Inference with predicted data (IPD)

Tyler H McCormick
University of Washington
PAA 2025

The New York Times

Three-Quarters of U.S. Adults Are Now Overweight or Obese

A sweeping new paper reveals the dramatic rise of obesity rates nationwide since 1990.

By Nina Agrawal

Nov. 14, 2024

THE LANCET

ARTICLES · Online first, November 14, 2024 · *Open Access*

National-level and state-level prevalence of overweight and obesity among children, adolescents, and adults in the USA, 1990–2021, and forecasts up to 2050

GBD 2021 US Obesity Forecasting Collaborators †

Affiliations & Notes Article Info Linked Articles (1)

Download PDF Cite Share Set Alert Get Rights Reprints

A Older adolescents (aged 15–24 years)

	Females					Males				
	1990	2021	2050	Percentage change 1990–2021	Percentage change 2022–50	1990	2021	2050	Percentage change 1990–2021	Percentage change 2022–50
USA	10·1% (9·0 to 11·3)	28·8% (25·4 to 32·5)	38·0% (30·8 to 44·2)	185·9% (139·4 to 237·1)	32·0% (15·2 to 45·9)	8·8% (8·0 to 9·6)	22·7% (20·3 to 25·1)	30·6% (24·9 to 35·3)	158·4% (123·9 to 197·4)	35·0% (17·3 to 47·8)
Alabama	10·4% (8·2 to 12·8)	38·0% (32·4 to 43·9)	**52·2% (41·6 to 60·3)**	**271·2% (181·5 to 382·9)**	37·7% (18·3 to 53·3)	7·3% (5·4 to 9·3)	25·8% (21·3 to 31·1)	36·4% (27·1 to 44·1)	**262·2% (153·0 to 401·6)**	41·4% (18·5 to 65·6)
Alaska	15·3% (11·3 to 19·8)	31·5% (25·4 to 38·1)	40·2% (28·6 to 49·6)	110·6% (46·9 to 191·9)	28·2% (5·2 to 46·7)	8·8% (6·1 to 12·4)	24·9% (19·9 to 30·3)	35·5% (23·7 to 45·4)	191·7% (85·0 to 331·0)	43·1% (8·9 to 73·1)
Arizona	8·0% (5·7 to 10·5)	29·2% (24·1 to 35·0)	41·4% (28·9 to 52·9)	**274·3% (159·9 to 424·7)**	42·1% (11·4 to 71·6)	8·4% (6·4 to 10·6)	23·2% (19·4 to 27·6)	32·6% (23·9 to 40·4)	180·2% (105·5 to 276·6)	41·3% (14·3 to 61·4)
Arkansas	11·3% (8·4 to 15·0)	39·2% (33·1 to 45·8)	**52·9% (41·1 to 63·8)**	**254·3% (152·2 to 402·9)**	36·1% (15·6 to 62·5)	11·6% (8·4 to 15·4)	25·6% (20·9 to 31·4)	32·7% (25·0 to 40·4)	126·6% (55·3 to 218·9)	27·8% (10·9 to 41·7)
California	7·8% (6·3 to 9·5)	24·0% (19·8 to 28·6)	30·5% (24·0 to 39·7)	**209·7% (131·1 to 303·0)**	28·2% (6·1 to 57·4)	8·8% (7·1 to 10·6)	21·5% (17·9 to 25·6)	27·6% (21·7 to 34·7)	147·1% (86·5 to 214·5)	28·8% (6·5 to 54·7)
Colorado	9·1% (6·5 to 12·2)	22·0% (17·9 to 26·8)	28·8% (20·5 to 36·0)	147·8% (68·6 to 253·5)	31·7% (5·7 to 46·8)	5·9% (4·1 to 8·0)	16·3% (13·1 to 19·6)	21·1% (14·8 to 27·2)	183·4% (91·3 to 314·1)	30·0% (4·7 to 53·0)
Connecticut	7·4% (5·2 to 10·1)	24·6% (19·8 to 30·1)	31·8% (23·6 to 42·1)	**241·4% (131·2 to 394·9)**	29·7% (5·2 to 59·6)	8·5% (6·6 to 10·7)	20·9% (17·0 to 25·0)	27·4% (20·7 to 34·4)	150·0% (82·2 to 235·6)	31·8% (9·2 to 59·4)
Delaware	12·4% (9·5 to 15·6)	31·7% (25·9 to 37·6)	42·3% (30·6 to 52·2)	159·9% (87·2 to 253·3)	34·2% (12·0 to 53·5)	10·8% (7·9 to 14·4)	20·3% (15·9 to 25·1)	27·8% (20·3 to 35·0)	93·0% (29·6 to 174·1)	37·8% (16·1 to 58·7)
Florida	10·7% (8·0 to 13·7)	26·7% (21·2 to 32·2)	33·9% (25·1 to 41·6)	154·3% (80·9 to 251·7)	28·1% (9·0 to 43·8)	7·5% (5·9 to 9·3)	19·4% (15·7 to 23·2)	24·8% (18·4 to 31·4)	161·8% (89·1 to 255·7)	27·8% (10·3 to 51·3)
Georgia	11·5% (9·3 to 14·3)	30·0% (25·1 to 35·5)	38·3% (29·1 to 47·7)	164·4% (98·3 to 242·3)	28·1% (9·9 to 51·8)	7·9% (6·0 to 10·3)	22·5% (18·8 to 26·6)	30·9% (22·5 to 38·7)	187·9% (106·3 to 289·9)	37·8% (9·5 to 63·0)
Hawaii	9·9% (7·6 to 12·7)	26·0% (20·9 to 31·5)	33·1% (26·3 to 40·7)	166·2% (91·1 to 261·7)	28·7% (13·8 to 47·5)	11·4% (9·0 to 14·3)	25·0% (20·8 to 29·8)	33·0% (27·1 to 40·5)	121·8% (61·4 to 191·0)	32·8% (22·9 to 51·4)
Idaho	10·4% (8·2 to 13·0)	28·7% (23·5 to 34·3)	39·1% (27·3 to 49·2)	179·9% (102·7 to 273·6)	36·8% (8·1 to 64·8)	6·7% (5·1 to 8·4)	20·1% (16·6 to 24·1)	29·9% (20·5 to 38·1)	**205·3% (118·4 to 317·5)**	49·1% (13·8 to 79·0)
Illinois	9·9% (7·9 to 12·3)	27·7% (23·2 to 33·2)	35·0% (26·8 to 43·3)	182·9% (110·4 to 272·3)	26·9% (8·3 to 50·3)	8·8% (7·0 to 10·7)	22·1% (18·0 to 26·0)	31·4% (21·7 to 39·4)	154·9% (92·3 to 233·8)	42·7% (11·4 to 65·6)
Indiana	11·2% (9·0 to 14·0)	35·4% (29·9 to 40·7)	**48·3% (37·6 to 57·9)**	**219·2% (142·4 to 311·8)**	37·4% (17·4 to 64·1)	9·6% (7·7 to 11·9)	26·2% (21·6 to 30·8)	37·3% (28·4 to 45·6)	176·3% (102·6 to 261·4)	42·8% (19·6 to 63·6)
Iowa	10·8% (8·4 to 13·6)	34·1% (29·1 to 39·3)	**48·6% (36·6 to 58·1)**	**221·4% (143·2 to 325·0)**	43·2% (16·7 to 65·4)	11·0% (8·7 to 13·7)	23·4% (19·6 to 27·4)	32·2% (22·8 to 40·1)	115·8% (64·3 to 182·6)	37·5% (10·7 to 62·3)
Kansas	13·6% (10·1 to 17·6)	31·8% (26·8 to 37·3)	39·8% (30·9 to 46·5)	138·9% (73·4 to 223·3)	25·3% (7·5 to 38·1)	8·2% (5·8 to 11·0)	25·5% (21·8 to 29·4)	36·0% (26·7 to 44·2)	**218·0% (123·4 to 349·7)**	41·6% (13·4 to 64·4)
Kentucky	12·4% (10·1 to 14·8)	36·4% (30·9 to 42·3)	**49·6% (39·0 to 58·7)**	197·1% (128·5 to 275·7)	37·2% (18·0 to 53·4)	10·1% (8·3 to 12·2)	25·7% (21·2 to 30·1)	34·9% (27·1 to 42·0)	157·1% (93·0 to 231·4)	36·8% (16·8 to 54·0)
Louisiana	14·7% (11·2 to 18·4)	34·8% (28·9 to 39·9)	43·2% (34·6 to 50·6)	140·9% (77·1 to 221·0)	24·7% (12·5 to 35·6)	8·9% (6·6 to 11·6)	26·4% (22·1 to 31·5)	36·1% (27·1 to 45·1)	**202·7% (116·0 to 320·8)**	37·2% (14·8 to 64·0)
Maine	13·0% (9·6 to 16·9)	27·8% (22·3 to 33·4)	35·3% (27·4 to 44·0)	118·1% (51·9 to 197·5)	27·7% (12·8 to 41·6)	9·1% (6·7 to 11·8)	22·9% (18·9 to 27·8)	31·8% (22·8 to 39·7)	157·0% (81·8 to 255·2)	38·7% (14·8 to 65·4)
Maryland	11·0% (8·6 to 13·6)	27·7% (23·1 to 32·7)	34·8% (26·0 to 42·2)	156·0% (86·3 to 237·6)	26·9% (7·0 to 41·3)	10·5% (8·3 to 13·1)	21·7% (18·2 to 26·0)	28·8% (20·8 to 36·0)	110·7% (56·2 to 179·2)	32·6% (7·5 to 55·0)
Massachusetts	7·4% (5·5 to 9·9)	22·3% (18·1 to 27·0)	30·3% (21·2 to 40·0)	209·6% (115·1 to 332·0)	36·5% (9·6 to 64·6)	8·2% (6·2 to 10·5)	17·9% (14·4 to 21·6)	22·2% (16·6 to 28·3)	123·3% (58·4 to 204·2)	24·2% (6·0 to 43·0)
Michigan	11·0% (8·8 to 13·3)	31·9% (26·8 to 37·1)	41·9% (32·9 to 49·5)	192·2% (123·3 to 275·0)	31·9% (13·1 to 47·3)	10·6% (8·4 to 13·0)	23·0% (19·3 to 26·9)	30·3% (23·1 to 37·7)	119·3% (65·8 to 182·0)	31·5% (11·1 to 54·3)
Minnesota	7·1% (5·7 to 8·8)	26·3% (21·9 to 31·0)	38·3% (27·6 to 46·6)	**276·7% (179·4 to 386·3)**	45·7% (14·3 to 69·4)	7·6% (6·0 to 9·3)	22·9% (19·3 to 27·2)	33·1% (23·3 to 41·1)	**207·2% (129·1 to 301·3)**	45·3% (13·4 to 69·1)
Mississippi	13·7% (10·5 to 17·0)	40·9% (35·1 to 46·5)	**53·5% (43·6 to 62·1)**	203·8% (128·0 to 296·7)	31·7% (13·8 to 46·8)	10·2% (7·7 to 13·1)	28·5% (24·0 to 33·8)	39·8% (30·6 to 47·8)	184·3% (108·5 to 284·0)	40·0% (18·1 to 58·2)
Missouri	10·2% (7·9 to 12·9)	32·5% (27·4 to 38·1)	42·5% (33·2 to 50·5)	**222·7% (137·7 to 333·3)**	31·4% (12·3 to 47·7)	9·2% (7·2 to 11·4)	25·2% (21·3 to 29·2)	35·9% (27·8 to 43·4)	178·4% (107·3 to 263·6)	42·7% (19·2 to 60·5)
Montana	9·1% (6·7 to 12·0)	26·3% (21·5 to 31·3)	38·5% (26·3 to 48·7)	194·5% (102·8 to 312·6)	46·9% (12·2 to 80·8)	9·1% (6·8 to 12·0)	20·6% (17·0 to 24·9)	28·7% (19·7 to 37·2)	131·6% (60·7 to 222·7)	39·8% (9·7 to 69·3)

