

The Explanatory Power of Causal Effects

James Stratton and Nicolaj Thor*

February 4, 2026

ABSTRACT

How much of the observed variation in an outcome Y does a variable X *causally* explain? We propose a causal R^2 (CR^2) to answer such questions and quantify the importance of X in determining Y . CR^2 is the fit of a causal model, identified by combining observational and experimental data, and captures the reduction in Y 's variance from eliminating X 's causal effect. In applications, class size predicts 8% of reading scores but causally explains only 3%; institutions explain one-fifth of cross-country income variation; and salt intake raises blood pressure similarly across genders, but explains far less among women.

*Emails: jstratton@g.harvard.edu and thor@brown.edu. The author order is random. We are grateful for comments from Adamson Bryant, Raj Chetty, Cole Davis, Cameron Deal, John Friedman, Peter Hull, Kosuke Imai, Larry Katz, Toru Kitagawa, Soonwoo Kwon, Florian Mudekereza, Sendhil Mullainathan, Mandy Pallais, Yechan Park, Jonathan Roth, Nico Rotundo, Jesse Shapiro, Elie Tamer, Winnie van Dijk, and audiences at Brown University, Harvard University, the University of Warwick, and the 2025 American Causal Inference Conference. We are also grateful to have participated in the Alexander and Diviya Magaro Peer Pre-Review Program at Harvard's Institute for Quantitative Social Science.

1 Introduction

Social scientists often seek to explain observed variation in outcomes. What are the “causes of inequality in incomes” (Mincer 1958)? Why do some cities grow while others stall (Saiz 2010; Glaeser 2011)? And what are the “fundamental causes of the large differences in income per capita across countries” (Acemoglu, Johnson, and Robinson 2001)? Suppose a researcher, in answer to such a question, proposes that a variable X is a main cause of observed variation in an outcome Y . Making and assessing this claim requires a measure of the share of variation in Y causally explained by X . Such a measure would quantify the share of wage dispersion explained by education, the share of variation in city growth explained by zoning, and the share of cross-country income differences explained by institutions. In this paper, we propose such a measure, which to our knowledge does not yet exist, and show its usefulness.

The R^2 would suffice as a measure if by “explain” we meant “predict”. Yet economists rarely use the R^2 precisely because we typically seek causal, not predictive, explanations: if X predicts, but does not cause, Y , then X does not “explain” variation in the sense that interests us. Since R^2 cannot distinguish between variables that causally affect an outcome *vs.* those merely associated with it, R^2 is not well-suited to the question. This problem holds for ordinary fit statistics more broadly: they use only the joint distribution of Y and X and hence cannot untangle correlation from causation.

In contrast, causal inference measures how X *affects* Y (*e.g.*, how much an extra year of school raises wages), but not whether *observed variation* in Y is explained by X . These two concepts are related, but different. A variable that has no causal effect cannot causally explain variation in an outcome. Yet a variable that causally affects the outcome can explain either a large or a small share of variation in the outcome, depending on its variance and covariance with other factors. To see the distinction between causal effect and causally explained variation in an extreme case, note that if all workers had the same education, education would explain no variation in wages, even with a large causal effect.¹ As such, the causal effect of a variable alone does not measure

¹While merely illustrative, this is not uncommon in economics: many RCTs test treatments that are new in the study population (*e.g.*, Miguel and Kremer 2004; Kleven et al. 2011; Breza and Chandrasekhar 2019). More generally, the causal inference literature asks how a change in X would affect Y . We instead ask how much observed variation in Y is due to variation in X . Even when X varies a lot in the population *and* has a large causal effect, it may account for little of the observed cross-sectional variation in Y .

how much variation in an outcome it explains.

We propose to measure the share of variation in Y causally explained by a variable X in three steps: first, estimate the causal effect of X on Y ; second, compute the “best causal model” for the outcome as the model that minimizes mean squared error in the population, under the constraint that modelled differences in Y match that causal effect; third, evaluate the fit of this model in the population.

Our approach parallels the R^2 . Recall that the (non-parametric) R^2 evaluates how well we can predict Y from X using as a model the conditional expectation function, $m(x) := \mathbb{E}[Y \mid X = x]$. In particular, if, for a model f , we define $\text{MSE}(f) = \mathbb{E}[(Y - f(X))^2]$, then the $R^2 = 1 - \text{MSE}(m) / \text{Var}[Y]$. That is, the R^2 evaluates the goodness-of-fit of the conditional expectation function, the best predictive model for Y . Instead, we seek a *causal model* for Y , which we define as a model constrained to match the average treatment effects of X : that is, for any x, \hat{x} , we have that $\mu(x) - \mu(\hat{x})$ is the average treatment effect of a change from \hat{x} to x . The best causal model, μ , minimizes mean squared error in the population subject to this constraint.² This model captures the relation between the variable and the outcome after purging any confounding: the modelled differences in Y are only due to the causal effect of X and not the effect of omitted variables. We define the **causal R-squared** (CR^2) as the fit of that causal model: $\text{CR}^2 = 1 - \text{MSE}(\mu) / \text{Var}[Y]$.

We use CR^2 to formalise and measure the share of variance causally explained by a variable, and base this choice on several interpretations. As an *analogue to the predictive R^2* , the CR^2 captures variation explained by the variable, albeit causally rather than merely predictively. As a *thought experiment*, the CR^2 corresponds to the reduction in outcome variance from eliminating the average causal effect of the feature. This reduction in variance is the share of variance caused by the variable: it exists only because of the variable’s causal effect.

The causal R^2 has intuitive and useful properties. It equals 0 if X has no causal effect or does not vary in the population, and 1 if X fully determines it. It is unitless and bounded above by the predictive R^2 , with equality if observables are independent of unobservables.

The CR^2 is simple to estimate. It combines (i) causal effects, and (ii)

For instance, a highly-effective tutoring program which targets struggling students will not explain much of the variation in test scores. For more details, see section 4.3.

²Up to a constant, this best causal model equals the *average potential outcome* function, $\mathbb{E}[Y(x)]$, which is the expected outcome in an experiment that assigns all units to treatment x .

the joint distribution of (Y, X) to determine fit of the best causal model. Accordingly, identification requires (i) an *experimental dataset* to identify the causal effects, in which X' is randomly assigned, and Y' is the resulting outcome, and (ii) an *observational dataset* from the population in which to evaluate the fit of the best causal model. Our baseline setting for the experimental dataset is a randomized experiment, but the approach extends to quasi-experimental designs that recover an average treatment effect. A plug-in estimator is consistent, and inference follows from the bootstrap or the Delta method.

We show the practical use of CR^2 in four applications. Using data from Kremer et al. (2011), who randomized spring protection in Kenya, we find that variation in spring protection *causally explains* about a third of the existing variation in water quality in the population—more than in the experimental population. In Project STAR, class size *predicts* 8% of the variation in reading scores, but *causally explains* only 3%: the predictive power of class size mostly comes from omitted variables, not the causal effect of class size. Applying CR^2 to the settler mortality instrument in Acemoglu, Johnson, and Robinson (2001), about one fifth of cross-country income variation is causally explained by differences in institutions' extractiveness—answering the question from which the authors set out: what are the “fundamental causes of the large differences in income per capita across countries”. Lastly, we assess the share of variation in blood pressure, an important indicator of cardiovascular health, explained by salt consumption. The causal effect of salt intake on blood pressure is similar for women and men. Yet salt intake causally explains 7% of the variation in men's blood pressure, but under 1% in women. This difference arises because women's blood pressure varies more for reasons unrelated to salt.

These applications, we hope, illustrate the usefulness of CR^2 . To make the claim that a variable plays a large role in explaining observed variation in an outcome, a researcher must measure how much variation that variable explains. The CR^2 provides such a measure. It allows researchers to make such a claim well-founded, to quantify it (*e.g.* to say that a variable causally explains 10% of the variation in an outcome), and to evaluate existing theories. These CR^2 values also gauge the value of future research: if the explained share is low, then our theory of the outcome is incomplete.

The measure's value is largely explanatory rather than about policy counterfactuals. For instance, the CR^2 is valuable in making and

assessing the claim that differences in education are a main cause of differences in income. However, a policy-maker considering proposals to raise incomes might only care about the causal effect, even if it explains little *existing* variation in incomes. Goldberger (1979) illustrates this point with eyesight as an example: most cross-sectional variation in eyesight is genetic, but there is still great value in prescribing glasses—a non-genetic intervention.³

There are also other plausible definitions of causal explanatory power, some of which we describe. The CR^2 has the advantage of being simple, unitless, portable, and easy to estimate and interpret as the share of causally explained variation.

Related literature. Existing work extends R^2 to settings other than causal models.⁴ A small literature uses R^2 to bound omitted variable bias (Oster 2019; Cinelli and Hazlett 2020, 2025) or external validity bias (Andrews and Oster 2019). Those papers use the predictive R^2 to assess causal effects; we instead develop a causal counterpart of the R^2 .

We also connect to the economics literature on completeness, which compares the predictive power of theory-constrained with unconstrained models (*e.g.*, Peysakhovich and Naecker 2017; Fudenberg et al. 2022; Apesteguia and Ballester 2021). We study causal, not predictive models, though our metric can be viewed as the performance of a model constrained to be causal.

We also relate to work on causal attribution and the causes of effects. Causal attribution asks: given two observed variables (X_1, X_2) , and a known potential outcome function, how much of the joint effect of (X_1, X_2) should be attributed to X_1 vs. X_2 (*e.g.*, Datta, Sen, and Zick 2016; Heskes et al. 2020; Jung et al. 2022; Weitze 2025)? In contrast, we compare the variation explained by observed variables with the *total* variation in Y . The causes of effects literature asks whether a unit’s outcome would have differed under another treatment (*e.g.*, Pearl 1999; Halpern and Pearl 2005; Yamamoto 2012; Dawid and Musio 2022); we ask what share

³That being said, a policymaker interested in reducing *inequality* in outcomes may use CR^2 to assess whether inequality is largely owing to a given cause that warrants more attention.

