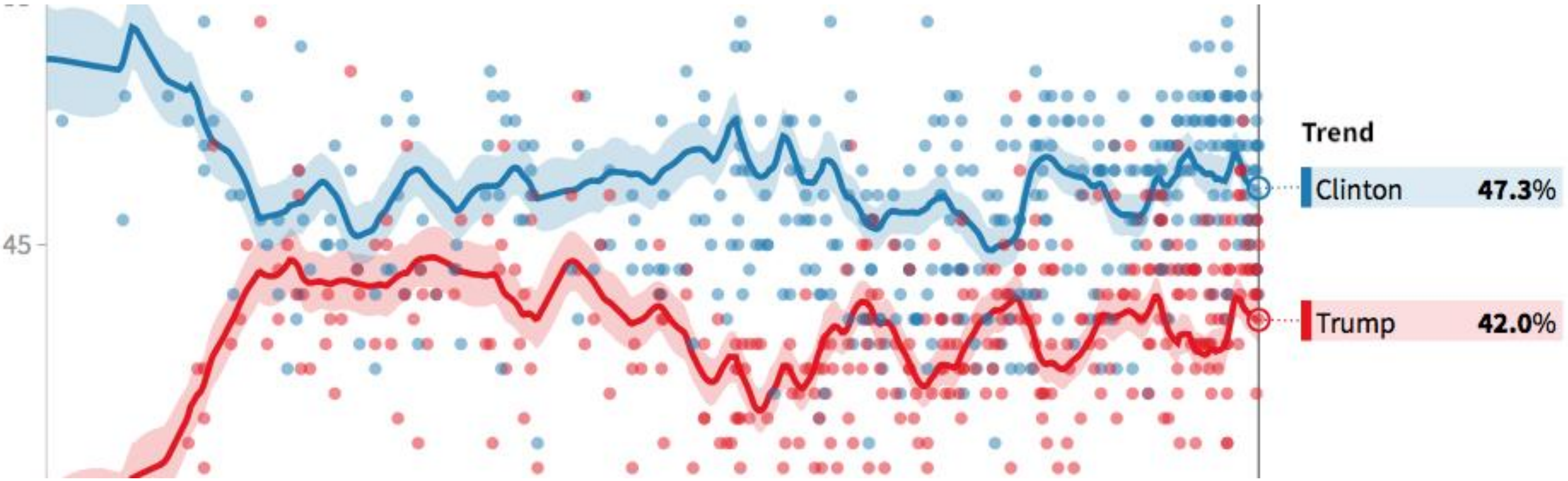


POOLED DATA IS LIKE A TIME SERIES OF CROSS-SECTIONS



ELECTION POLLS ARE AN EXAMPLE OF POOLED DATA

A DIFFERENT GROUP OF PEOPLE IS SAMPLED AT EACH TIME



IN POOLED REGRESSION, JUST FORGET YOU HAVE PANEL DATA AND RUN AN ORDINARY LINEAR REGRESSION INSTEAD

THIS MEANS WE STACK ALL OBSERVATIONS AND TREAT THEM AS INDEPENDENT OBSERVATIONS


AWESOME! CAN I DO THAT?

IGNORING A RELEVANT VARIABLE TYPICALLY LEADS TO BIASED ESTIMATES OF THE COEFFICIENTS...

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

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

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

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

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 $\boldsymbol{\beta}$

IGNORING A RELEVANT VARIABLE TYPICALLY LEADS TO BIASED ESTIMATES OF THE COEFFICIENTS…

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

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

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

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

Taking the expected value…

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

…unless $\mathbf{X}^{\mathbf{T}}\mathbf{U} = 0$,
In which case Pooled Regression yields unbiased estimates of the coefficients!

..

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

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

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

WHAT DOES $X^T U = 0$ MEAN IN PRACTICE?