A Older adolescents (aged 15–24 years)

	Females					Males				
	1990	2021	2050	Percentage change 1990–2021	Percentage change 2022–50	1990	2021	2050	Percentage change 1990–2021	Percentage change 2022–50
USA	10·1% (9·0 to 11·3)	28·8% (25·4 to 32·5)	38·0% (30·8 to 44·2)	185·9% (139·4 to 237·1)	32·0% (15·2 to 45·9)	8·8% (8·0 to 9·6)	22·7% (20·3 to 25·1)	30·6% (24·9 to 35·3)	158·4% (123·9 to 197·4)	35·0% (17·3 to 47·8)
Alabama	10·4% (8·2 to 12·8)	38·0% (32·4 to 43·9)	**52·2% (41·6 to 60·3)**	**271·2% (181·5 to 382·9)**	37·7% (18·3 to 53·3)	7·3% (5·4 to 9·3)	25·8% (21·3 to 31·1)	36·4% (27·1 to 44·1)	**262·2% (153·0 to 401·6)**	41·4% (18·5 to 65·6)
Alaska	15·3% (11·3 to 19·8)	31·5% (25·4 to 38·1)	40·2% (28·6 to 49·6)	110·6% (46·9 to 191·9)	28·2% (5·2 to 46·7)	8·8% (6·1 to 12·4)	24·9% (19·9 to 30·3)	35·5% (23·7 to 45·4)	191·7% (85·0 to 331·0)	43·1% (8·9 to 73·1)
Arizona	8·0% (5·7 to 10·5)	29·2% (24·1 to 35·0)	41·4% (28·9 to 52·9)	**274·3% (159·9 to 424·7)**	42·1% (11·4 to 71·6)	8·4% (6·4 to 10·6)	23·2% (19·4 to 27·6)	32·6% (23·9 to 40·4)	180·2% (105·5 to 276·6)	41·3% (14·3 to 61·4)
Arkansas	11·3% (8·4 to 15·0)	39·2% (33·1 to 45·8)	**52·9% (41·1 to 63·8)**	**254·3% (152·2 to 402·9)**	36·1% (15·6 to 62·5)	11·6% (8·4 to 15·4)	25·6% (20·9 to 31·4)	32·7% (25·0 to 40·4)	126·6% (55·3 to 218·9)	27·8% (10·9 to 41·7)
California	7·8% (6·3 to 9·5)	24·0% (19·8 to 28·6)	30·5% (24·0 to 39·7)	**209·7% (131·1 to 303·0)**	28·2% (6·1 to 57·4)	8·8% (7·1 to 10·6)	21·5% (17·9 to 25·6)	27·6% (21·7 to 34·7)	147·1% (86·5 to 214·5)	28·8% (6·5 to 54·7)
Colorado	9·1% (6·5 to 12·2)	22·0% (17·9 to 26·8)	28·8% (20·5 to 36·0)	147·8% (68·6 to 253·5)	31·7% (5·7 to 46·8)	5·9% (4·1 to 8·0)	16·3% (13·1 to 19·6)	21·1% (14·8 to 27·2)	183·4% (91·3 to 314·1)	30·0% (4·7 to 53·0)
Connecticut	7·4% (5·2 to 10·1)	24·6% (19·8 to 30·1)	31·8% (23·6 to 42·1)	**241·4% (131·2 to 394·9)**	29·7% (5·2 to 59·6)	8·5% (6·6 to 10·7)	20·9% (17·0 to 25·0)	27·4% (20·7 to 34·4)	150·0% (82·2 to 235·6)	31·8% (9·2 to 59·4)
Delaware	12·4% (9·5 to 15·6)	31·7% (25·9 to 37·6)	42·3% (30·6 to 52·2)	159·9% (87·2 to 253·3)	34·2% (12·0 to 53·5)	10·8% (7·9 to 14·4)	20·3% (15·9 to 25·1)	27·8% (20·3 to 35·0)	93·0% (29·6 to 174·1)	37·8% (16·1 to 58·7)
Florida	10·7% (8·0 to 13·7)	26·7% (21·2 to 32·2)	33·9% (25·1 to 41·6)	154·3% (80·9 to 251·7)	28·1% (9·0 to 43·8)	7·5% (5·9 to 9·3)	19·4% (15·7 to 23·2)	24·8% (18·4 to 31·4)	161·8% (89·1 to 255·7)	27·8% (10·3 to 51·3)
Georgia	11·5% (9·3 to 14·3)	30·0% (25·1 to 35·5)	38·3% (29·1 to 47·7)	164·4% (98·3 to 242·3)	28·1% (9·9 to 51·8)	7·9% (6·0 to 10·3)	22·5% (18·8 to 26·6)	30·9% (22·5 to 38·7)	187·9% (106·3 to 289·9)	37·8% (9·5 to 63·0)
Hawaii	9·9% (7·6 to 12·7)	26·0% (20·9 to 31·5)	33·1% (26·3 to 40·7)	166·2% (91·1 to 261·7)	28·7% (13·8 to 47·5)	11·4% (9·0 to 14·3)	25·0% (20·8 to 29·8)	33·0% (27·1 to 40·5)	121·8% (61·4 to 191·0)	32·8% (22·9 to 51·4)
Idaho	10·4% (8·2 to 13·0)	28·7% (23·5 to 34·3)	39·1% (27·3 to 49·2)	179·9% (102·7 to 273·6)	36·8% (8·1 to 64·8)	6·7% (5·1 to 8·4)	20·1% (16·6 to 24·1)	29·9% (20·5 to 38·1)	205·3% (118·4 to 317·5)	49·1% (13·8 to 79·0)
Illinois	9·9% (7·9 to 12·3)	27·7% (23·2 to 33·2)	35·0% (26·8 to 43·3)	182·9% (110·4 to 272·3)	26·9% (8·3 to 50·3)	8·8% (7·0 to 10·7)	22·1% (18·0 to 26·0)	31·4% (21·7 to 39·4)	154·9% (92·3 to 233·8)	42·7% (11·4 to 65·6)
	11·2%	25·4%	48·2%	219·2%	27·4%	9·6%	26·2%	27·3%	176·3%	42·8%

as those with a B.M.I. at or over 30. The authors acknowledged that B.M.I. is an imperfect measure that may not capture variations in body structure across the population. But from a scientific perspective, experts said, B.M.I. is correlated with other measures of body fat and is a practical tool for studying it at a population level.

Is correlation enough?

Question for discussion: Is BMI a prediction?

Exercise: The article claims "The authors acknowledged that B.M.I. is an imperfect measure that may not capture variations in body structure across the population. But from a scientific perspective, experts said, B.M.I. is correlated with other measures of body fat and is a practical tool for studying it at a population level."

This is a powerful claim if BMI is, indeed, a prediction model. It says that to make statistical inference (e.g. that the distribution of BMI has changed over time) we don't need the true outcome, we just need something that is **correlated** with the true outcome. That's easy to do with a prediction model, so that's awesome!

Let's just check

First, generate an outcome Y that is a linear function of 3 X variables plus gaussian noise, so Y=Xbeta+epsilon where beta is the true regression parameter and epsilon is the noise.

Now, take a variable ystar that is correlated with Y with some correlation rho.