⁴These settings include survival analysis (*e.g.*, Harrell et al. 1982), non-linear models (*e.g.*, McFadden 1973; Nagelkerke 1991; Li and Wang 2019), and Bayesian models (*e.g.*, Gelman and Pardoe 2006; Gelman et al. 2019). Recent work discusses “generalized” (R^2) (Wang, Jiang, and Liu 2017) and “out-of-sample” (Hawinkel, Waegeman, and Maere 2024) R^2 . Most relatedly, heritability in genetics, and the attributable fraction in epidemiology measure the share of variation in a characteristic attributable to some source—genetics for heritability (Visscher, Hill, and Wray 2008), and a risk factor for the attributable fraction (Porta 2014). Both measures assess predictive, not causal, relations.

of *population* variance comes from a given cause.

Gelman and Imbens (2013) distinguish forward causal questions (effects of causes) from reverse causal questions (apportioning outcomes to causes), which they view as model-checking. In their view, asking about the causes of an outcome involves assessing how well an existing model can explain the outcome, and whether it misses important determinants (p. 3):

“[I]f we ask, Why do incumbents get more contributions than challengers, ... get some measure for candidate quality ... and still see a large and statistically significant difference between the funds given to incumbents and challengers, then it seems we need more explanation.”

The causal R^2 formalizes this idea: it measures how well a causal model explains the outcome, and how much remains unexplained.

Outline. Section 2 gives an example to introduce the main ideas. Section 3 defines CR^2 . Section 4 discusses its properties. Section 5 covers identification, estimation, and inference. Section 6 presents applications. Section 7 concludes. Proofs are in the appendix; extra results in the Online Appendix.

2 An illustrative example

We first illustrate our approach in a simple example about the link between students’ test scores and class sizes. Let Y_i stand for student i ’s test score, C_i class size, and U_i other factors affecting scores. To build intuition, we begin with a constant effects model:

$$(1) \quad Y_i = \alpha + \beta C_i + \gamma U_i,$$

$$(2) \quad \begin{pmatrix} C_i \\ U_i \end{pmatrix} \sim \mathcal{N}(\mu, \Sigma), \quad \mu = \begin{pmatrix} \mu_C \\ 0 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix},$$

where (1) is the structural equation for Y_i , and (2) is the joint distribution of class size and the unobservable factor, which together determine the distribution of test scores.⁵ Scores and class size are observable; U_i is not.

What share of variation in test scores is explained by class size? If we construe this question *in a predictive sense*, we can answer it through *ob-*

⁵That is, $Y \sim \mathcal{N}(\alpha + \beta\mu_C, \beta^2 + \gamma^2 + 2\beta\gamma\rho)$, with $\text{Cov}[Y_i, C_i] = \beta + \gamma\rho$.

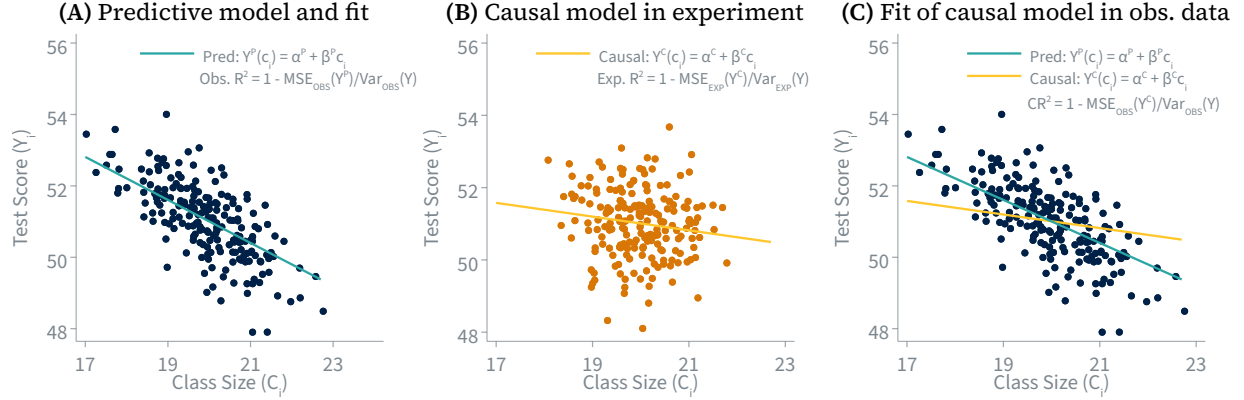
observational data: draws of (Y_i, C_i) . Figure 1(A) shows example realizations. The line of best fit converges, as the sample grows, to the conditional expectation $Y^P(c_i) := \mathbb{E}[Y_i | C_i = c_i] = \alpha^P + \beta^P c_i$, for $(\alpha^P, \beta^P) := (\alpha - \gamma\rho\mu_C, \beta + \gamma\rho)$. As is well-known, Y^P is the best predictive model: it minimizes squared prediction error. Its fit—the standard R^2 —measures the share of variation in scores that class size *predicts*.

Nonetheless, this R^2 does not capture how much variation class size *causally explains*: indeed, R^2 can be high even if class size has no causal effect ($\beta = 0$) but is strongly correlated with unobservables which affect the outcome (ρ, γ large in magnitude). More broadly, since the causal effect of C on Y is not identified by the joint distribution of (Y, C) , the observational data alone lack the information needed to compute variance *causally explained*.

Suppose then that the analyst also has access to data from an experiment that draws students from the population, randomly assigns class sizes (leaving other factors unchanged), and records the ensuing test scores. Figure 1(B) presents example realizations. Since class size is randomized, the best fit line converges to the average potential outcome $\mathbb{E}[Y_i(c_i)] = \alpha + \beta c_i$. We call this best fit line a *causal model* Y^C . It typically differs from the predictive model Y^P because of omitted variable bias: in Figure 1(B), Y^C is flatter than Y^P , consistent with positive selection into smaller classes ($\rho < 0$).

The best causal model recovers the true causal effect β . Because of this, it is tempting to interpret the R^2 in the experimental data as the share of variation in Y that is causally explained by variation in C . Unfortunately, this “experimental R^2 ” reflects the share of variation in Y that is causally explained by C in the experiment, but not in the population. These quantities differ. For one thing, in practice, often the *marginal distribution of treatment* differs: experimental treatment shares are often chosen to maximize precision, not to match the population distribution (Neyman 1934; Duflo, Glennerster, and Kremer 2007; Athey and Imbens 2017). Even when an experiment has been reweighted to match the marginal distribution of treatment in the population, the *outcome distribution* in the experiment generally differs from the population distribution. This is because the *joint distribution of treatment and unobservables*, which generates the distribution of the outcome, by definition differs between the experiment and the population: randomization in the experiment ensures that C is independent of U , whereas

Figure 1. Example: fit of predictive and causal models relating test scores to class size



Notes: This figure shows an example of the fit of predictive and causal models relating test scores (Y_i) to class size (C_i). Panel (A) shows sample realizations of observational data. The line of best fit is the conditional expectation function, *i.e.* the best predictive model. Its fit is the predictive R^2 . Panel (B) shows sample realizations of experimental data. The line of best fit is the average potential outcome function, *i.e.* the best causal model. Its fit in the experimental data is the “experimental R^2 ”. Panel (C) again shows the sample realizations of observational data and the predictive model from Panel (A). It also plots the causal model, the average potential outcome function from Panel (B), the experiment. The fit of this best causal model in the observational data is the causal R^2 .

the two may covary in the population.⁶ Since we seek to explain the observed outcome variation in the population, the experimental R^2 never suffices.

More broadly, from the observational data alone, we can estimate the joint distribution of Y and C , but not the causal effect of C . From the experimental data alone, we can estimate the causal effect of C , but not its explanatory power for Y in the population. We therefore propose to measure the share of variance causally explained by class size by estimating the causal effects of the variable of interest in experimental data, and then evaluating the fit of the corresponding best causal model in observational data (Panel (C)). We call this measure the causal R-squared, $CR^2 = 1 - \frac{MSE(Y^C)}{Var[Y]}$, and show that it is a useful formalisation of the concept of causally explained variation.

To interpret it, take a student with score y_i and class size c_i . The causal model for test scores suggests a score of $\alpha^C + \beta c_i$; the residual $y_i - (\alpha^C + \beta c_i)$ (the vertical distance between the realized outcome and the

⁶In the population, $Var[Y] = \beta^2 + \gamma^2 + 2\beta\gamma\rho$. In the experiment, $Var_E[Y] = \beta^2 Var_E[C] + \gamma^2$, where $Var_E[\cdot]$ is variance in the experiment. If the experiment is representative of class size in the population, $Var_E[Y] = \beta^2 + \gamma^2$.

causal model) is the component of student i 's score that is unexplained by class size. If students' scores differed *only* because of differences in class size ($\gamma = 0$), then this residual would be zero for all units: class size would fully explain variation in scores. If other differences between students affect scores, the residual captures the remaining variation. Summing these squared residuals and dividing by the outcome variance gives the share of variance not causally explained; the complement is the share explained. This parallels the usual R^2 , but the variation is explained causally, not just predictively. We show in section 3.3 that the interpretation of CR^2 as causally explained variation also results from several thought experiments.

3 Defining the causal R-squared

We now formalize the approach of the last section and define the causal R^2 more generally. We first describe our setting, and define causal models and model fit. We then introduce the non-parametric and linear CR^2 .

3.1 Setting

We use the standard Rubin Causal Model with two primitives: K feature variables (indexed by k) and a potential outcome function. Each *feature variable* X_k is a real-valued random variable taking values in \mathcal{X}_k . The feature vector is $X := (X_1, \dots, X_K)$ taking values in $\mathcal{X} := \mathcal{X}_1 \times \dots \times \mathcal{X}_K$, with joint distribution P_X . The feature variables and the *potential outcome function* $Y: \mathcal{X} \rightarrow \mathcal{Y}$ together determine the real-valued *outcome* $Y := Y(X)$.