IT MEANS THAT THE UNOBSERVED VARIABLES (U) ARE UNRELATED TO THE OBSERVED CHARACTERISTICS OF THE ENTITIES (X)

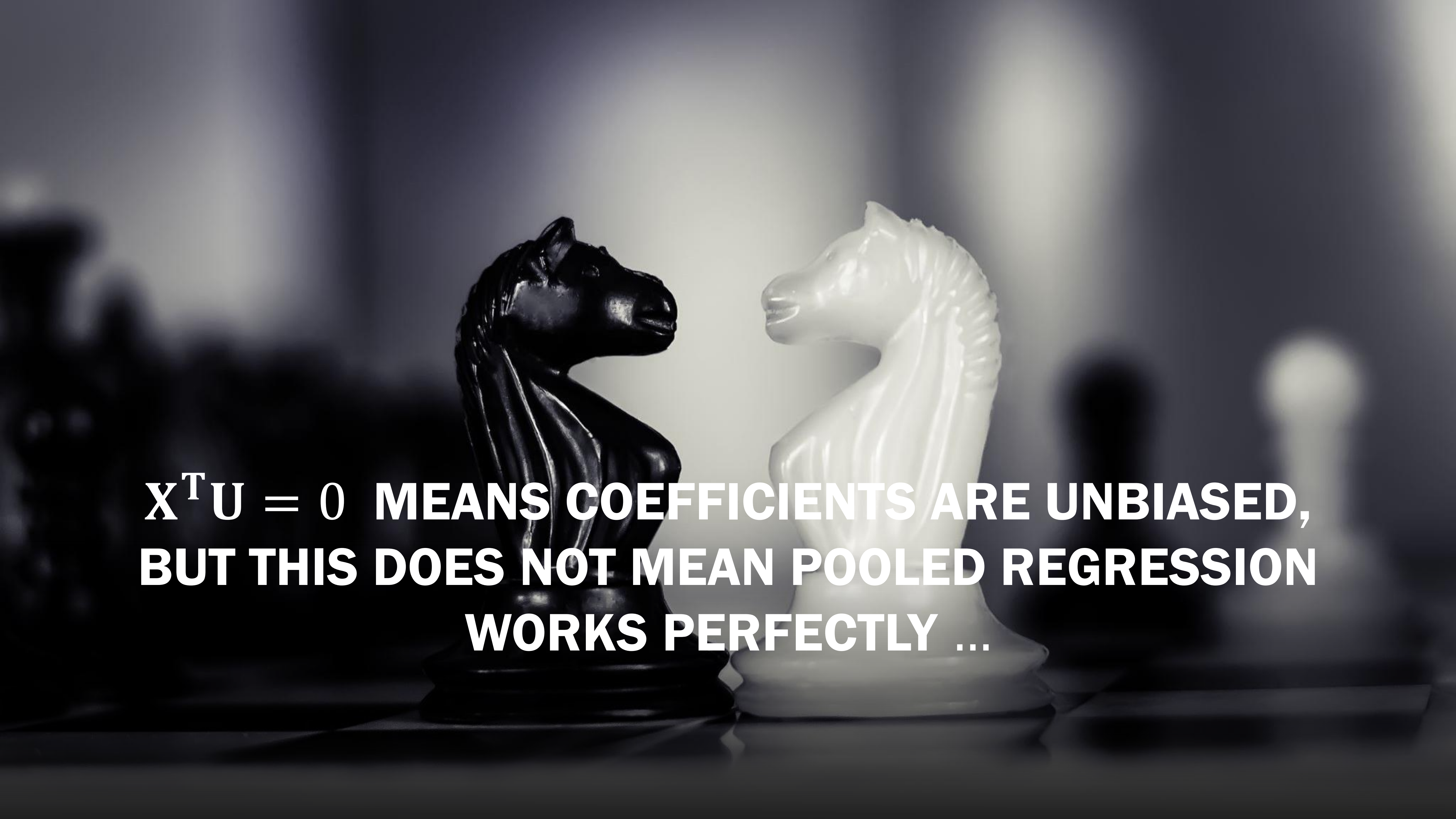
EXERCISE

You will be given some examples of regression. For each regression…

- Think of possible sources of unobserved heterogeneities (U).
- What theoretical argument would you need to make in order to convince a reviewer that pooled data is a valid approach to tackle this regression problem? (i.e. what would $X^T U = 0$ mean in this case?)