Create a simulation that repeatedly generates draws of Y, X, and ystar and runs the regression of ystar on X.

Plot the coverage of the 95% confidence interval using the true value of beta.

Do this for lots of values of rho. Your final plot should have the rho on the X axis and coverage for each beta (individually) on the Y. You'll know your simulation is correct if you have 95% coverage when rho=1.

Let's just check

First, generate an outcome Y that is a linear function of 3 X variables plus gaussian noise, so Y=Xbeta+epsilon where beta is the true regression parameter and epsilon is the noise.

Now, take a variable ystar that is correlated with Y with some correlation rho.

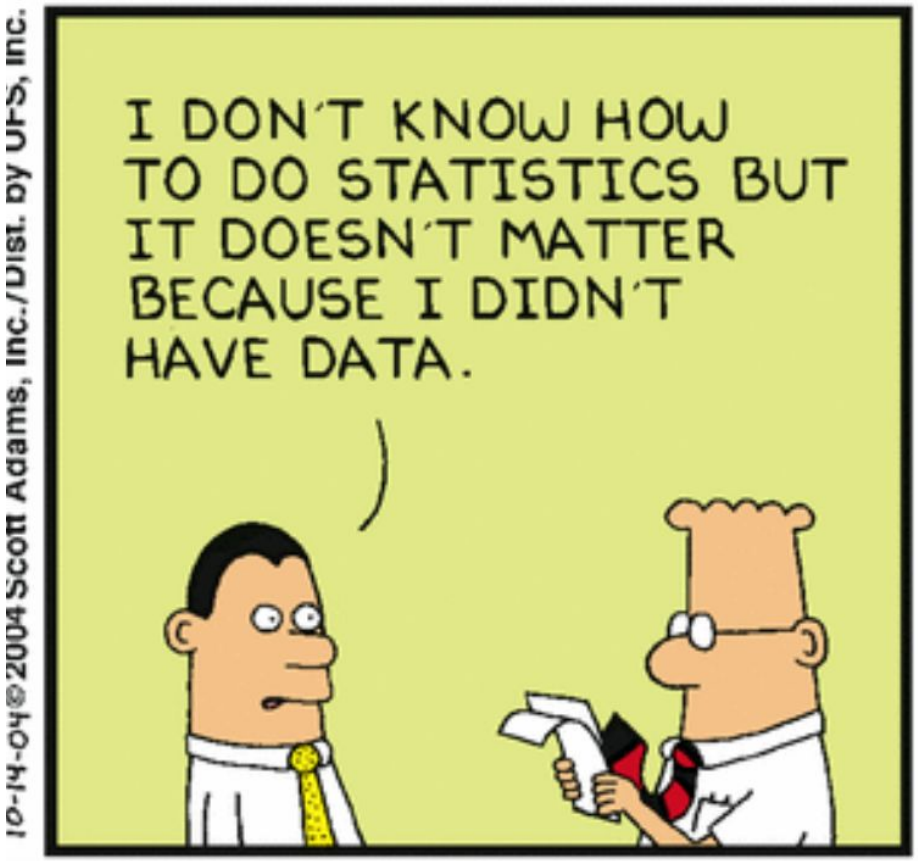
Create a simulation that repeatedly generates draws of Y, X, and ystar and runs the regression of ystar on X.

Plot the coverage of the 95% confidence interval using the true value of beta.

Do this for lots of values of rho. Your final plot should have the rho on the X axis and coverage for each beta (individually) on the Y. You'll know your simulation is correct if you have 95% coverage when rho=1.

Exercise: Write a paragraph to Ninan Agrawal explaining what you found.

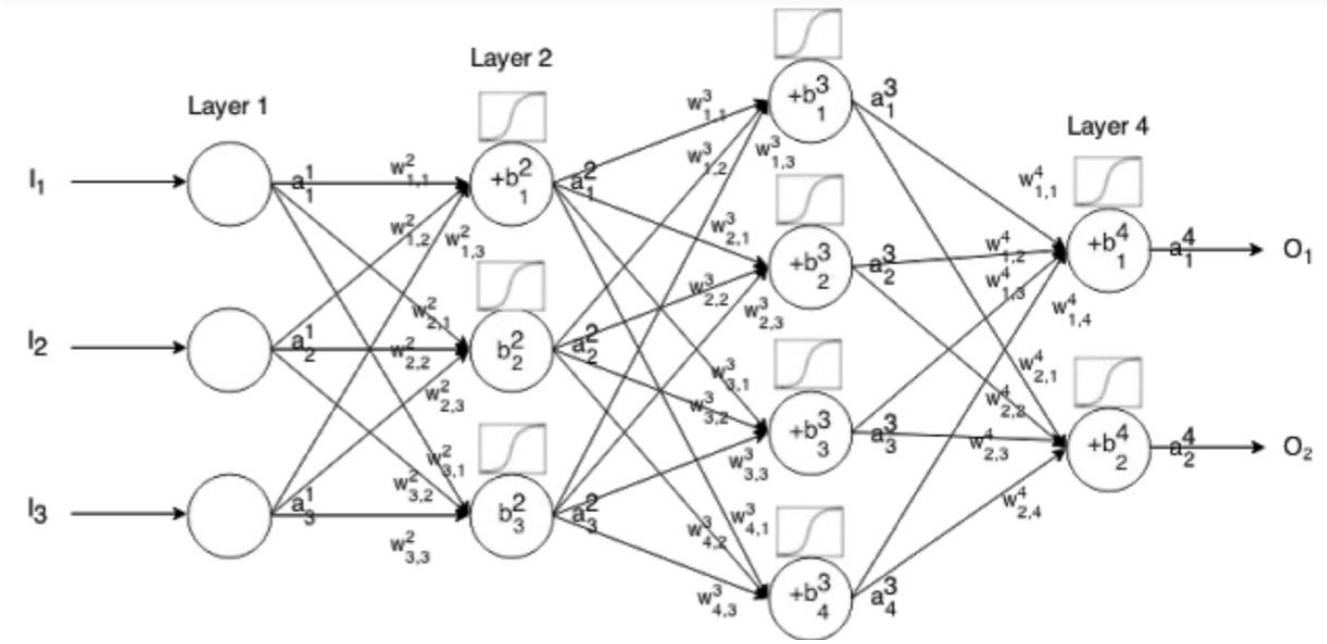
A problem of the past!



https://dilbert.com/search_results?terms=Statistical+Analysis

Making predictions is easy to do; hard to do well

$$f(DataYouCanGet) = (DataYouWant)\text{-}ish$$



This happens everywhere!

Machine learning identifies interacting genetic variants contributing to breast cancer risk: A case study in Finnish cases and controls

Hamid Behravan,[#1] Jaana M. Hartikainen,[1] Maria Tengström,[2,3] Katri Pylkäs,[4] Robert Winqvist,[4] Veli-Matti Kosma,[#1,5] and Arto Mannermaa[#1,5]

nature > nature genetics > articles > article

nature genetics

Article | Published: 01 May 2019

A transcriptome-wide association study of high-grade serous epithelial ovarian cancer identifies new susceptibility genes and splice variants

Machine learning models in electronic health records can outperform conventional survival models for predicting patient mortality in coronary artery disease

Andrew J. Steele, Spiros C. Denaxas, Anoop D. Shah, Harry Hemingway, Nicholas M. Luscombe

Performance of four computer-coded verbal autopsy methods for cause of death assignment compared with physician coding on 24,000 deaths in low- and middle-income countries

Nikita Desai, Lukasz Aleksandrowicz, Pierre Miasnikof, Ying Lu, Jordana Lerian, Peter Byass, Stephen Tollman, Paul Mee, Dewan Alam, Suresh Kumar Rathi, Abhishek Singh, Rajesh Kumar,

Enterotypes of the human gut microbiome

Manimozhiyan Arumugam[1*], Jeroen Raes[1,2*], Eric Pelletier[3,4,5], Denis Le Paslier[3,4,5], Takuji Yamada[1], Daniel R. Mende[1], Gabriel R. Fernandes[1,6], Julien Tap[1,7], Thomas Bruls[3,4,5], Jean-Michel Batto[7], Marcelo Bertalan[8], Natalia Borruel[9], Francesc Casellas[9], Leyden Fernandez[10], Laurent Gautier[8], Torben Hansen[11,12], Masahira Hattori[13], Tetsuya Hayashi[14], Michiel Kleerebezem[15], Ken Kurokawa[16], Marion Leclerc[7], Florence Levenez[7], Chaysavanh Manichanh[9], H. Bjørn Nielsen[8], Trine Nielsen[11], Nicolas Pons[7], Julie Poulain[3], Junjie Qin[17], Thomas Sicheritz-Ponten[8,18], Sebastian Tims[15], David Torrents[19,20], Edgardo Ugarte[9], Erwin G. Zoetendal[15], Jun Wang[11,21,22,23], Francisco Guarner[9], Oluf Pedersen[11,21,22,23], Willem M. de Vos[15,24],

nature > nature genetics > analyses > article

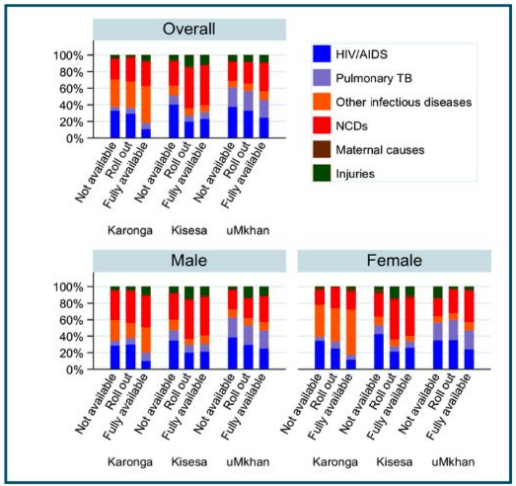
nature genetics

Analysis | Published: 16 January 2017

Case-control association mapping by proxy using family history of disease

Jimmy Z Liu, Yaniv Erlich & Joseph K Pickrell

Nature Genetics 49, 325–331 (2017) | Download Citation



Paradigm shift, really ??