We interpret $Y(x_i)$ as the outcome for a *unit* i with realized feature vector x_i . We thus treat the potential outcome function $Y(\cdot)$ as a *population* object, rather than defining a unit-specific $Y_i(\cdot)$. This is without loss, since we define X to include all variables affecting Y .⁷ We thus allow for heterogeneous effects of any X_k , but the sources of heterogeneity are other variables in X . This emphasizes that the goal is understanding total variation in Y , which comes fully from variation in X . We make

⁷Similar notation appears in, e.g., Vytlačil (2002), Pearl (2009), and Hernán and Robins (2025). To see that this is without loss, adopt Vytlačil's (2002) notation: let $(\Omega, \mathcal{F}, \mathbb{P})$ be the probability space, and for each $x \in \mathcal{X}$, let $Y_x(\omega)$ be the random variable corresponding to the unit-specific potential outcome under x . Now redefine X to include any latent factors or indeed the unit index itself: $\tilde{X}(\omega) = (X(\omega), \omega)$. Then we can define the population-level $Y(\cdot)$ by $Y(\tilde{X}(\omega)) = Y_{X(\omega)}(\omega)$ for all ω .

the stable unit treatment value assumption (SUTVA) that unit j 's feature vector does not affect i 's outcome.

The potential outcome function $Y(\cdot)$ and the distribution of feature variables P_X together pin down the joint distribution of the outcome and feature variables $P_{Y,X}$, with CDF:

$$F_{Y,X}(y, x_1, \dots, x_K) = \int_{\mathcal{X}} \mathbb{1}_{\{Y(v_1, \dots, v_K) \leq y\}} \mathbb{1}_{\{v_1 \leq x_1, \dots, v_K \leq x_K\}} dP_X(v_1, \dots, v_K).$$

Below, all expectations and variances are with regard to $P_{Y,X}$ unless otherwise specified.

Features may be observed or unobserved. Write the *observed features* as $X^O := (X_1, \dots, X_O)$ for $O \leq K$, and the *unobserved features* as $X^U := (X_{O+1}, \dots, X_K)$.⁸ Denote by P_{Y,X^O} the marginal distribution of (Y, X^O) induced by $P_{Y,X}$. We assume that features are either observable *and* manipulable (X^O), or unobservable (X^U). In practice, researchers may have covariates: features which are observed, but not manipulated—either because they are inherently difficult to manipulate (*e.g.*, sex or race) or because they are not the experiment's focus. We discuss incorporating these covariates in Online Appendix D.

We assume that the outcome and features have finite second moments, and that the outcome has positive variance. For section 5 on identification, estimation, and inference, we assume finite fourth moments.

3.2 Predictive and causal models

We seek to understand the relation between the outcome and the observed features. Formally, we describe a relation as a **model**: a real-valued function of \mathcal{X}^O . We evaluate models using quadratic loss, though our approach can straightforwardly be extended to other loss functions.

Definition 1. The *risk* of a model M is $\mathcal{R}(M) := \mathbb{E}[(M(X^O) - Y(X))^2]$.

If we observe all features ($O = K$), a natural model is the potential outcome function $Y(\cdot)$ itself, which achieves zero risk; we call $Y(\cdot)$ the *true model*. At the opposite extreme, when no features are observed, the risk-minimizing model is $\mathbb{E}[Y]$, with risk $\text{Var}[Y]$; we call this the *baseline*

⁸We could model unobservables as at most one-dimensional without loss. We nonetheless treat unobservables as possibly multi-dimensional to stress that they are distinct causes.

model \bar{Y} . In the predominant case, when some, but not all, features are observed, the relevant model depends on the analyst’s goals. An analyst seeking to predict the outcome Y using the association between observed features and the outcome wants a predictive model. If instead the analyst seeks to causally explain the outcome from observed features, she uses a causal model. Below, we formally define these models.

Definition 2. Given an outcome Y and a vector of observed features X^O , the *best predictive model* $Y_{X^O}^P$ is the model which minimizes risk:

$$Y_{X^O}^P := \arg \min_M \mathcal{R}(M).$$

Equivalently, $Y_{X^O}^P(X^O) = \mathbb{E}[Y \mid X^O]$ almost surely.

This best predictive model encodes whether X^O predicts Y , but not whether X^O *causally explains* Y , as it conflates causal and non-causal sources of association between X^O and Y . In particular, if X^O is correlated with unobserved determinants of Y , then $Y_{X^O}^P$ can change with X^O even when X^O has no causal effect on Y . Therefore, if we want a *causal explanation*, we do not use a predictive model, but a model that uses only the causal relation between X^O and Y .

Formally, we define this as follows. For any x^O , recall that the *average potential outcome* is $\mathbb{E}[Y(x^O)] := \mathbb{E}_{P_{X^U}}[Y(x^O, X^U)]$. We say a model M is a *causal model* if, for any $x^O, \hat{x}^O \in \mathcal{X}^O$, we have $M(\hat{x}^O) - M(x^O) = \mathbb{E}[Y(\hat{x}^O)] - \mathbb{E}[Y(x^O)]$. A causal model is a model consistent with causal effects: for any two values of the observable features, x^O, \hat{x}^O , the difference in output from the causal model, $M(\hat{x}^O) - M(x^O)$ is exactly the expected increase in the outcome caused by manipulating X^O from x^O to \hat{x}^O , *i.e.* the average treatment effect. This definition implies that all causal models are equal to the average potential outcome plus a constant since any difference between model values at two different realisations of X^O is pinned down by the causal effects.

Among these models, we say the best causal model in explaining the outcome in the population is the one which chooses this constant as to minimise risk.⁹

⁹Alternatively, one could set the constant to match the expected average potential outcome. This approach coincides with ours when the potential outcome function is separable: $Y(x^O, x^U) = h(x^O) + g(x^U)$. We do not, in general, pursue this approach, for three reasons. First, the expected value of the best causal model would generally differ from the expected outcome, and hence the best causal model would be poorly calibrated as a model of the observed outcomes. Second, in some settings, researchers can identify causal effects but not average potential outcomes. Third, it would be harder

Definition 3. Given an outcome Y and a vector of observed features X^O , the *best causal model* $Y_{X^O}^C$ is the causal model which minimizes risk:

$$Y_{X^O}^C := \arg \min_M \mathcal{R}(M), \text{ s.t } M \text{ is a causal model.}$$

Equivalently, the best causal model is almost surely the average potential outcome function, re-centered to the expected value of the outcome, as the following lemma shows.

Lemma 1. *Every causal model M is the average potential outcome plus a constant: $M(x^O) = \mathbb{E}[Y(x^O)] + c$, for some constant $c \in \mathbb{R}$. The best causal model chooses $c^* = (\mathbb{E}[Y] - \mathbb{E}[\mathbb{E}[Y(x^O)]])$ such that the model is re-centred to the expected value of the outcome:*

$$Y_{X^O}^C(x^O) = \mathbb{E}[Y(x^O)] + \left(\mathbb{E}[Y] - \mathbb{E}[\mathbb{E}[Y(x^O)]] \right).$$

The best causal model solves the same risk-minimisation problem as the best predictive model, but under the additional constraint that differences in model outputs must represent causal effects. The best predictive model uses the association between X and Y in the population to minimise risk. However, differences between predicted Y at different X are due to both causal effects and omitted variable bias from unobservables. Causal models purge any such confounding. In that way, causal and predictive models encapsulate how Y changes with X^O , depending on whether we consider the change in Y associated with a change in X^O , or caused by a change in X^O .

3.3 Fit of the best causal model: the CR^2

We now define goodness-of-fit.

Definition 4. The *fit* of a model M is its proportional reduction in risk relative to the baseline model:

$$G(M) = \frac{\mathcal{R}(\bar{Y}) - \mathcal{R}(M)}{\mathcal{R}(\bar{Y})} = \frac{\text{Var}[Y] - \mathcal{R}(M)}{\text{Var}[Y]}.$$

The fit of the baseline model is 0; the fit of the true model is 1. The fit of the best predictive model is the familiar non-parametric predictive

to read the fit of such a model as the share of causally explained variance, which is our goal. In particular, our measure equals the reduction in variance from eliminating the average effect of X^O , which is not true in the alternative.

$$R^2: R^2(X^O) := G(Y_{X^O}^P).$$

Definition 5. Given outcome Y and observed features X^O , the *non-parametric causal* R^2 for the relationship between Y and X^O is the fit of the best causal model for Y :

$$CR^2(X^O) := G(Y_{X^O}^C).$$

Interpreting CR^2 as the share of variance causally explained. In our view, the causal R^2 measures the share of variance in an outcome Y causally explained by the features X^O . We show three different perspectives that lead to this interpretation of CR^2 .

First, this view mirrors the standard view of the predictive R^2 . The predictive R^2 , which is interpreted as the share of variance in Y predicted by X^O , is the complement of the fraction of variance unexplained (FVU): $R^2 = 1 - \frac{\text{Var}[Y - Y_{X^O}^P]}{\text{Var}[Y]}$. Intuitively, for each realization (y_i, x_i^O) , the best predictive model $Y_{X^O}^P$ leaves a residual $y_i - Y_{X^O}^P(x_i^O)$; the R^2 compares the variance of these residuals with the variance of the outcome. Similarly, the causal R^2 is $1 - \frac{\text{Var}[Y - Y_{X^O}^C]}{\text{Var}[Y]}$, that is, the complement of the fraction of variance *causally* unexplained. For each realization (y_i, x_i^O) , the causal effects of x_i^O leave a residual $y_i - Y_{X^O}^C(x_i^O)$ unexplained; CR^2 compares the variance of these residuals with the variance of the outcome. We therefore interpret it as *causally* explained variation. It is the share of variation that observables can predict after purging any confounding.

Another way to see this is to first decompose the outcome Y as

$$Y = Y' + (Y_{X^O}^C(X^O) - \mathbb{E}[Y]),$$

for $Y' := Y - (Y_{X^O}^C(x^O) - \mathbb{E}[Y])$. The term $(Y_{X^O}^C(x^O) - \mathbb{E}[Y])$ reflects how the observed features X^O cause Y to deviate from its mean. Y' thus represents the residual effect of other features: the part not caused by X^O . To view it this way, note that, in expectation, Y' does not vary as we manipulate X^O . The causal R^2 is then the percentage reduction in variance when moving from Y to Y' : that is, the reduction in the outcome's variance when we purge it of the causal effects of the observed features.