EXERCISE

1. Child Mortality ~ Democracy
2. Spending ~ Income
3. Wages ~ Education level
4. Crime Rate ~ Unemployment



$X^T U = 0$ MEANS COEFFICIENTS ARE UNBIASED, BUT THIS DOES NOT MEAN POOLED REGRESSION WORKS PERFECTLY …

$$\mathbf{Y}_{it} = \mathbf{X}_{it}\boldsymbol{\beta} + \underbrace{\mathbf{U}_{\mathbf{i}}}_{\text{Ignored in pooled regression}} + \epsilon_{it}$$

Pooled regression's residual

$$\mathbf{Y}_{it} = \mathbf{X}_{it}\boldsymbol{\beta} + \boxed{\mathbf{U}_{\mathrm{i}} + \epsilon_{it}}$$

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

$$\boldsymbol{\epsilon}_{it}^{\textbf{Pooled}} = \mathbf{U}_{\mathrm{i}} + \boldsymbol{\epsilon}_{it}$$

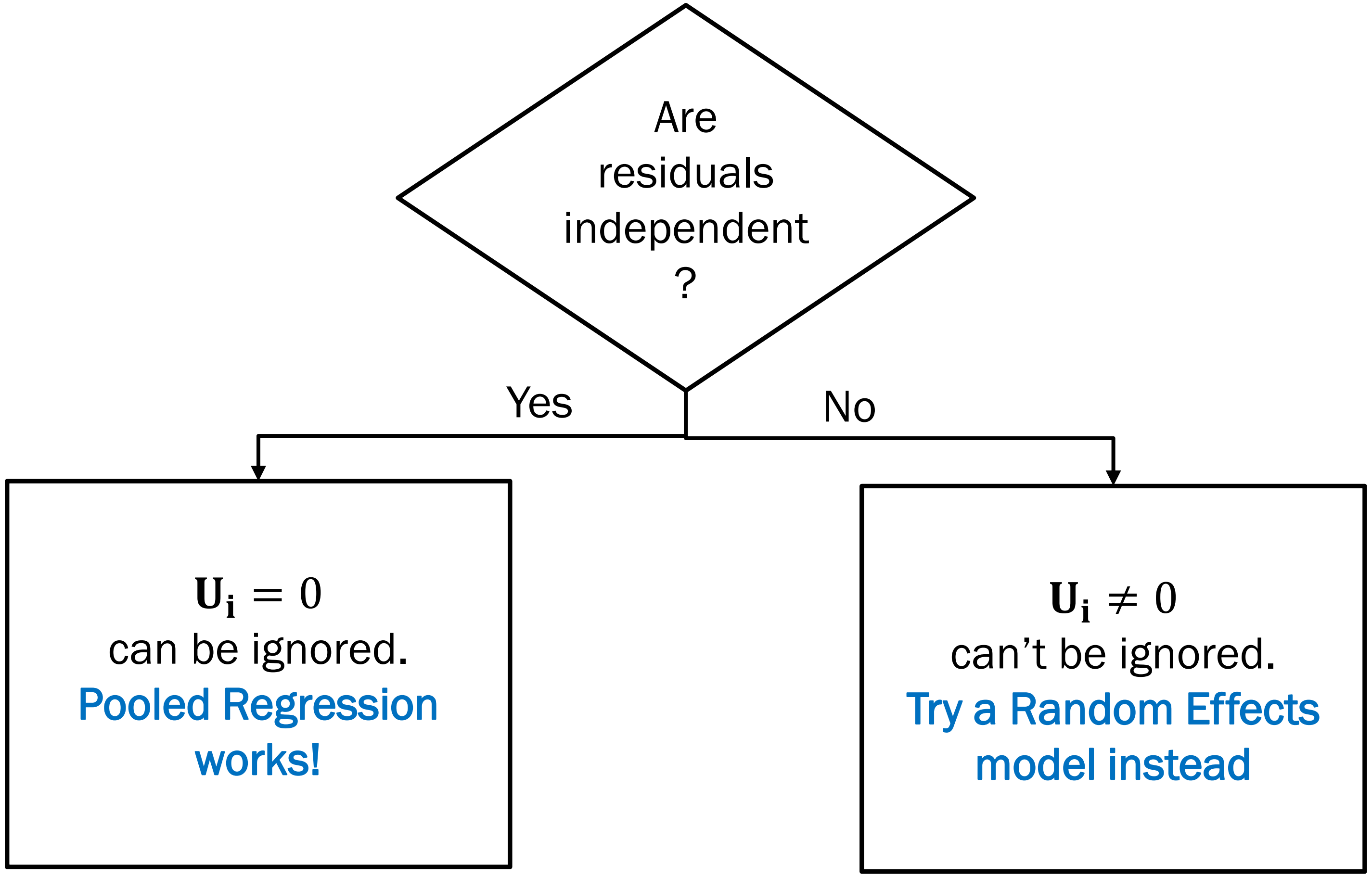
$$\mathrm{Y}_{it} = \mathrm{X}_{it}\beta + \mathrm{U_i} + \epsilon_{it}$$