SOME FURTHER RESULTS ON ERRORS IN DOUBLE SAMPLING TECHNIQUE*

By CHAMELI BOSE
Statistical Laboratory, Calcutta

1. The author (1942, 1943) studied some types of double sampling technique which consists in obtaining an expression for a character y, sometimes difficult or uneconomic to measure directly, in terms of an appreciably correlated character x, easier to obtain. Such problems were studied also by Neyman (1938), and Cochran (1939); but their methods of approach were different. Snedecor and King (1942) obtained a formula for the variance of the forecasting equation for

Sampling Techniques

third edition

1977

WILLIAM G. COCHRAN

Professor of Statistics, Emeritus
Harvard University

12.6 REGRESSION ESTIMATORS

In some applications of double sampling the auxiliary variate x_i has been used to make a regression estimate of $\bar{Y}$. In the first (large) sample of size n', we measure only x_i; in the second, a random subsample of size $n = \nu n' = n'/k$ where the fraction ν is chosen in advance, we measure both x_i and y_i. The estimate of $\bar{Y}$ is

$$\bar{y}_{lr} = \bar{y} + b(\bar{x}' - \bar{x}) \tag{12.48}$$

where $\bar{x}'$, $\bar{x}$ are the means of the x_i in the first and second samples and b is the least squares regression coefficient of y_i on x_i, computed from the second sample.

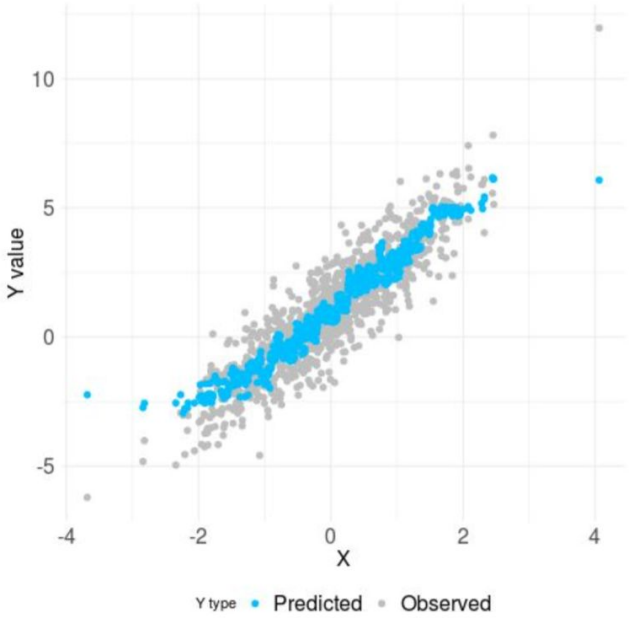
Where does this leave us?

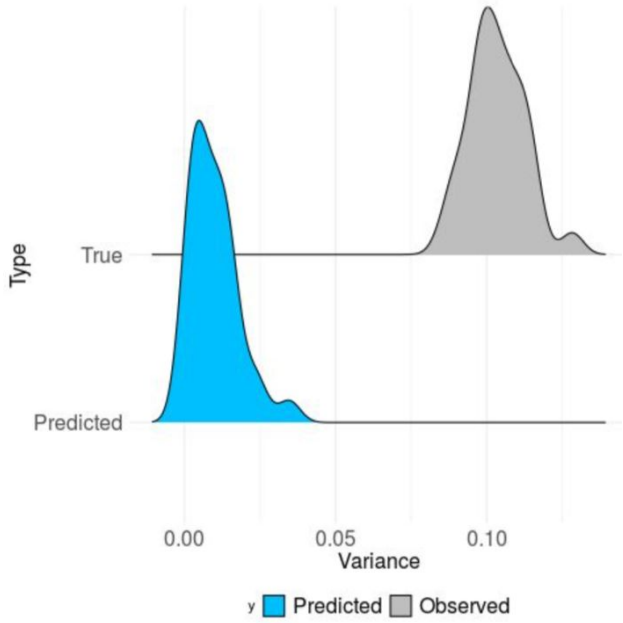
Many modern statistical problems involve using the data you can easily access to ***predict*** the variables you want.

Predictions then get used for downstream analysis or policy decision-making.

What are the implications “post-prediction” inference?

$$
\begin{aligned}
X &\sim N(0,1) \\
Z &\sim N(0,1) \\
\mu &= 1 + \beta X \\
Y &\sim N(\mu,1) \\
E[Y|X,Z] &= \beta_0 + \beta_1 X + \beta_2 Z \\
Y_p &= \hat{\beta}_0 + \hat{\beta}_1 X + \hat{\beta}_2 Z
\end{aligned}
$$





Models

y_i is the observed outcome for person i
x_i is covariate(s) of interest
z_i are additional covariates