Second, CR^2 measures the proportional reduction in outcome variance in a thought experiment that eliminates the average causal effect of X^O on Y . Formally, we compare the observed variance of Y to the variance under a counterfactual in which the average causal effect of X^O has

been removed from Y —*i.e.*, for any x^O, \hat{x}^O , we have $\mathbb{E}[Y(x^O)] - \mathbb{E}[Y(\hat{x}^O)] = 0$ —holding all else constant. In our view, this reduction in variance is the variance that X^O causally explains. It is variance in Y that does not exist but for the causal effects of X^O .

One way to make this thought experiment concrete is to consider a government policy that removes any differences in outcomes arising from the effect: that restores individuals' outcomes to what they would have been without the average causal effect of X . That is, for an individual i with features x_i^O and pre-compensation outcome y_i , the post-compensation outcome is $\hat{y}_i' = y_i - Y_{X^O}^C(x_i^O)$.¹⁰ The ensuing outcome is $Y' = Y - Y_{X^O}^C$, with variance $\text{Var}[Y - Y_{X^O}^C] = \mathbb{E}[(Y - Y_{X^O}^C)^2] - \mathbb{E}[Y - Y_{X^O}^C]^2 = \mathcal{R}(Y_{X^O}^C)$.¹¹ The variance of the outcome when differences in the outcome due to causal effects are eliminated, or there were no causal effects to begin with, is $\mathcal{R}(Y_{X^O}^C)$. So, the CR^2 yields the proportional reduction in variance from eliminating the average causal effect of X^O .

Third, CR^2 can be interpreted as the improvement in model accuracy from knowing the causal effects of X^O . Recall that the predictive R^2 for the relation between Y and X^O can be seen as the percentage reduction in mean squared error from predicting a unit i 's outcome y_i given (i) the realized value of x_i^O and (ii) the joint distribution (Y, X^O) , relative to the constant predictor \bar{Y} . Analogously for the CR^2 , consider an analyst who models unit i 's outcome given (i) the realized value of x_i^O and (ii) the causal effects of X^O on Y , but *not* the joint distribution (Y, X^O) . That is, the analyst is asked to model y_i as best possible using the causal effects only, *without* exploiting correlations from the observational joint distribution—*i.e.* seek a causal rather than predictive explanation. The analyst's best model is $Y_{X^O}^C$ and the CR^2 is the percentage improvement in model accuracy when constrained to use purely causal variation in Y driven by X^O .

3.4 Parametric CR_{lin}^2

In practice, researchers often estimate parametric models: for instance, the standard linear R^2 measures fit of a linear predictive model. We can analogously define a parametric CR^2 when estimating the best causal

¹⁰We could remove these differences in other ways since the outcome levels may be changed while holding gaps between people fixed. For the thought experiment, all those ways are equivalent as such level changes preserve variance.

¹¹The last equality follows because the constant in the best causal model maximises fit: thus, the mean residual is zero, and the second term drops out.

model in a restricted function class \mathcal{F} , which may be motivated either by prior knowledge about the relation between Y and X^O , or by a need for tractability. We assume \mathcal{F} includes at least all constant models, and that it contains two models M and M' which differ, in the sense that $\mathbb{E}[(M(X_i^O) - M'(X_i^O))^2] > 0$. We also assume \mathcal{F} is closed under addition of constants, in the sense that, if \mathcal{F} includes a model M , it also includes $M + \alpha$ for any constant α .

Given a function class \mathcal{F} , the *average potential outcome function under \mathcal{F}* , $\mathbb{E}_{\mathcal{F}}[Y(\cdot)]$, is the best approximation within \mathcal{F} to the average potential outcome $\mathbb{E}[Y(\cdot)]$; that is,

$$\mathbb{E}_{\mathcal{F}}[Y(\cdot)] := \arg \min_{M \in \mathcal{F}} \mathbb{E}[\mathbb{E}[(M(X^O) - Y(X^O))^2]].$$

For instance, if the function class \mathcal{F} is the class of linear models, the average potential outcome function under \mathcal{F} corresponds to the fitted values from a regression of the outcome on the observed features in an experiment in which treatment is randomly assigned according to the marginal distribution of observables in the population. A model M is a *causal model under \mathcal{F}* if $M(x^O) - M(\hat{x}^O) = \mathbb{E}_{\mathcal{F}}[Y(x^O)] - \mathbb{E}_{\mathcal{F}}[Y(\hat{x}^O)]$. The *best causal model under \mathcal{F}* , $Y_{\mathcal{F},X^O}^C$, is

$$Y_{\mathcal{F},X^O}^C := \arg \min_{M \in \mathcal{F}} \mathcal{R}(M), \text{ such that } M \text{ is a causal model under } \mathcal{F}.$$

Similarly, we define the *best predictive model under \mathcal{F}* , $Y_{\mathcal{F},X^O}^P$, as

$$Y_{\mathcal{F},X^O}^P := \arg \min_{M \in \mathcal{F}} \mathcal{R}(M).$$

We assume throughout that these best predictive and causal models are unique. We call \mathcal{F} *well-specified* if it includes the best causal model $Y_{X^O}^C$.¹² Otherwise, \mathcal{F} is *misspecified*.¹³

The standard predictive R^2 is the goodness-of-fit of the best predictive

¹²The function class \mathcal{F} is well-specified if it includes the best causal model $Y_{X^O}^C$ given the observable features, even if that best causal model omits the effects of some unobservables. For instance, say X^O is observed, but its effects on Y varies with an unobserved X^U . If the average potential outcome given X^O is linear, the class of functions $X^O \rightarrow \alpha + \beta X^O$ is well-specified, even though β averages over heterogeneity induced by X^U .

¹³The (unrestricted) best causal model was defined pointwise, and so did not depend on the marginal distribution of X^O ; this remains true if \mathcal{F} is well-specified. Otherwise, the marginal distribution of X^O , used in the outer expectation, will affect the best model. The best model might then be seen as an approximation of the unrestricted best causal model.

model under \mathcal{F} : $R_{\mathcal{F}}^2(X^O) := G(Y_{\mathcal{F},X^O}^P)$. We define $CR_{\mathcal{F}}^2$ as the goodness-of-fit of the best causal model under \mathcal{F} : $CR_{\mathcal{F}}^2(X^O) := G(Y_{\mathcal{F},X^O}^C)$. The non-parametric CR^2 is nested as \mathcal{F} unrestricted.

When \mathcal{F} is the class of linear functions, the best predictive model is the linear projection of Y on X^O ; the best causal model is the (re-centred) linear projection of Y on X^O when X^O is randomly assigned in accordance with its population marginal distribution. Define the *linear causal* R^2 , $CR_{\text{lin}}^2(X^O)$, as the fit of this model. Applying our definition of goodness-of-fit, it follows that, in the well-specified case, this linear causal R^2 simply replaces the observational OLS coefficients $(\alpha^{\text{OBS}}, \beta^{\text{OBS}})$ in the predictive R^2 with corresponding experimental coefficients $(\alpha^{\text{EXP}}, \beta^{\text{EXP}})$:

$$R_{\text{lin}}^2(X^O) = 1 - \frac{\mathbb{E}[(Y - \alpha^{\text{OBS}} - (\beta^{\text{OBS}})^{\top} X^O)^2]}{\text{Var}[Y]},$$

$$CR_{\text{lin}}^2(X^O) = 1 - \frac{\mathbb{E}[(Y - \alpha^{\text{EXP}} - (\beta^{\text{EXP}})^{\top} X^O)^2]}{\text{Var}[Y]}.$$

4 Properties of the causal R^2

Having defined the causal R^2 , we now discuss its properties.

4.1 Basic properties

We begin with some basic properties needed to interpret CR^2 as the share of variance causally explained. Where the outcome is unclear, we specify it by writing $CR_{\mathcal{F}}^2(X^O \rightarrow Y)$.

Proposition 1 (basic properties). *For any function class \mathcal{F} :*

- (i) *If X^O has no causal effect on the outcome (that is, $Y(x^O, x^U) = Y(x^{O'}, x^U)$ for all $(x^O, x^{O'}, x^U)$), then $CR_{\mathcal{F}}^2(X^O) = 0$.*
- (ii) *If \mathcal{F} is well-specified, and X^O fully determines the outcome (that is, $Y(x^O, x^U) = Y(x^O, x^{U'})$ for all $(x^O, x^U, x^{U'})$), then $CR_{\mathcal{F}}^2(X^O) = 1$.*
- (iii) *If X^O does not vary, then $CR_{\mathcal{F}}^2(X^O) = 0$.*
- (iv) *If \mathcal{F} is well-specified, and X^U does not vary, then $CR_{\mathcal{F}}^2(X^O) = 1$.*
- (v) *Define $Y' = Y + \varepsilon$, for ε unobserved mean-zero random noise independent of, and causally unaffected by, X . Then $|CR_{\mathcal{F}}^2(X^O \rightarrow Y)| \geq |CR_{\mathcal{F}}^2(X^O \rightarrow Y')|$.*
- (vi) *CR^2 is not symmetric.*

Parts (i)–(iv) are limiting cases. If X^O does not affect Y , or if X^O does not vary, it explains no variation, and hence $\text{CR}_{\mathcal{F}}^2(X^O) = 0$. This is also true if X^O has no *average* effect on Y , such that the average potential outcome function $\mathbb{E}[Y(x^O)]$ is constant in x^O . If instead X^U does not vary or X^O fully determines the outcome—such that no variation in Y remains after setting X^O —then $\text{CR}_{\mathcal{F}}^2(X^O) = 1$, if \mathcal{F} is well-specified.¹⁴

Part (v) is a comparative static: weakening the relation between the outcome and observed features by adding noise attenuates their explanatory power.

Part (vi) says that CR^2 is not symmetric. Let there be one observable feature. Y may explain a certain share of X^O , but X^O might explain a different share of Y . This is intuitive since causal relations are not symmetric: X^O may cause Y , but not *vice versa*.

We consider these basic properties necessary for a measure of causally explained variation. Some alternatives do not satisfy them. For instance, the predictive R^2 in general violates (i) and (vi). Online Appendix A discusses other seemingly intuitive measures, and shows they lack one or more of the basic properties.

The CR^2 is also invariant to some transformations. This is desirable: for instance, education should explain the same share of variation in wages, no matter whether wages are measured in dollars or cents, or whether education is measured in years or days.

Proposition 2 (invariance to transformations).