$$\epsilon_{it}^{\mathbf{Pooled}} = \mathbf{U_i} + \epsilon_{it}$$

Residuals are serially correlated

OK, SO CAN I USE POOLED REGRESSION?

BREUSCH-PAGAN LAGRANGE MULTIPLIERS TEST



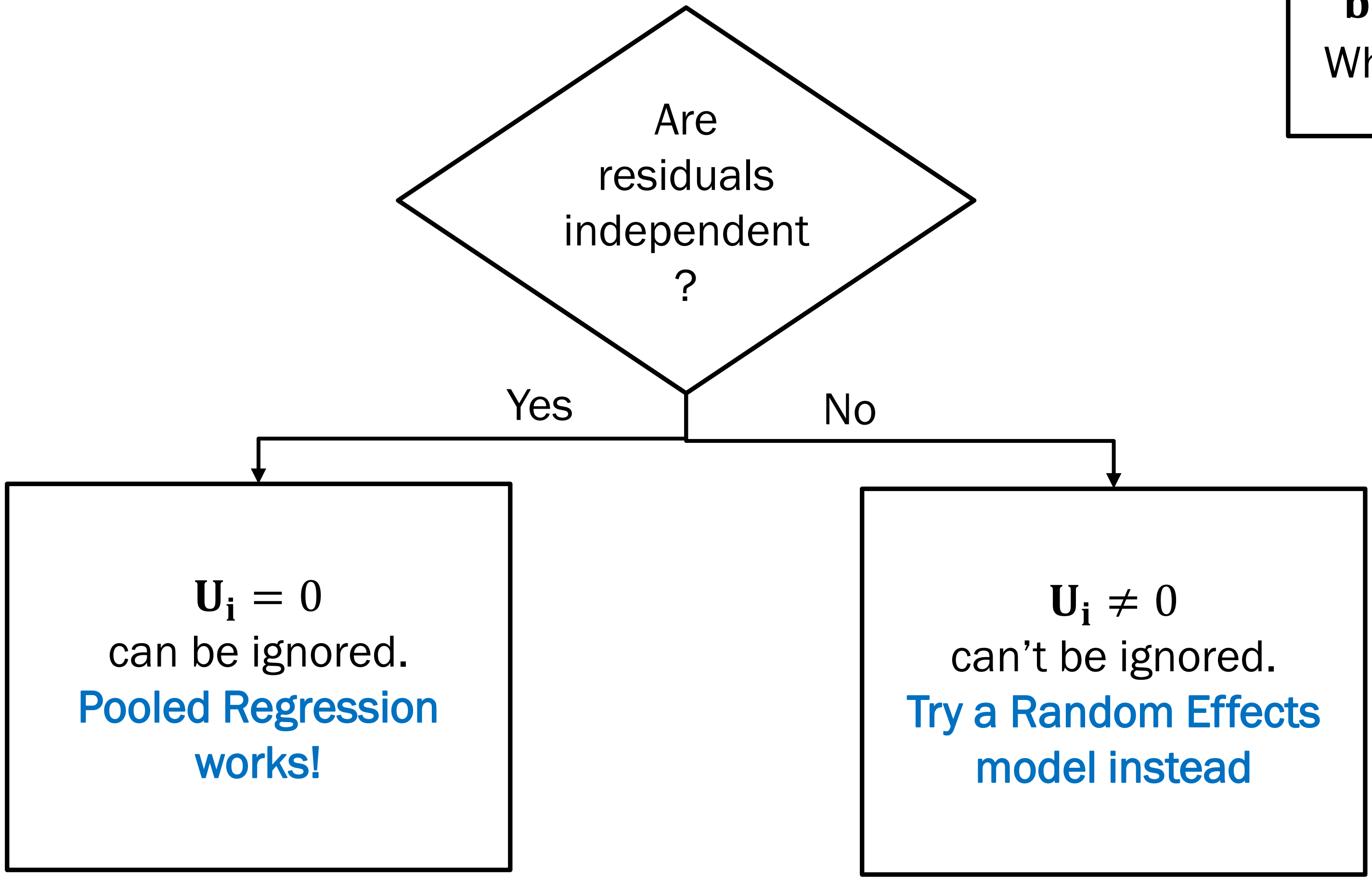
BREUSCH-PAGAN LAGRANGE MULTIPLIERS TEST

Recall that

$\boldsymbol{\epsilon}^{\mathbf{Pooled}} = \mathbf{U_i} + \boldsymbol{\epsilon}$

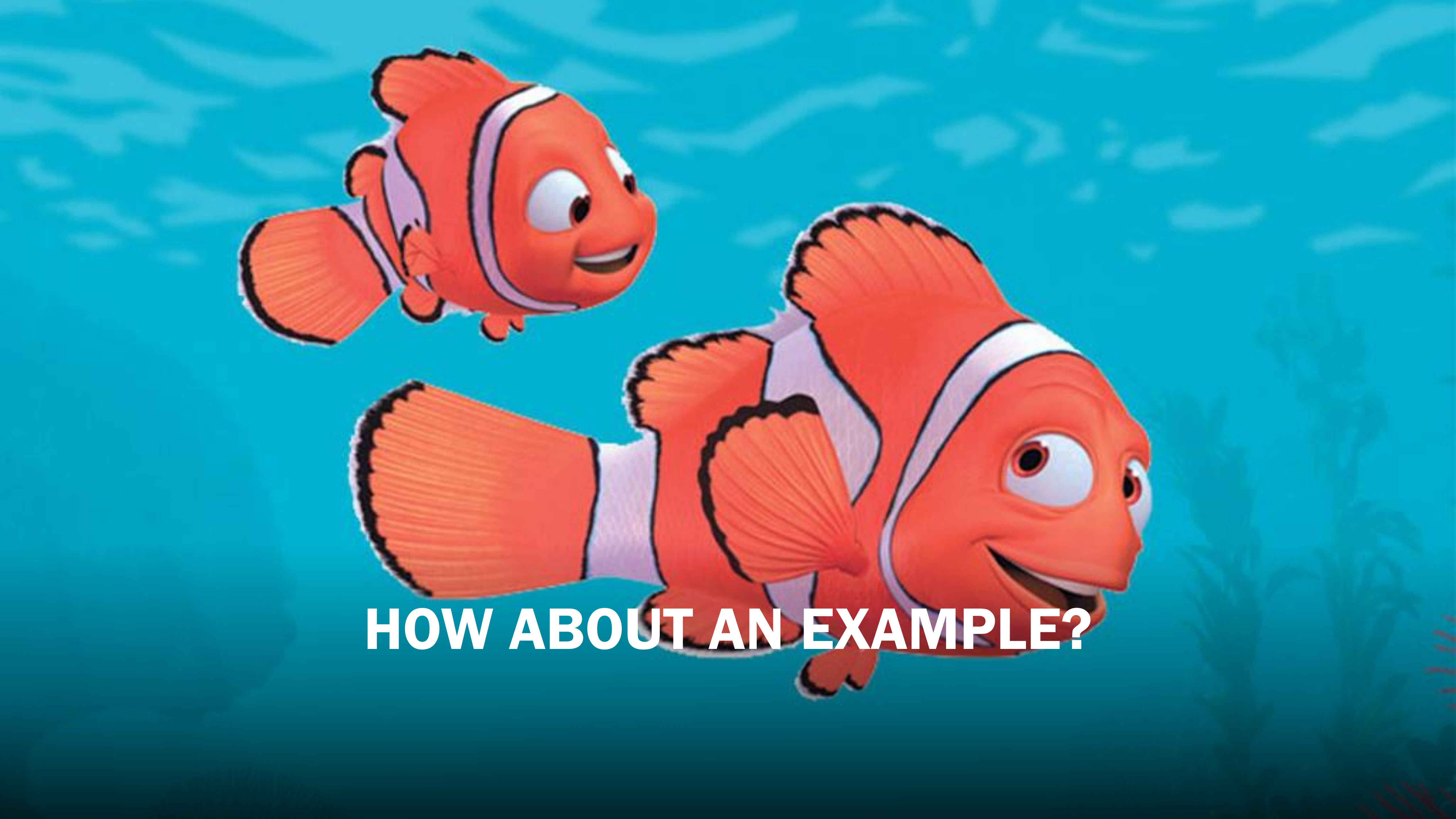
$\mathbf{b} = \boldsymbol{\beta} + \left(\mathbf{X^T X}\right)^{-\mathbf{1}} \mathbf{X^T U}$