- **Generative/"state of nature" model**

$$g\left(E[y_i|x_i, z_i]\right) = h(x_i, z_i)$$

- **Prediction model**

$$y_i^p = f(x_i, z_i)$$

- **Inference model**

$$g\left(E[y_i|x_i]\right) = M_{x_i}\vec{\beta}$$

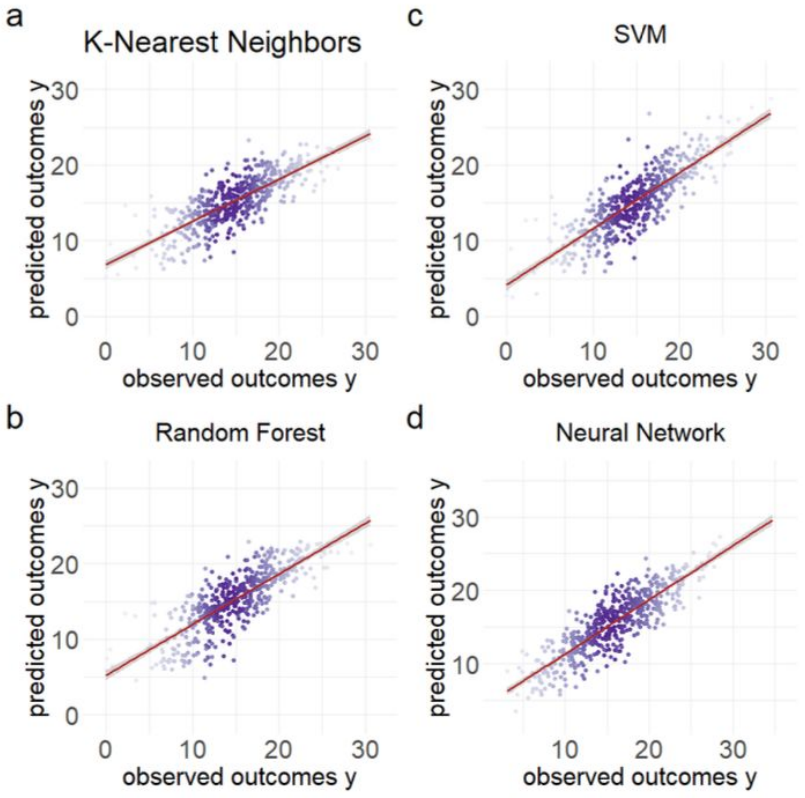
- **Post-prediction Inference model**

$$g\left(E[y_i^p|x_i]\right) = M_{x_i}\vec{\beta}_p$$

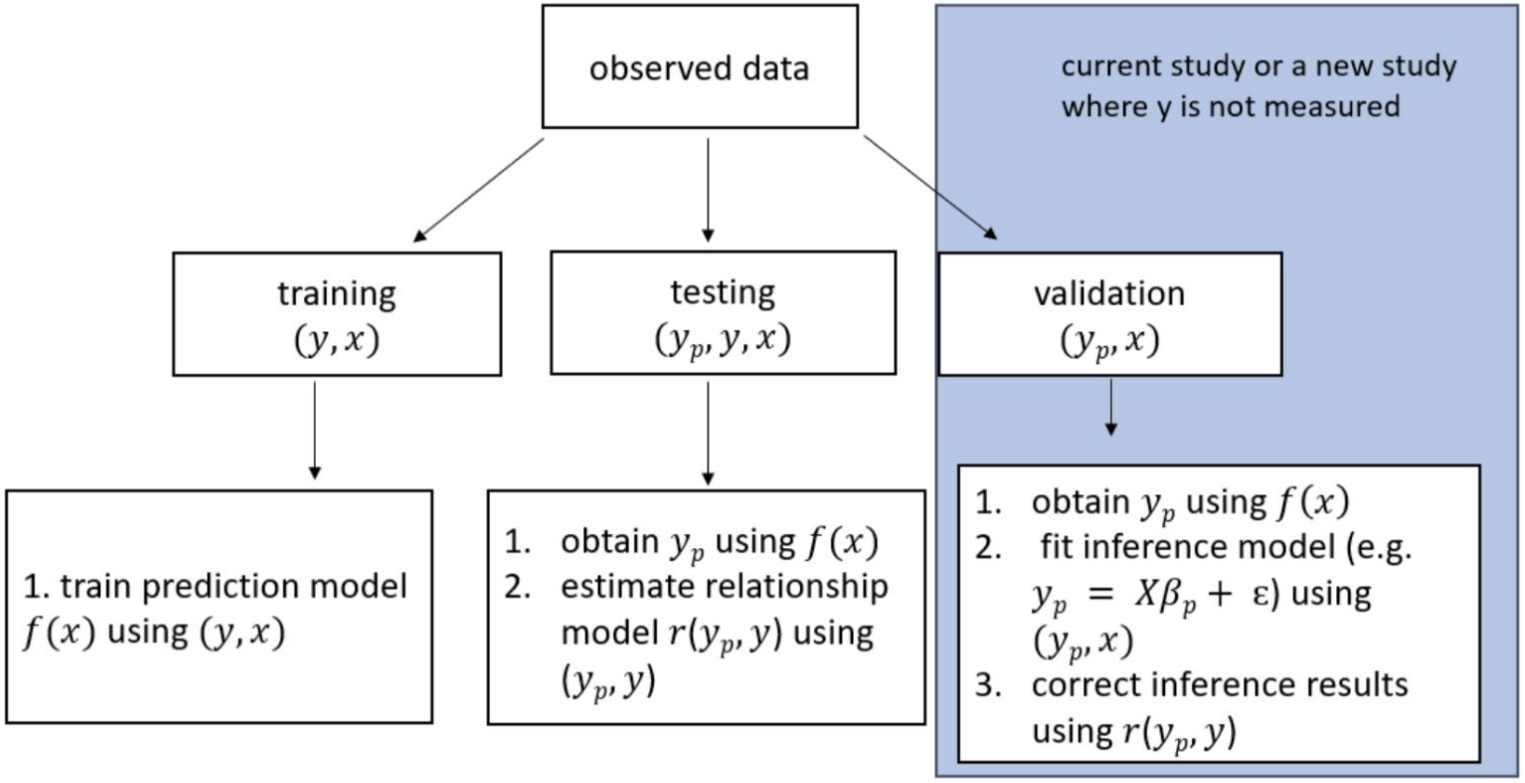
Our approach: relationship model

$E[y_i^p|x_i]$ is hard to model and depends on the ML algorithm

$$\mathbf{E}[y^p|x] = \mathbf{E}[\mathbf{E}[y^p|x,y]|x]$$
$$\approx \mathbf{E}[\mathbf{E}[y^p|y]|x]$$



Post-prediction inference



y: observed outcome x: observed covariate y_p: predicted outcome
$f(x)$: prediction model $r(y_p, y)$: relationship model

Another approach, prediction-powered inference

The previous approach relies on the *relationship* between the observed and predicted y.

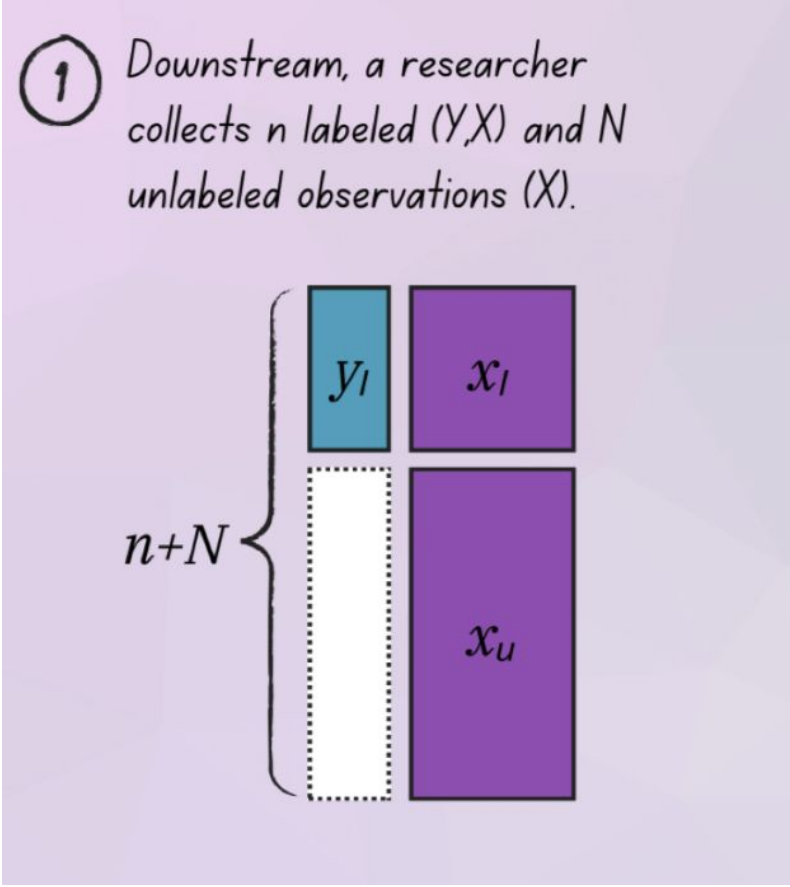
Key observation: For many models, that relationship is pretty simple (even if the prediction model really isn't!)

Another approach, prediction-powered inference

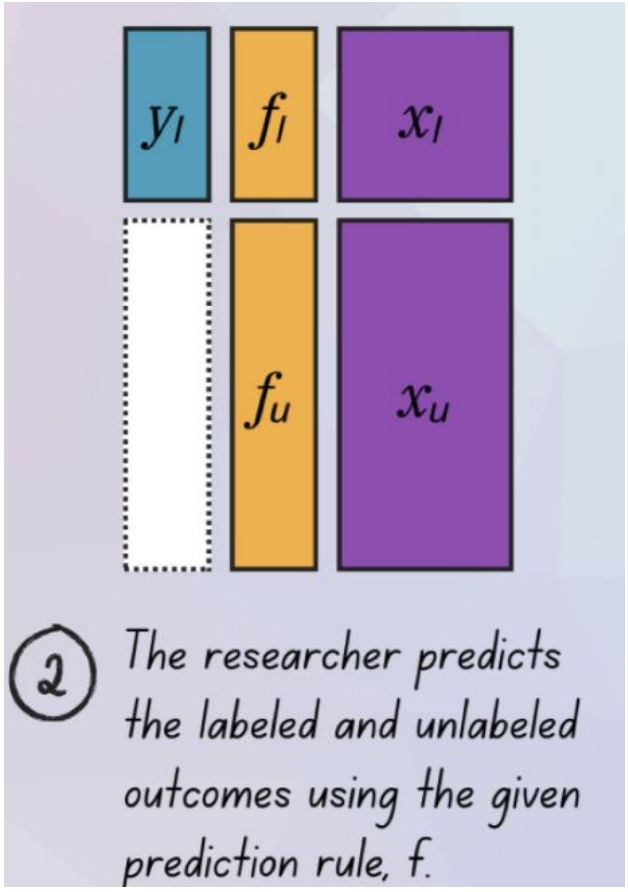
Prediction-powered inference takes a different approach and, instead, relies on a correction in the space of the covariates.

It explicitly computes a “correction” factor for the regression coefficient.

Another approach, prediction-powered inference



Another approach, prediction-powered inference



Another approach, prediction-powered inference

3 The researcher conducts inference on predicted data (here, prediction-powered inference) by correcting the estimate and standard error in the unlabeled data using the relationship between the true and predicted outcomes and the features in the labeled data.

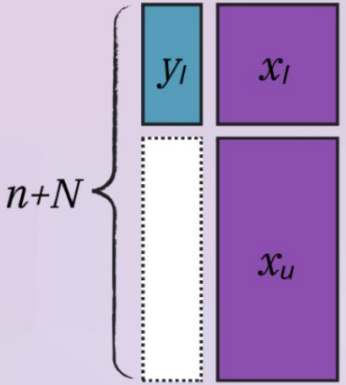
$$\hat{\theta}^{naive} : f_u \sim x_u$$

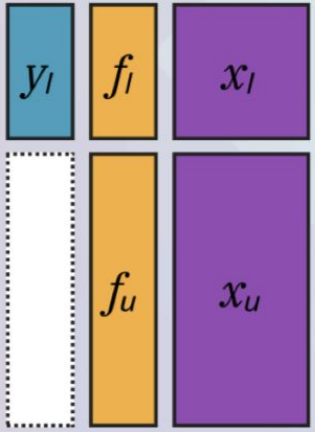
$$\hat{\Delta} : y_l - f_l \sim x_l$$

$$\hat{\theta}^{PPI} = \hat{\theta}^{naive} + \hat{\Delta}$$

Another approach, prediction-powered inference

1. Downstream, a researcher collects n labeled (Y,X) and N unlabeled observations (X).





2. The researcher predicts the labeled and unlabeled outcomes using the given prediction rule, f.

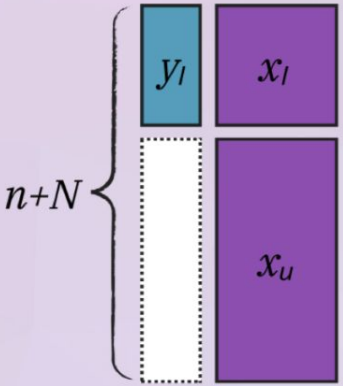
3. The researcher conducts inference on predicted data (here, prediction-powered inference) by correcting the estimate and standard error in the unlabeled data using the relationship between the true and predicted outcomes and the features in the labeled data.