- (i) CR^2 is invariant to affine transformations of the outcome, and to strictly monotone transformations of the features.
- (ii) CR_{lin}^2 is invariant to affine transformations of the outcome and features.

4.2 Comparison of predictive and causal R^2

Next, we describe the relation between the causal and predictive R^2 .

Proposition 3 (relation between predictive and causal R^2). *For any function class \mathcal{F} :*

- (i) *The causal R^2 is bounded above by the predictive R^2 : $\text{CR}_{\mathcal{F}}^2(X^O) \leq R_{\mathcal{F}}^2(X^O)$.*
- (ii) *If observables are independent of unobservables ($X^O \perp\!\!\!\perp X^U$), the causal and predictive R^2 coincide: $\text{CR}_{\mathcal{F}}^2(X^O) = R_{\mathcal{F}}^2(X^O)$.*

¹⁴The “well-specified” assumption parallels the predictive R^2 : even if X fully predicts Y , the R^2 from a linear regression of Y on X is less than 1 if the predictive relation between Y and X is not linear.

Part (i) says the predictive R^2 weakly exceeds the causal R^2 . Intuitively, the best predictive model maximises fit; the best causal model must perform weakly worse, since any predictive power from confounding is purged. (Recall that the best causal model is the solution of the same risk-minimising problem that gives rise to the best predictive model, but chosen from a smaller set of candidate models.) This result has practical use: even without knowing the causal effect of a variable, we can conclude that it has little causal explanatory power from the fact that it has little predictive power.

Part (ii) says that the predictive and causal R^2 coincide when observables are independent of unobservables: in the absence of confounding, the best predictive model equals the best causal model, and hence they have the same fit. For the linear CR^2 , orthogonality suffices for this result.

4.3 Possibility of negative values

Unlike the predictive R^2 , CR^2 may be negative and non-monotonic.

Proposition 4 (possibility of negativity and non-monotonicity). *For any function class \mathcal{F} :*

- (i) $R_{\mathcal{F}}^2$ is bounded between 0 and 1, whereas $CR_{\mathcal{F}}^2$ is bounded above by 1, but may be negative.
- (ii) $R_{\mathcal{F}}^2$ increases as more features are observed, whereas $CR_{\mathcal{F}}^2$ may fall.

Part (i) says that CR^2 may be negative. This happens when the observables *suppress*, rather than create, variance in the outcome, in the sense that the outcome’s variance would be smaller after eliminating the (average) causal effect of the observables. Consider the introductory example. Suppose smaller classes causally increase test scores, but students with worse unobservables tend to sort into smaller classes—for instance, because of programs to assist struggling students—and this sorting is strong enough that class size and test scores are *positively* correlated. Then the best causal model, which predicts that class size and test scores are negatively related, fits the population data *worse* than the baseline constant model of no relation. Since CR^2 is the proportional reduction in risk relative to this baseline model, the CR^2 is negative.

A negative CR^2 indicates that observed features suppress variation in the outcome: there “should” be more variation in the outcome than we observe. A similar phenomenon arises in Kitagawa-Oaxaca-Blinder

decompositions, which define the share of a gap between groups (*e.g.*, the gap in mean wages among men *vs.* women) “explained” by an observable (*e.g.*, education) as the reduction in the gap after residualizing the outcome on observables in a pooled regression. An observable may explain a negative component of variation: for instance, college completion is positively correlated with wages, but women complete college at higher rates, so the share of the gender wage gap explained by college is negative (Blau and Kahn 2017). Intuitively, education *suppresses* the wage gap—it contributes negatively. A similar intuition applies to CR^2 .

The non-monotonicity of the CR^2 has the same intuition. When no features are observed, the CR^2 is zero, but it may be negative when some features are added.¹⁵

4.4 Properties of CR_{lin}^2

The linear CR^2 is of applied interest because researchers often use linear models. It has a simple expression in terms of summary statistics.

Proposition 5 (summary statistics expression in linear case). *If \mathcal{F}_{lin} is well-specified:*

$$CR_{\text{lin}}^2(X^O) = R_{\text{lin}}^2(X^O) - (\tilde{\beta}_{X^O}^P - \tilde{\beta}_{X^O}^C)^\top \rho_{X^O} (\tilde{\beta}_{X^O}^P - \tilde{\beta}_{X^O}^C),$$

where $\tilde{\beta}_{X^O}^P$ are the coefficients from an OLS regression of standardized Y on standardized X^O in observational data, $\tilde{\beta}_{X^O}^C$ is the corresponding vector from an OLS regression in an experiment that randomly assigns X^O (with the standardization again with respect to the observational mean and variance), and ρ_{X^O} is the correlation matrix of observed features in observational data.

With a single observed feature variable, this simplifies to $CR_{\text{lin}}^2(X^O) = R_{\text{lin}}^2(X^O) - (\tilde{\beta}_{X^O}^P - \tilde{\beta}_{X^O}^C)^2$. That is, the difference between the predictive and causal linear R^2 is the squared difference between the observational and experimental slopes for the relation between the outcome and the observed feature.

Hence, if the model is well-specified, CR_{lin}^2 can be computed from standard summary statistics. Often, a researcher can compute CR_{lin}^2 from published results, even without access to the underlying micro-data.

¹⁵Online Appendix E shows there are no other monotone measures of the share of variation causally explained. Non-monotonicity is a constitutive feature of any measure of share of variation causally explained.

5 Identification, estimation, and inference

We now turn to estimation from finite data. We begin by describing the available data, before discussing identification and estimation, and then turning to inference. Since estimation is mostly standard, we relegate simulations to Online Appendix C.

5.1 Data setting

The CR^2 is defined from the joint distribution of the outcome and observable features, and the potential outcome function, which pins down causal effects. The joint distribution is easy to obtain from observational data, but the causal effects are not identified from observational data without further assumptions. In practice, we expect identifying causal effects to be the main hurdle to computing CR^2 . A range of identification approaches may identify these effects. For simplicity, our baseline is identification via a randomized experiment. We discuss extensions to quasi-experiments in 5.5.

The analyst has access to two samples: an observational sample (O) and an experimental sample (E).¹⁶ Following Athey et al. (2025), we think of the data as a single sample of $N = N_O + N_E$ units, with N_O units in the observational sample and N_E units in the experimental sample. For each unit i , denote by $S_i \in \{O, E\}$ the sample to which i belongs.

Each unit i is characterized by (Y_i, X_i^O, X_i^U, S_i) , where $Y_i = Y(X_i^O, X_i^U)$. We do not observe X_i^U for any unit. The observational sample consists of N_O i.i.d. draws (Y_i, X_i^O) from the joint distribution of the outcome and observables, R_{Y, X^O} . This sample thus identifies R_{Y, X^O} , but not the causal effects of X^O . To identify these effects, the analyst relies on an experiment (Rubin 1978).

Definition 6. An *experiment* $E := (x_t^O, \Pr(x_t^O))_{t=1}^T$ consists of:

- (i) T *treatment arms*, where each arm $x_t^O \in \mathcal{X}^O$ is an assignment of observed features; and
- (ii) an *assignment mechanism* $(\Pr(x_t^O))_{t=1}^T$, where each $\Pr(x_t^O) \in (0, 1)$ is

¹⁶This is an example of a “data fusion” setting (Rässler 2012; Bareinboim and Pearl 2016), also called “auxiliary data” (Hellerstein and Imbens 1999; Chen, Hong, and Tarozzi 2008) or “data combination” (Ridder and Moffitt 2007; Pearl and Bareinboim 2022). Data fusion has been used to generalize causal effects (e.g., Colnet et al. 2024), improve precision (e.g., Rosenman et al. 2023), estimate causal effects *via* surrogates (e.g., Athey et al. 2025), and correct for biases in observational data (e.g., Kallus, Puli, and Shalit 2018; Athey, Chetty, and Imbens 2025).

the probability with which a unit in the experiment is assigned to treatment arm x_t^O , with $\sum_{t=1}^T \Pr(x_t^O) = 1$.

The analyst draws an i.i.d. sample of size N_E from $P_{Y,X}$ and independently assigns each unit to a treatment arm via the assignment mechanism. She sets the unit’s observed features according to their treatment arm, and observes the ensuing outcome. Given treatment x_t^O , the outcome $y_i = Y(x_t^O, X_i^U)$ depends on the random variable X_i^U and so is random; denote its distribution by $P_{Y|x_t^O}^E$.

Formally, the analyst draws N observations (Y_i, X_i^O, S_i) from a superpopulation in which, with probability $p \in [0, 1]$, the unit is drawn from the observational distribution, and with probability $(1 - p)$ from the experimental distribution. For $p = 1$, the sample is an *observational sample*; for $p = 0$, an *experimental sample*; for $p \in (0, 1)$, a *combination of observational and experimental samples*. Table 1 summarizes these two data sources and their advantages and drawbacks. The observational sample is drawn directly from the population, and hence its feature distribution matches the population’s, but the observed features may not be independent of unobservables, barring identification of causal effects. In the experimental sample, randomization ensures $X_i^O \perp\!\!\!\perp X_i^U$, allowing identification of causal effects. However, the feature and outcome distributions generally differ from the population distributions.

Table 1. Summary of observational and experimental samples

| | Sample S_i | Outcome Y_i observed? | First O features X_i^O observed? | Other features X_i^U observed? | Features, outcome match pop. dist.? | Random assignment? ($X_i^O \perp\!\!\!\perp X_i^U$) |
|-------------|--------------|----------------------------|--|--|--|---|
| Obs. sample | O | ✓ | ✓ | × | ✓ | × |
| Exp. sample | E | ✓ | ✓ | × | × | ✓ |

Notes: This table summarizes the observational and experimental samples. The first row describes the observational sample; the second row describes the experimental sample. The outcome and observable features are observed in both samples, and the unobservables in neither. The first difference between the two samples is that the feature and outcome distributions in the observational data matches the population distribution whereas it does not in the experiment. The second difference is that observables are independent of unobservables in the experiment but not generally in the observational data.