What happens if $\mathbf{U} = 0$?



BREUSCH-PAGAN LAGRANGE MULTIPLIERS TEST

$$\begin{cases} H_0: & \mathbf{U} = 0 & \text{Use Pooled Regression} \\ H_a: & \mathbf{U} \neq 0 & \text{Try Random Effects} \end{cases}$$



HOW ABOUT AN EXAMPLE?

IS INVESTMENT DETERMINED BY COMPANY VALUE?
THE GRUNFELD DATASET

```
library("plm")

library("stargazer")

pooled <- plm(inv ~ value + capital, data=Grunfeld, model='pooling')
```

	Dependent variable:
	inv
	Pooled
value	0.116***
	(0.006)
capital	0.231***
	(0.025)
Constant	-42.714***
	(9.512)
Observations	200
R2	0.812
Adjusted R2	0.811
F Statistic	426.576*** (df = 2; 197)
Note:	*p<0.1; **p<0.05; ***p<0.01

WHAT IF WE TRY TO CAPTURE U_i BY USING A DUMMY FOR EACH i?

IS INVESTMENT DETERMINED BY COMPANY VALUE?

THE GRUNFELD DATASET

library("plm")

library("stargazer")

pooled <- plm(inv ~ value + capital, data=Grunfeld, model='pooling’)

ols <- lm(inv~value + capital + factor(firm, data=Grunfeld)