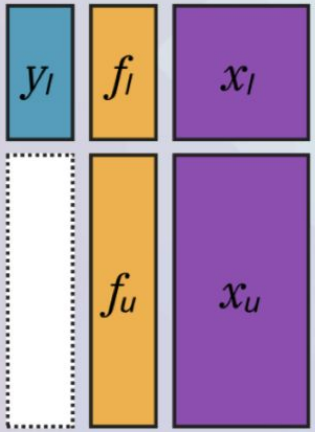
$$\hat{\theta}^{naive} : f_u \sim x_u$$

$$\hat{\Delta} : y_l - f_l \sim x_l$$

Another approach, prediction-powered inference

1. Downstream, a researcher collects n labeled (Y,X) and N unlabeled observations (X).





2. The researcher predicts the labeled and unlabeled outcomes using the given prediction rule, f.

3. The researcher conducts inference on predicted data (here, prediction-powered inference) by correcting the estimate and standard error in the unlabeled data using the relationship between the true and predicted outcomes and the features in the labeled data.

$$\hat{\theta}^{naive}: f_u \sim x_u$$

$$\hat{\Delta}: y_l - f_l \sim x_l$$

$$\hat{\theta}^{PPI} = \hat{\theta}^{naive} + \hat{\Delta}$$

Another approach, prediction-powered inference

We use PPI++, which gives the corrected parameter estimates by optimizing the objective function:

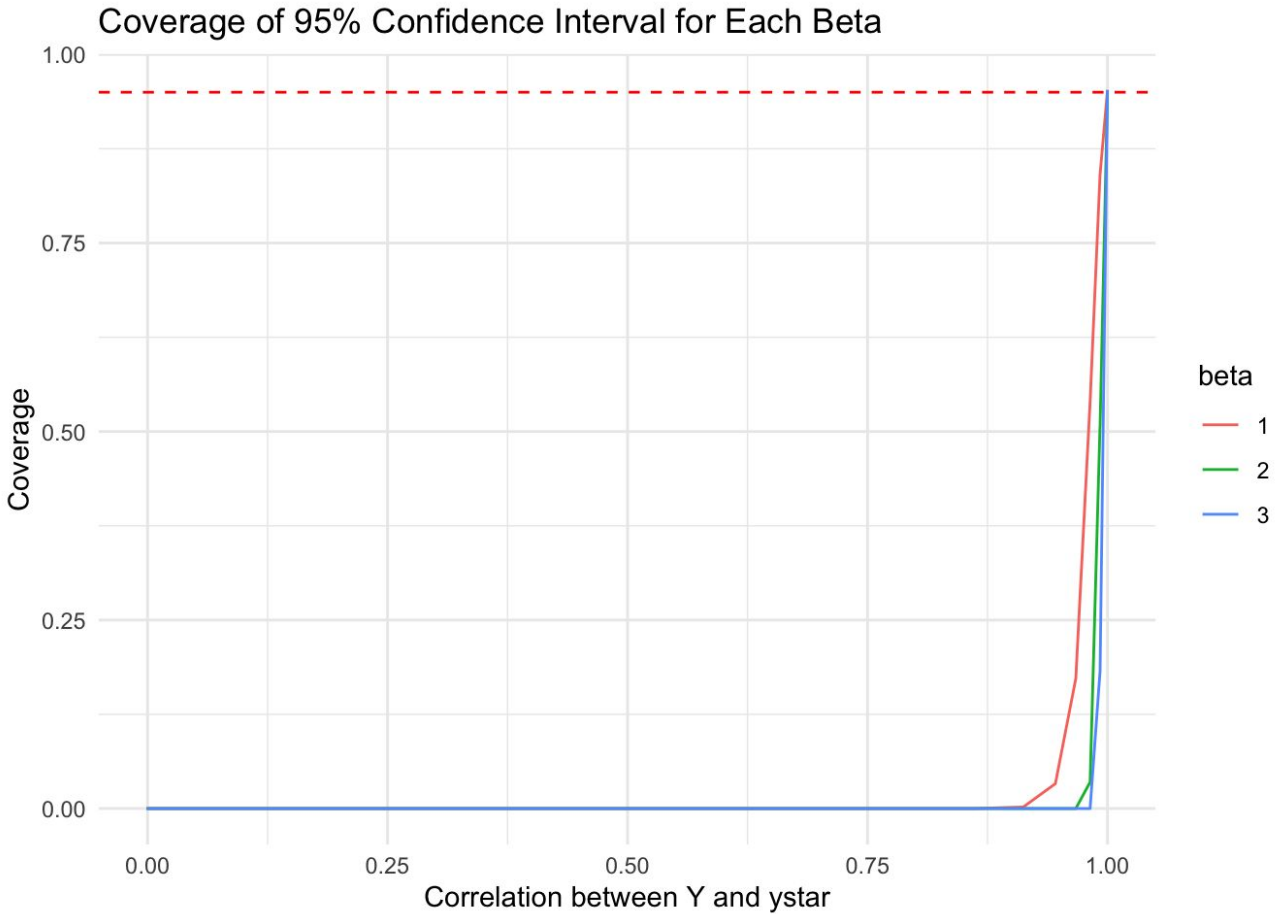
$$L(\theta) = \mathbb{E}[l_\theta(X, Y)], \quad L^f(\theta) = \mathbb{E}[l_\theta(X, f(X)]$$

$$L^{PP}(\theta) = L_n(\theta) + \lambda(\tilde{L}^f_N(\theta) - L^f_n(\theta)).$$

Where n is the number of labaled observations, N is the number of unlabaled observations, $l_\theta(X, Y)$ is the loss function, Y is the true output and $f(X)$ is the predicted output.

$\lambda \in [0, 1]$ is the tuning parameter controlling the interpolation between classical and prediction-powered inference.

Summary

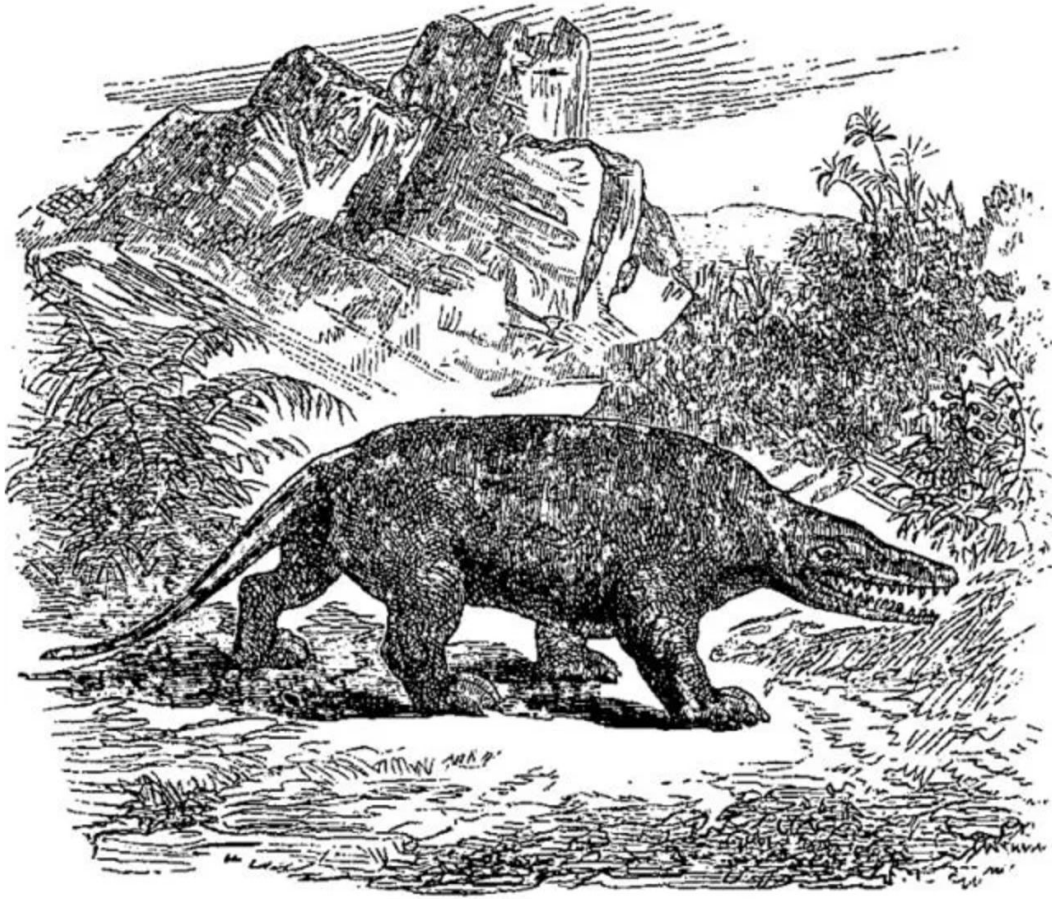

Coverage of 95% Confidence Interval for Each Beta
1.00
0.75
0.50
0.25
0.00
Coverage
0.00
0.25
0.50
0.75
1.00
Correlation between Y and ystar
beta
1
2
3

Correcting for predicted outcomes

Tyler H McCormick
University of Washington
CSSS 594

First, a thought exercise

How do you draw a dinosaur?



Mégalosaure restauré.