This data combination arises in several settings. First, the analyst may conduct an experiment on a given population, and separately draw an observational sample from that population, *e.g.*, an observational study informs a subsequent randomized intervention. This is common: Epanomeritakis and Viviano (2025) report that over 30% of experimen-

tal papers in journals of the American Economic Association from 2015 to 2025 also include observational evidence. A second possibility is that the experiment includes a “status quo” arm, which may be treated as observational data. Finally, in some cases a researcher has observational data with as-if random variation in the features in a subsample; this subsample can be used as the experimental data, and the broader sample can be used as the observational data.

5.2 Identifying CR^2

We now turn to identifying CR^2 . An experiment is **full-rank** if $(1, x_t^O)_{t=1}^T$ has full rank, and **full-support** if every $x^O \in \text{supp } P_{X^O}$ appears as some treatment arm x_t^O .

Proposition 6 (identification). *Fix an outcome Y and observables X^O .*

- (i) *For any \mathcal{F} , $CR_{\mathcal{F}}^2(X^O)$ is not identified by an observational or experimental sample alone.*
- (ii) *The non-parametric causal R^2 , $CR^2(X^O)$, is identified by a combination of observational and experimental samples if and only if the experiment is full-support.*
- (iii) *Under a well-specified linear model, the linear causal R^2 , $CR_{\text{lin}}^2(X^O)$, is identified by a combination of observational and experimental samples if and only if the experiment is full-rank.*

Part (i) is intuitive given our discussion of the advantages and drawbacks of each sample: identifying the best causal model requires an experimental dataset; evaluating its goodness-of-fit requires an observational dataset. In consequence, either dataset alone is not enough.

The analyst can do better with the combined data. Part (ii) says that, with a full-support experiment, she can identify the non-parametric CR^2 . Intuitively, without any structure on treatment effects, the analyst must manipulate over the entire support of X^O . This is plausible only if the support of X^O is discrete, and in practice only if the support is quite small. For instance, with a single, binary observed feature, any non-trivial experiment suffices; with several observed features, each with finite support, a factorial experiment is needed (see, *e.g.*, Mukerjee and Wu 2006).

Part (iii) is less stringent, requiring only that the experiment independently manipulates each feature. If the analyst has prior reason to expect the causal model to be linear, she can thus identify CR_{lin}^2 . On the

other hand, if she simply uses linearity as a convenient functional form, then she can compute CR_{lin}^2 as a linear approximation to CR^2 .

5.3 Estimation and inference

We now construct a plug-in estimator for the causal R^2 , when it is identified.

Definition 7. Say the analyst has a combination of observational and experimental samples. Given an outcome Y , observed features X^O , and function class \mathcal{F} , define the *plug-in estimator* $\widehat{\text{CR}}_{\mathcal{F}}^2(X^O)$ as:

$$\widehat{\text{CR}}_{\mathcal{F}}^2(X^O) := 1 - \frac{\widehat{\mathcal{R}}(\hat{Y}_{\mathcal{F},X^O}^C)}{\widehat{\text{Var}}[Y]}$$

where $\hat{Y}_{\mathcal{F},X^O}^C := \hat{m}^C + \hat{c}^*$,

$$\hat{m}^C := \arg \min_{M \in \mathcal{F}} \frac{1}{N_E} \sum_{i:s_i=E} (y_i - M(x_i^O))^2,$$

$$\bar{y}_O := \frac{1}{N_O} \sum_{i:s_i=O} y_i,$$

$$\hat{c}^* := \bar{y}_O - \frac{1}{N_O} \sum_{i:s_i=O} \hat{m}^C(x_i^O),$$

$$\widehat{\mathcal{R}}(M) := \frac{1}{N_O} \sum_{i:s_i=O} (y_i - M(x_i^O))^2,$$

$$\widehat{\text{Var}}[Y] := \frac{1}{N_O} \sum_{i:s_i=O} (y_i - \bar{y}_O)^2.$$

The plug-in estimator replaces each quantity in the definition of $\text{CR}_{\mathcal{F}}^2$ with its finite-sample counterpart.¹⁷ For instance, suppose \mathcal{F} is the space of linear models. In that case, estimating the causal R^2 boils down to (1) estimating the best linear approximation to the potential outcome function, \hat{m}^C , which corresponds to the fitted values of the outcome in a linear regression of the outcome on the feature in experimental data; (2) re-centering this causal model; (3) evaluating the mean squared error from the re-centered model compared with the variance of the outcome.

Since each component of $\widehat{\text{CR}}_{\mathcal{F}}^2(X^O)$ converges to its population counterpart, Slutsky's Theorem shows Proposition 7.

¹⁷When the analyst's sample is a combination of observational and experimental data, $N_E > 0$ and $N_O > 0$ with probability one for N large, so the fractions are well-defined.

Proposition 7. *Suppose the conditions for identification in parts (ii) or (iii) of Proposition 6 are satisfied. Then the plug-in estimator is consistent for $\text{CR}_{\mathcal{F}}^2(X^O)$.*

Nonetheless, $\widehat{\text{CR}}_{\mathcal{F}}^2$ is generally downward-biased in small samples, since estimation error in the best causal model tends to inflate the model’s risk.¹⁸ In Online Appendix C, we present simulations in which this bias vanishes reasonably quickly as the sample size grows.

Inference for $\widehat{\text{CR}}_{\mathcal{F}}^2$ follows from the Delta method or the bootstrap. Since inference proceeds by mostly standard arguments, we sketch the main ideas here, and leave a detailed discussion to Online Appendix B. The Delta method approach is convenient for $\widehat{\text{CR}}_{\text{lin}}^2$, and begins by writing the plug-in estimator in terms of other sample quantities:

$$\widehat{\text{CR}}_{\text{lin}}^2(\hat{\theta}) = 1 - \frac{\hat{\theta}_1}{\hat{\theta}_2}, \text{ for } \hat{\theta} = \begin{pmatrix} \hat{\theta}_1 \\ \hat{\theta}_2 \end{pmatrix} := \begin{pmatrix} \widehat{\mathcal{R}}_{\text{lin}} \\ \widehat{\text{Var}}[Y] \end{pmatrix}$$

One can then show that $\hat{\theta}$ is asymptotically normal, derive its asymptotic covariance, and apply the Delta method to compute the asymptotic variance of $\widehat{\text{CR}}_{\text{lin}}^2$. Outside the linear case, it is easier to construct bootstrapped standard errors, randomly sampling with replacement from both the observational and experimental samples.

5.4 Measurement error

In practice, the analyst may measure the features or outcome with error. We show how measurement error affects the CR^2 . We focus on the linear CR^2 and a single observed feature, and consider *classical measurement error*: for $Z \in \{X^O, Y\}$, the observed value is $Z' = Z + \varepsilon$ with ε mean zero and uncorrelated with (X^O, Y) . There are four cases to consider, depending on whether noise appears in the outcome or feature, and whether it arises in the experimental or observational data.¹⁹

Proposition 8 (measurement error). *Say there is a single observable feature. Denote by CR_{CME}^2 the linear CR^2 when there is classical measurement error (CME), and by CR^2 the true value.*

¹⁸In contrast, the predictive R^2 is generally upward-biased due to overfitting.

¹⁹It may be surprising to consider measurement error in the experimentally-assigned feature. This would occur, *e.g.*, if treatment status is sometimes recorded incorrectly.

- (i) CME in the outcome in the observational data attenuates the CR^2 : $CR_{CME}^2 = \frac{\text{Var}[Y]}{\text{Var}[Y] + \text{Var}[\varepsilon]} \times CR^2$.
- (ii) CME in the feature in the observational data reduces CR^2 : $CR_{CME}^2 = CR^2 - \beta^2 \frac{\text{Var}[\varepsilon]}{\text{Var}[Y]}$, where β is the experimental regression coefficient on X^O .
- (iii) CME in the outcome in the experiment does not affect CR^2 : $CR_{CME}^2 = CR^2$.
- (iv) CME in the feature in the experiment can increase or reduce the causal R^2 , though it is still bounded above by the R^2 in observational data.

Proposition 8 can be used to sign the bias from measurement error. It also motivates correcting CR^2 for measurement error, under further assumptions. For instance, consider case (i), and suppose the analyst observes two noisy measurements of each unit’s outcome, $Y'_{i,1} = Y_i + \varepsilon_{i,1}$ and $Y'_{i,2} = Y_i + \varepsilon_{i,2}$, where $(\varepsilon_{i,1}, \varepsilon_{i,2})$ are mean-zero noise terms, independent of one another, the outcome, and the features. Then, it is well-known (Spearman 1904; Griliches 1974) that the “raw predictive R^2 ” (R_{raw}^2) between Y' and X^O is attenuated relative to the “signal predictive R^2 ” (R_{signal}^2) between Y and X^O , and that the analyst can recover R_{signal}^2 by dividing R_{raw}^2 by the reliability $r := \text{Corr}[Y'_{i,1}, Y'_{i,2}]$.²⁰ By similar logic, in the well-specified case, $CR_{\text{signal}}^2 = CR_{\text{raw}}^2/r$; hence, an analyst with multiple outcome measures may correct for measurement error. We do this in our application to Kremer et al. (2011) in section 6.1.

5.5 External validity

We have assumed the experiment randomly samples from the population. Some randomized controlled trials satisfy this condition, but others do not: they are conducted on subpopulations that differ from the general population. Similarly, some quasi-experimental designs identify the causal model only among a sub-population. For instance, in a standard instrumental variables design with a binary treatment, the instrument shifts treatment only among compliers, identifying a local average treatment effect (LATE) rather than the average treatment effect (ATE) (Imbens and Angrist 1994; Angrist, Imbens, and Rubin 1996).

Identifying CR^2 in this setting thus requires a transportability assumption: the causal effects in the experiment must coincide with those in the population. This generalization problem is familiar: it arises

²⁰This follows from noting that $r = \text{Var}[Y]/\text{Var}[Y']$, and $R_{\text{raw}}^2 = \frac{\text{Var}[Y'] - \mathbb{E}[(Y' - Y^P(X^O))^2]}{\text{Var}[Y']} = R_{\text{signal}}^2 \frac{\text{Var}[Y]}{\text{Var}[Y']}$.

whenever an analyst estimates a LATE or treatment on treated effect (“TOT”) but seeks to interpret it as the ATE. One approach is to argue that the experimental and target populations are similar, perhaps by comparing the observable characteristics of the two. In a randomized experiment, the experimental subpopulation is observed directly; in instrumental variables designs, Abadie’s (2003) weighting theorem identifies the distribution of complier covariates. If these distributions are similar, we may have more confidence in generalizing the causal model. If they differ, the analyst can compute covariate-specific LATEs and reweight them to recover the ATE, under the assumption that effect heterogeneity is fully captured by observables (Angrist and Fernández-Val 2013; Hartman et al. 2015). Finally, sometimes, structural latent-index models can extrapolate causal effects beyond compliers (*e.g.*, Heckman, Tobias, and Vytlačil 2001, 2003; Heckman 2010; Angrist 2004).