	Dependent variable:	
	inv	
	panel linear Pooled	OLS OLS with Dummies
	(1)	(2)
value	0.116*** (0.006)	0.110*** (0.012)
capital	0.231*** (0.025)	0.310*** (0.017)
factor(firm)2		172.203*** (31.161)
factor(firm)3		-165.275*** (31.776)
factor(firm)4		42.487 (43.910)
factor(firm)5		-44.320 (50.492)
factor(firm)6		47.135 (46.811)
factor(firm)7		3.743 (50.565)
factor(firm)8		12.751 (44.053)
factor(firm)9		-16.926 (48.453)
factor(firm)10		63.729 (50.330)
Constant	-42.714*** (9.512)	-70.297 (49.708)
Observations	200	200
R2	0.812	0.944
Adjusted R2	0.811	0.941
Residual Std. Error		52.768 (df = 188)
F Statistic	426.576*** (df = 2; 197)	288.500*** (df = 11; 188)
Note:		*p<0.1; **p<0.05; ***p<0.01

IS INVESTMENT DETERMINED BY COMPANY VALUE?

THE GRUNFELD DATASET

```
library("plm")

library("stargazer")

pooled <- plm(inv ~ value + capital, data=Grunfeld, model='pooling’)

ols <- lm(inv~value + capital + factor(firm, data=Grunfeld)
```

	Dependent variable:	
	inv	
	panel linear Pooled	OLS OLS with Dummies
	(1)	(2)
value	0.116***	0.110***
	(0.006)	(0.012)
capital	0.231***	0.310***
	(0.025)	(0.017)
factor(firm)2		172.203***
		(31.161)
factor(firm)3		-165.275***
		(31.776)
factor(firm)4		42.487
		(43.910)
factor(firm)5		-44.320
		(50.492)
factor(firm)6		47.135
		(46.811)
factor(firm)7		3.743
		(50.565)
factor(firm)8		12.751
		(44.053)
factor(firm)9		-16.926
		(48.453)
factor(firm)10		63.729
		(50.330)
Constant	-42.714***	-70.297
	(9.512)	(49.708)
Observations	200	200
R2	0.812	0.944
Adjusted R2	0.811	0.941
Residual Std. Error		52.768 (df = 188)
F Statistic	426.576*** (df = 2; 197)	288.500*** (df = 11; 188)
Note:		*p<0.1; **p<0.05; ***p<0.01

Significant dummies! The U_is cannot be ignored. Pooled regression is not adequate here.

IS INVESTMENT DETERMINED BY COMPANY VALUE?

THE GRUNFELD DATASET

library("plm")

library("stargazer")

pooled <- plm(inv ~ value + capital, data=Grunfeld, model='pooling’)

ols <- lm(inv~value + capital + factor(firm, data=Grunfeld)

	Dependent variable:	
	inv	
	panel linear Pooled	OLS OLS with Dummies
	(1)	(2)
value	0.116***	0.110***
	(0.006)	(0.012)
capital	0.231***	0.310***
	(0.025)	(0.017)
factor(firm)2		172.203***
		(31.161)
factor(firm)3		-165.275***
		(31.776)
factor(firm)4		42.487
		(43.910)
factor(firm)5		-44.320
		(50.492)
factor(firm)6		47.135
		(46.811)
factor(firm)7		3.743
		(50.565)
factor(firm)8		12.751
		(44.053)
factor(firm)9		-16.926
		(48.453)
factor(firm)10		63.729
		(50.330)
Constant	-42.714***	-70.297
	(9.512)	(49.708)
Observations	200	200
R2	0.812	0.944
Adjusted R2	0.811	0.941
Residual Std. Error		52.768 (df = 188)
F Statistic	426.576*** (df = 2; 197)	288.500*** (df = 11; 188)
Note:		*p<0.1; **p<0.05; ***p<0.01

Coefficients changed. Pooled Regression estimates were biased!

IS INVESTMENT DETERMINED BY COMPANY VALUE?