From “Les animaux d’autrefois” by Victor Meunier, 1869

https://medium.com/northwest-jammin/the-evolution-of-dinosaur-art-db4a8694a8c6

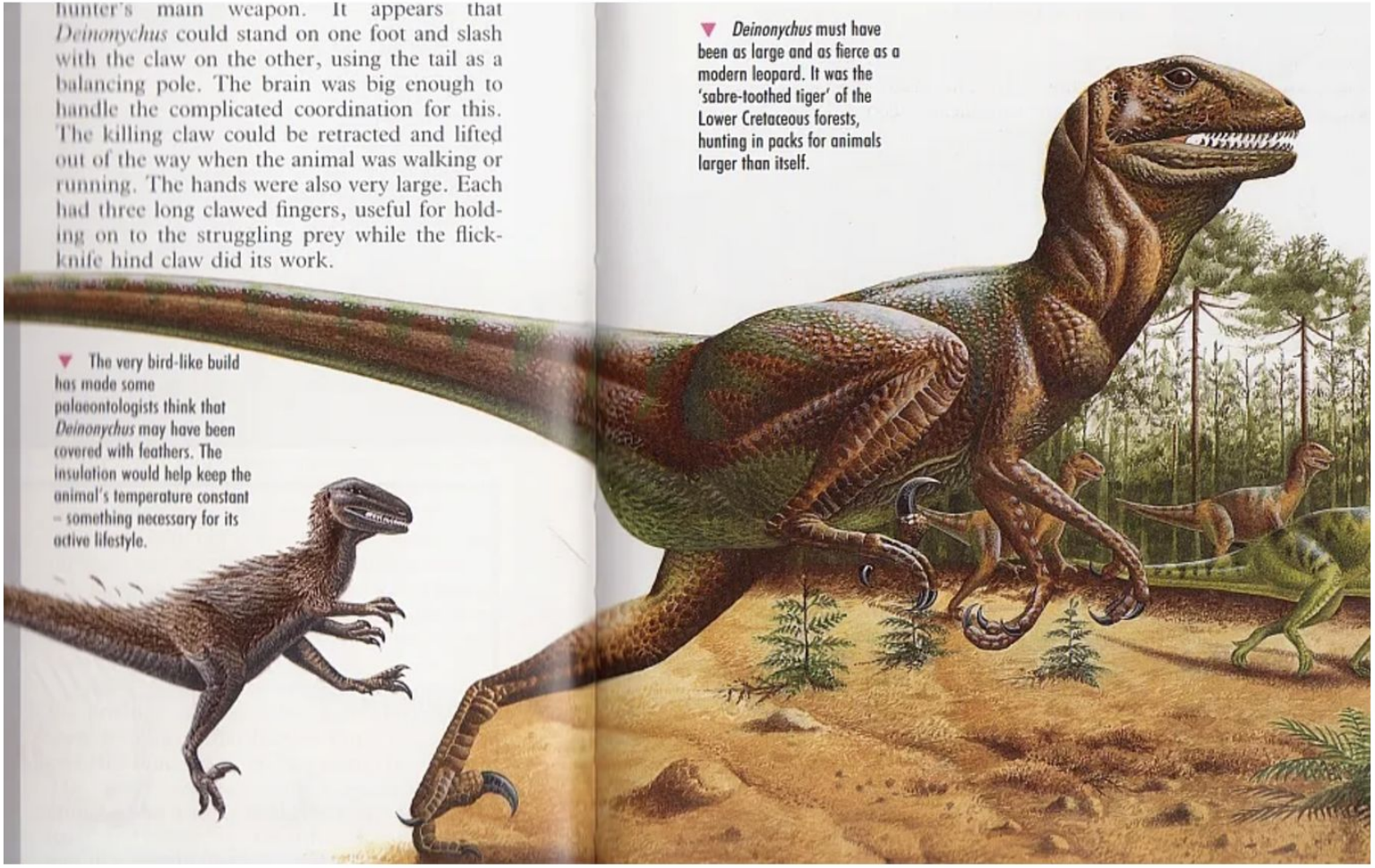
Leaping Laelaps — Charles R. Knight, 1897

Fantasia, 1940

hunter's main weapon. It appears that *Deinonychus* could stand on one foot and slash with the claw on the other, using the tail as a balancing pole. The brain was big enough to handle the complicated coordination for this. The killing claw could be retracted and lifted out of the way when the animal was walking or running. The hands were also very large. Each had three long clawed fingers, useful for holding on to the struggling prey while the flick-knife hind claw did its work.

▼ The very bird-like build has made some palaeontologists think that *Deinonychus* may have been covered with feathers. The insulation would help keep the animal's temperature constant – something necessary for its active lifestyle.

▼ *Deinonychus* must have been as large and as fierce as a modern leopard. It was the 'sabre-toothed tiger' of the Lower Cretaceous forests, hunting in packs for animals larger than itself.



From The Illustrated Dinosaur Encyclopedia by Dougal Dixon, 1988

Jurassic Park, 1993

https://www.nationalgeographic.com/animals/article/160405-dinosaurs-feathers-birds-museum-new-york-science

How do you draw a dinosaur?

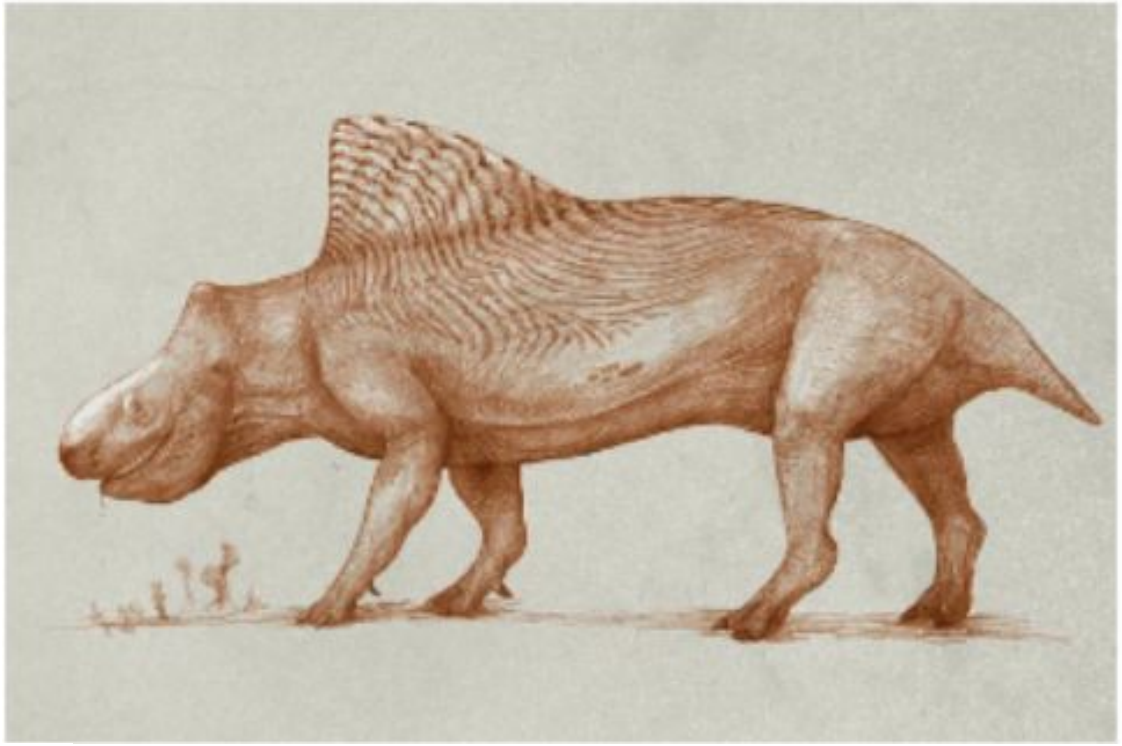
Tldr, nobody knows

Yet!!