6 Applications of the CR²

We illustrate our measure in four settings, summarized in Table 2. We describe the main results here, leaving the details to Online Appendix F.

Table 2. Summary of applications

| Population | Outcome (Y) | Observed cause of interest (X^O) |
|---|-----------------|--------------------------------------|
| 1. Springs in Western Kenya (Kremer et al., 2011) | E Coli level | Spring protection |
| 2. Elementary school students in Tennessee (STAR) | Test scores | Class size |
| 3. Former colonies (Acemoglu et al., 2001) | GDP per capita | Expropriation risk |
| 4. Adults in the U.S. and U.K. (DASH, NDNS) | Blood pressure | Sodium intake |

Notes: This table summarizes the four applications described in this section.

6.1 Application 1: Share of variation in spring water quality explained by spring protection

Our first application closely resembles our baseline data setting. Kremer et al. (2011) conduct a randomized controlled trial in Kenya’s Western Province evaluating the effects of spring protection on water quality. The authors define “spring protection” as “seal[ing] off the source of a naturally occurring spring and encas[ing] it in concrete so that water

flows out from a pipe rather than seeping from the ground” (p. 149). They measure water quality primarily by the level of *Escherichia coli* bacteria. We ask what share of variation in water quality across Kenyan springs is causally explained by variation in spring protection.

The experimental sample consists of springs in the authors’ experiment. The observational sample consists of nearby springs not involved in the experiment. Protection status is binary. We assess the predictive relation between water quality and spring protection through an OLS regression in the observational data of the form

$$(3) \quad \ln E \text{ Coli}_s = \alpha^P + \beta^P \text{Protection}_s + \epsilon_s^P,$$

where $\ln E \text{ Coli}_s$ is the natural logarithm of the *E. coli* level in spring s , and Protection_s is an indicator for spring s being protected. The estimated coefficients from this regression, $(\hat{\alpha}^P, \hat{\beta}^P)$, define our estimated best predictive model, $\ln \widehat{E \text{ Coli}}_s^P(\text{Protection}_s) = \hat{\alpha}^P + \hat{\beta}^P \text{Protection}_s$. We assess the causal relation through a corresponding OLS regression in the experimental data:

$$(4) \quad \ln E \text{ Coli}_s = \alpha^C + \beta^C \text{Protection}_s + \epsilon_s^C.$$

The estimated causal effect of spring protection determines our best causal model,

$$(5) \quad \ln \widehat{E \text{ Coli}}_s^C(\text{Protection}_s) = \hat{c}^* + \beta^C \text{Protection}_s,$$

where \hat{c}^* is the re-centering constant such that the causal model is in expectation equal to the mean of the outcome. Panel (A) of Table 3 presents the estimated values $(\hat{\alpha}^P, \hat{\beta}^P)$ and $(\hat{c}^*, \hat{\beta}^C)$, which are reasonably similar; indeed, we cannot reject at the 95% level that $\hat{\beta}^C = \hat{\beta}^P$.

We then assess the goodness-of-fit of these models (Panels B–C). The best predictive model reduces mean squared error by around one-sixth. The best causal model reduces mean squared error by only a marginally smaller amount. In consequence the predictive and causal R^2 are very similar, as the first two bars in Panel A of Figure 2 show. In section 2, we said that it may be tempting to calculate the R^2 in the experiment as a measure of causally explained variation. We plot this measure here and find that it would understate by about 5 percentage points how much spring protection determines variation in water quality.

Kremer et al. (2011) document substantial error in measuring *E. coli*

levels. Proposition 8 suggests this measurement error will attenuate the share of variation explained. Under the assumption that the measurement error is independent of both spring protection and true water quality, we can correct for noise by dividing the share of variance explained by the test-retest reliability of *E. coli* measurements, which Kremer et al. (2011) estimate to be 0.46. The third and fourth bars in Figure 2(A) display this signal CR^2 , indicating that spring protection explains about one third of variation in water quality, both causally and predictively.

Table 3. Share of variance in E Coli causally explained by variance in spring protection

| | Obs. data (1) | Exp. data (2) |
|--|------------------|------------------|
| A. Best predictive and causal models | | |
| Constant | 4.82 (0.144) | 3.64 (0.089) |
| Spring protection | -1.98 (0.273) | -1.47 (0.158) |
| B. Outcome variance and mean squared error | | |
| Var[ln E Coli] | 4.88 | 4.46 |
| MSE[Best predictive model for ln E Coli] | 4.08 | - |
| MSE[Best causal model for ln E Coli] | 4.13 | 3.98 |
| C. Share of variance explained | | |
| % of var. predictively explained in pop. (obs. R^2) | 16.42% | - |
| % of var. causally explained in exp. (exp. R^2) | - | 10.71% |
| % of var. causally explained in pop. (CR^2) | 15.38% | - |

Notes: This table summarizes our analysis of variation in water quality (*E. Coli* level) causally explained by spring protection. Panel (A) reports estimates of equations (3) and (4), with corresponding standard errors. Panel (B) reports the mean squared error of these models. Panel (C) reports the implied share of variance explained.

We thus conclude that spring protection is a primary determinant of observed variation in water quality. We emphasize the difference between our question and the one in Kremer et al. (2011). Kremer et al. (2011) report the causal effect of spring protection on water quality, which speaks to the gains from hypothetically expanding protection. We instead ask what share of *observed* differences in water quality are explained by differences in spring protection, which speaks to its empirical importance as an explanation of differences in water quality across springs. A natural question is why people have more access to clean

water in some areas of the country than others. Our large estimated CR^2 indicates that spring protection is a main cause of these differences.

6.2 Application 2: Share of variation in test scores explained by class sizes

Our second application draws on the Tennessee Student-Teacher Achievement Ratio Experiment (“Project STAR”), which randomly assigned over 10,000 elementary school students to classes of different size. A rich literature has used STAR data to estimate causal effects of class size on test scores and later-life outcomes (Krueger 1999; Chetty et al. 2011; Dynarski, Hyman, and Schanzenbach 2013). We ask what share of variation in test scores is causally explained by class size.

Our experimental sample consists of STAR data made publicly available by Achilles et al. (Tennessee’s Student Teacher Achievement Ratio (STAR) project). Following prior literature, we assess the causal effect of class size in a two-stage least-squares regression of the form:

$$(6) \quad \text{ClassSize}_i = \pi^C + \rho^C \text{Small}_i + \nu_i^C,$$

$$(7) \quad \text{Score}_{i,s} = \alpha_s^C + \beta_s^C \widehat{\text{ClassSize}}_i + \varepsilon_{i,s}^C,$$

where $\text{Score}_{i,s}$ is student i ’s mean score in subject s (reading or math) over grades 1–3, ClassSize_i is mean class size in grades 1–3, and Small_i is an indicator for experimental assignment to a small class. The second-stage coefficient $\hat{\beta}^C$ is the causal effect of class size, which pins down the set of causal models. As discussed in section 5.5, this causal effect is among *compliers*, and so could differ from the causal effect in the population. In practice, we find very high compliance in the STAR setting, consistent with prior work (Krueger 1999). As a result, the compliers are very similar to the broader population on observables (Appendix Table 1).

Achilles et al. (Tennessee’s Student Teacher Achievement Ratio (STAR) project) also publish an observational sample, consisting of students in Tennessee schools that were matched to STAR schools, but did not participate in the experiment. We assess the predictive relation between students’ test scores and class size using the OLS regression:

$$(8) \quad \text{Score}_{i,s} = \alpha_s^P + \beta_s^P \text{ClassSize}_i + \varepsilon_{i,s}^P.$$

The resulting coefficients, $(\hat{\alpha}_s^P, \hat{\beta}_s^P)$, define our estimated best predictive

model. The predictive relation between class size and test scores is about five times as much as the causal effect of class size (Appendix Table 3), consistent with substantial omitted variable bias in (8).

Finally, we assess the goodness-of-fit of the best predictive and causal models. Panel B of Figure 2 shows that around 8% (5%) of variation in reading (math) scores is predicted by variation in class size, whereas only about 3% (2%) of variation is causally explained. For math scores, we cannot reject at the 95% level that the CR^2 is equal to zero; that is, that class size causally explains none of the variation in test scores. The causal and predictive R^2 differ due to the difference in the experimental and observational regression coefficients, *i.e.*, omitted variable bias.

We conclude that the predictive power of class size primarily reflects omitted variable bias, and that class size is not a major cause of test score variation, at least among the matched Tennessee schools. We note the distinction between this result and prior work. Existing literature has used the STAR experiment to estimate the causal effect of class size on outcomes. This effect is enough to answer the forward-looking policy question of how outcomes would change if we changed class sizes. The CR^2 instead indicates how much class size can explain *existing* differences in student outcomes. Researchers and policymakers spend much effort to understand what causes different test scores among students (Dynarski and Michelsmore 2017). The small causal CR^2 indicates that class size is not a major source of these differences.