THE GRUNFELD DATASET

library("plm")

library("stargazer")

pooled <- plm(inv ~ value + capital, data=Grunfeld, model='pooling’)

ols <- lm(inv~value + capital + factor(firm, data=Grunfeld)

	Dependent variable:	
	inv	
	panel linear Pooled	OLS OLS with Dummies
	(1)	(2)
value	0.116***	0.110***
	(0.006)	(0.012)
capital	0.231***	0.310***
	(0.025)	(0.017)
factor(firm)2		172.203***
		(31.161)
factor(firm)3		-165.275***
		(31.776)
factor(firm)4		42.487
		(43.910)
factor(firm)5		-44.320
		(50.492)
factor(firm)6		47.135
		(46.811)
factor(firm)7		3.743
		(50.565)
factor(firm)8		12.751
		(44.053)
factor(		-16.926
		(48.453)
factor(		63.729
		(50.330)
Constan		-70.297
		(49.708)
Observations	200	200
R2	0.812	0.944
Adjusted R2	0.811	0.941
Residual Std. Error		52.768 (df = 188)
F Statistic	426.576*** (df = 2; 197)	288.500*** (df = 11; 188)
Note:		*p<0.1; **p<0.05; ***p<0.01

Why did the R^2 increase?

IS INVESTMENT DETERMINED BY COMPANY VALUE?

THE GRUNFELD DATASET

```
library("plm")

library("stargazer")

pooled <- plm(inv ~ value + capital, data=Grunfeld, model='pooling’)

ols <- lm(inv~value + capital + factor(firm), data=Grunfeld)
```

	Dependent variable:	
	inv	
	panel linear Pooled	OLS OLS with Dummies
	(1)	(2)
value	0.116***	0.110***
	(0.006)	(0.012)
capital	0.231***	0.310***
	(0.025)	(0.017)
factor(firm)2		172.203***
		(31.161)
factor(firm)3		-165.275***
		(31.776)
factor(firm)4		42.487
		(43.910)
factor(firm)5		-44.320
		(50.492)
factor(firm)6		47.135
		(46.811)
factor(firm)7		3.743
		(50.565)
factor(firm)8		12.751
		(44.053)
factor(firm)9		-16.926
		(48.453)
factor(firm)10		63.729
		(50.330)
Constant	-42.714***	-70.297
	(9.512)	(49.708)
Observations	200	200
R2	0.812	0.944
Adjusted R2	0.811	0.941
Residual Std. Error		52.768 (df = 188)
F Statistic	426.576*** (df = 2; 197)	288.500*** (df = 11; 188)
Note:		*p<0.1; **p<0.05; ***p<0.01

I’m not interested in all these dummy coefficients. They only make my model look more complex! (continues....)

IS INVESTMENT DETERMINED BY COMPANY VALUE?

THE GRUNFELD DATASET

library("plm")

library("stargazer")

pooled <- plm(inv ~ value + capital, data=Grunfeld, model='pooling’)

ols <- lm(inv~value + capital + factor(firm, data=Grunfeld)

	Dependent variable:	
	inv	
	panel linear Pooled	OLS OLS with Dummies
	(1)	(2)
value	0.116***	0.110***
	(0.006)	(0.012)
capital	0.231***	0.310***
	(0.025)	(0.017)
factor(firm)2		172.203***
		(31.161)
factor(firm)3		-165.275***
		(31.776)
factor(firm)4		42.487
		(43.910)
factor(firm)5		-44.320
		(50.492)
factor(firm)6		47.135
		(46.811)
factor(firm)7		3.743
		(50.565)
factor(firm)8		12.751
		(44.053)
factor(firm)9		-16.926
		(48.453)
factor(firm)10		63.729
		(50.330)
Constant	-42.714***	-70.297
	(9.512)	(49.708)
Observations	200	200
R2	0.812	0.944
Adjusted R2	0.811	0.941
Residual Std. Error		52.768 (df = 188)
F Statistic	426.576*** (df = 2; 197)	288.500*** (df = 11; 188)
Note:		*p<0.1; **p<0.05; ***p<0.01

Can’t I have a model that looks like this, simple, only with the coefficients I’m interested in…. But unbiased???


YES, OF COURSE!


IT'S CALLED FIXED EFFECTS


STAY TUNED...