Statistical inference

Albrecht Dürer

Prediction

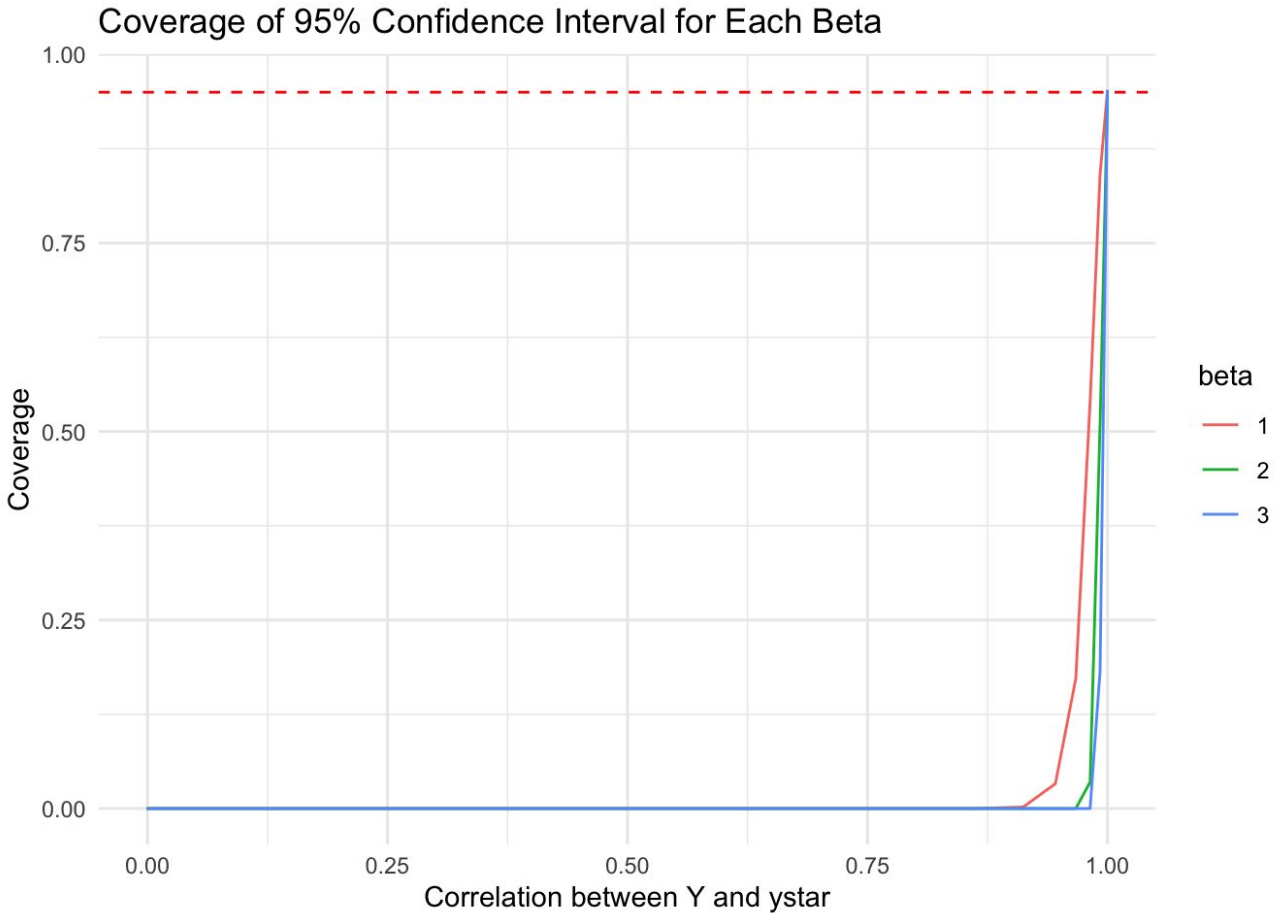


C.M. Kosemen

Solving modern scientific questions–

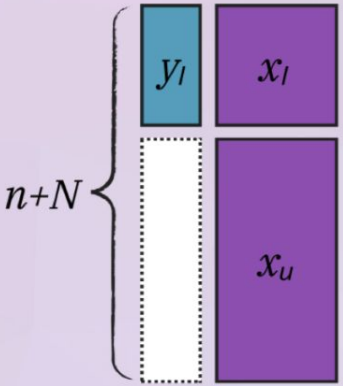
Figure 1: Artist renderings of a rhinoceroses based on limited information. Left: Albrecht Dürer's *The Rhinoceros*, woodcutting (1515); Right: C.M. Kosemen's re-imagining of a rhinoceros based on its skeleton.

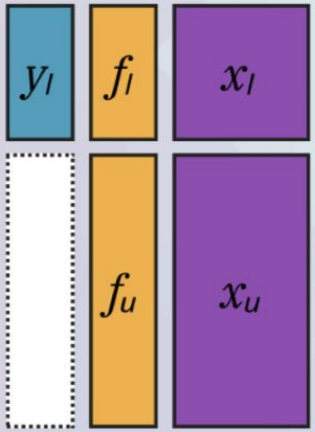
Summary


Coverage of 95% Confidence Interval for Each Beta
1.00
0.75
0.50
0.25
0.00
Coverage
0.00
0.25
0.50
0.75
1.00
Correlation between Y and ystar
beta
1
2
3

IPD: Recap

1. Downstream, a researcher collects n labeled (Y,X) and N unlabeled observations (X).





2. The researcher predicts the labeled and unlabeled outcomes using the given prediction rule, f.

3. The researcher conducts inference on predicted data (here, prediction-powered inference) by correcting the estimate and standard error in the unlabeled data using the relationship between the true and predicted outcomes and the features in the labeled data.

$$\hat{\theta}^{naive} : f_u \sim x_u$$

$$\hat{\Delta} : y_l - f_l \sim x_l$$

$$\hat{\theta}^{PPI} = \hat{\theta}^{naive} + \hat{\Delta}$$

ipd: Inference on Predicted Data

Overview

`ipd` is an open-source `R` software package for the downstream modeling of an outcome and its associated features where a potentially sizable portion of the outcome data has been imputed by an artificial intelligence or machine learning (AI/ML) prediction algorithm. The package implements several recent proposed methods for inference on predicted data (IPD) with a single, user-friendly wrapper function, `ipd`. The package also provides custom `print`, `summary`, `tidy`, `glance`, and `augment` methods to facilitate easy model inspection.

Background

Using predictions from pre-trained algorithms as outcomes in downstream statistical analyses can lead to biased estimates and misleading conclusions. The statistical challenges encountered when drawing inference on predicted data (IPD) include:

1. Understanding the relationship between predicted outcomes and their true, unobserved counterparts.
2. Quantifying the robustness of the AI/ML models to resampling or uncertainty about the training data.
3. Appropriately propagating both bias and uncertainty from predictions into downstream inferential tasks.

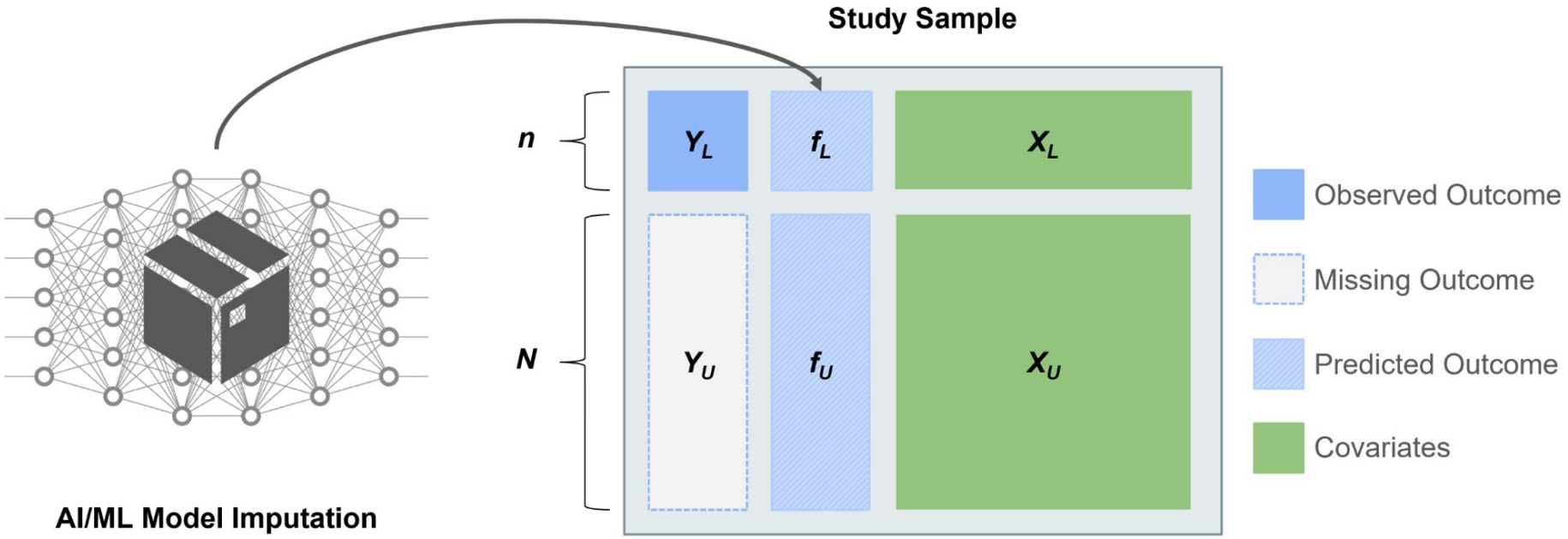
Several works have proposed methods for IPD, including post-prediction inference (PostPI) by Wang et al., 2020, prediction-powered inference (PPI) and PPI++ by Angelopoulos et al., 2023a and Angelopoulos et al., 2023b, and post-prediction adaptive inference (PSPA) by Miao et al., 2023. Each method was developed to perform inference on a quantity such as the outcome mean or quantile, or a regression coefficient, when we have:

IPD

Using predictions from pre-trained algorithms as outcomes in downstream statistical analyses can lead to biased estimates and misleading conclusions. The statistical challenges encountered when drawing inference on predicted data (IPD) include:

1. Understanding the relationship between predicted outcomes and their true, unobserved counterparts.
2. Quantifying the robustness of the AI/ML models to resampling or uncertainty about the training data.
3. Appropriately propagating both bias and uncertainty from predictions into downstream inferential tasks.

1. A dataset consisting of our outcome and features of interst, where the outcome is only observed for a small ‘labeled’ subset and missing for a, typically larger, ‘unlabeled’ subset.
2. Access to an algorithm to predict the missing outcome in the entire dataset using the fully observed features.



1.1 PostPI Bootstrap Correction (Wang et al., 2020)

```
#-- Specify the Formula

formula <- Y - f ~ X1

#-- Fit the PostPI Bootstrap Correction

nboot <- 200

ipd::ipd(formula,

  method = "postpi_boot", model = "ols", data = dat_ols, label = "set",

  nboot = nboot) |>

  summary()
```

2. Prediction-Powered Inference (PPI; Angelopoulos et al., 2023)

```
#-- Fit the PPI Correction

ipd::ipd(formula,

  method = "ppi", model = "ols", data = dat_ols, label = "set") |>

  summary()
```

3. PPI++ (Angelopoulos et al., 2023)

```
#-- Fit the PPI++ Correction

ipd::ipd(formula,

  method = "ppi_plusplus", model = "ols", data = dat_ols, label = "set") |>

  summary()
```

Method	Mean Estimation	Quantile Estimation	Linear Regression	Logistic Regression	Poisson Regression	Multiclass Regression
PostPI	❌	❌	✅	✅	❌	❌
PPI	✅	✅	✅	✅	❌	❌
PPI++	✅	✅	✅	✅	❌	❌
PSPA	✅	✅	✅	✅	✅	❌
PSPS	❌	❌	❌	❌	❌	❌
PDC	❌	❌	❌	❌	❌	❌
Cross-PPI	❌	❌	❌	❌	❌	❌
PPBoot	❌	❌	❌	❌	❌	❌
DSL	❌	❌	❌	❌	❌	❌