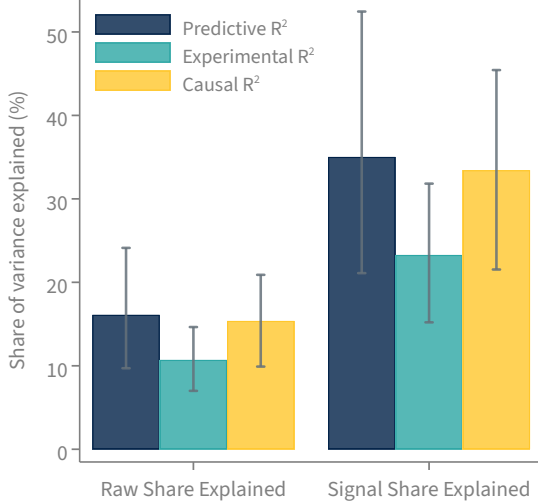
6.3 Application 3: Share of variation in national income explained by institutions

Our next application studies the determinants of cross-country income differences. Acemoglu, Johnson, and Robinson (2001) examine the effects of institutions on national income. Indeed, the authors explicitly set out to understand the sources of observed variation in incomes: the article’s opening sentence asks, “What are the fundamental causes of the large differences in income per capita across countries?” The authors measure institutional quality using an index of expropriation risk. Instrumenting for expropriation risk using the mortality rates of early European settlers, the authors show that expropriation risk has large effects on GDP per capita.²¹

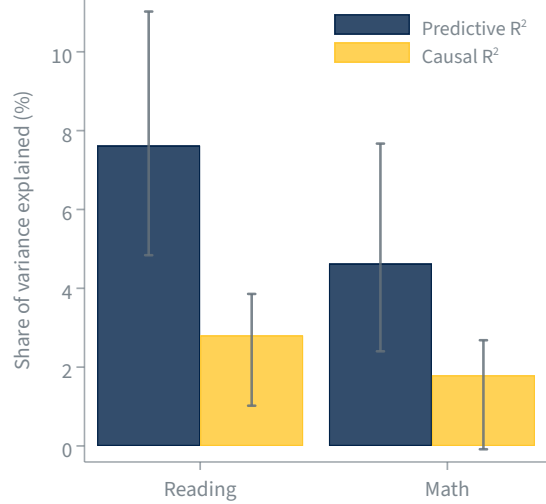
²¹This instrument has been controversial (McArthur and Sachs 2001; Glaeser et al. 2004; Albouy 2012; Conley and Kelly 2025). We take the instrument as given and

Figure 2. Causal R² in Applications

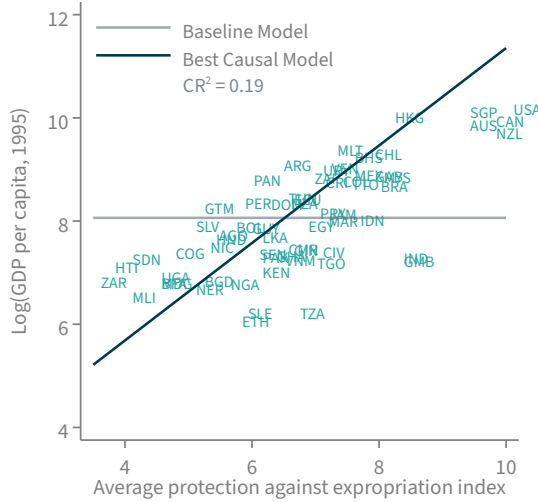
(A) Share of variance in water quality explained by spring protection



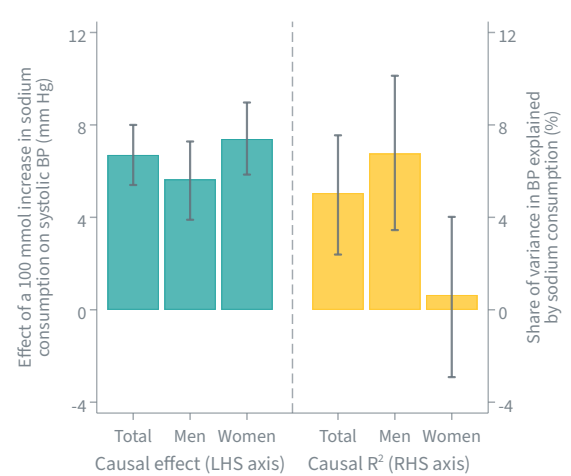
(B) Share of variance in test scores explained by class size



(C) Share of variance in national income causally explained by expropriation risk



(D) Share of variance in blood pressure causally explained by salt intake



Notes: This figure shows the results in our applications. Panel (A) reports the share of variation in water quality in Kenyan springs causally explained by spring protection; Panel (B) test scores by class size; Panel (C) national income by institutions (expropriation risk); and Panel (D) systolic blood pressure by salt intake, along with the causal effect of salt on blood pressure. Panels (A), (B), and (D) show 95% confidence intervals computed using 2,000 bootstrapped samples.

We build on this analysis by asking what share of variation in national income is causally explained by differences in expropriation risk. Acemoglu, Johnson, and Robinson (2001) estimate the causal effect of expropriation risk on national incomes in a two-stage least-squares regression:

$$(9) \quad \text{ExpropriationRisk}_c = \pi^C + \rho^C \text{SettlerMortality}_c + \nu_c^C,$$

$$(10) \quad \ln \text{GDP}_c = \alpha^C + \beta^C \widehat{\text{ExpropriationRisk}}_c + \varepsilon_c^C,$$

where $\text{ExpropriationRisk}_c$ is the measured expropriation risk index for country c , $\text{SettlerMortality}_c$ is the rate of early European settler mortality in c , and $\ln \text{GDP}_c$ is the natural logarithm of GDP per capita in 1995 in country c . Our estimated best causal model is then $\text{GDP}_c(\text{ExpropriationRisk}_c) = \hat{c}^* + \hat{\beta}^C \text{ExpropriationRisk}_c$. To use the treatment effect among compliers, we make an external validity assumption as described in section 5.5 and assess its plausibility in Online Appendix F.

We then evaluate the best causal model's goodness-of-fit in the observational data of Acemoglu, Johnson, and Robinson (2001). Panel C of Figure 2 shows a scatterplot of countries' log GDP per capita vs. their average expropriation risk. The navy line shows the best causal model. Its fit gives a causal R^2 of 0.19: that is, the results in Acemoglu, Johnson, and Robinson (2001) suggest that around one-fifth of variation in national income is causally explained by cross-country differences in institutions' expropriation risk. As a consequence of the small sample size, the estimated causal R^2 is very noisy: even at the 10% level, we cannot reject that expropriation risk explains none of the variation in income. Note that this best causal model differs from the line of best fit: the R^2 from a regression of log GDP per capita on expropriation risk is around 0.54. That is, although institutions are a main cause of differences in national income, most of their predictive power comes from omitted variables.

In our view, the causal R^2 allows us to make significant progress towards answering the question from which Acemoglu, Johnson, and Robinson (2001) start: are differences in institutions one of the fundamental causes of differences in national income? The authors find a large causal effect: reducing expropriation risk substantially increases GDP per capita. Nonetheless, this large causal effect is consistent with

examine only its implications for the share of variance in national incomes that is causally explained by institutions.

either a small, or a large, role for institutions in producing *observed* variation: for instance, if there is very limited variation in institutions, then institutions may explain little despite their large causal effects. Although our estimated CR^2 is noisy, the point estimate suggests that institutions are a substantial cause of observed income dispersion: they explain about one fifth of cross-country differences.

Lastly, we may ask not only whether institutions are a large cause of differences in national income, but whether they are a large cause *vis-à-vis* other factors, such as culture (Guiso, Sapienza, and Zingales 2006; Tabellini 2010), geography (Diamond 1997; Nunn and Puga 2012), or disease (Gallup and Sachs 2001; Bleakley 2010). We illustrate how the CR^2 can be used for this comparison with a back-of-the-envelope calculation of the power of childhood malaria in explaining cross-country income differences, which we detail in Online Appendix F. We use estimates of the causal effect of childhood malaria from Bleakley (2010). The best causal model implied by his findings explains 8.8% of income differences in the data of Acemoglu, Johnson, and Robinson (2001). This calculation makes several assumptions, including on the transportability of causal effects to other countries. This exercise illustrates that CR^2 can be used to compare the explanatory power of competing accounts of variation in an outcome. The point estimates suggest that malaria in childhood explains about half as much of the variation in cross-country income differences as institutions, though we urge caution as the estimates are quite noisy.

6.4 Application 4: Share of variation in blood pressure explained by sodium intake

Finally, we consider CR^2 in a stylised health setting. High blood pressure is a major cause of death in high-income countries and excess salt consumption is considered a leading cause of high blood pressure (Institute of Medicine 2010). We examine the share of variation in blood pressure causally explained by salt consumption.

Our experimental sample consists of participants in the DASH-Sodium experiment, a randomized controlled trial which evaluated the effects of salt intake on blood pressure. Sacks et al. (2001) use this experimental data to estimate an OLS regression of the form:

$$(11) \quad BP_i = \alpha^C + \beta^C \text{Sodium}_i + \epsilon_i^C,$$

where BP_i is person i 's systolic blood pressure. The estimated coefficient $\hat{\beta}^C$ yields our best causal model. The authors also run separate OLS regressions by gender, finding a somewhat larger effect of sodium intake on systolic blood pressure among women.

Evaluating the best causal model's goodness-of-fit requires population data on sodium intake and blood pressure. We use data from the UK National Diet and Nutrition Survey, a long-running, nationally-representative survey assessing the diets and health of British citizens which collected urinary sodium data for the period 2008–2012 (University Of Cambridge, MRC Epidemiology Unit and NatCen Social Research, National Diet and Nutrition Survey Years 1-11, 2008-2019). We follow a public health literature on sodium intake and blood pressure which essentially treats the U.S. and U.K. populations as transportable in this context (Scientific Advisory Committee on Nutrition 2003; Jones et al. 2020).

We find that salt consumption causally explains 5.1% of the variation in systolic blood pressure in the population, and 6.8% of that variation in men, but only 0.6% among women (p -value for difference < 0.01). This is striking because the causal effect of sodium on blood pressure is, if anything, slightly larger for women.

Sodium turns out to explain a much smaller share among women for three reasons. First, the variance of sodium is about 40% larger among men, and hence, given similar causal effects, it generates more variation in blood pressure for men. Second, the variance of blood pressure among men is about 20% smaller: there is less total variation to explain, and so the share of variation generated by sodium is larger among men. Third, among men, the causal and observational "effects" of sodium are similar: there is very little omitted variable bias. Among women, the causal effect is around 80% larger than the observational effect: sodium intake is *negatively* related to unobserved determinants of blood pressure. In consequence, sodium intake explains almost none of the observed variation among women, but a meaningful amount among men.

We thus conclude that sodium intake is a more important determinant of observed blood pressure among men than among women. This conclusion highlights the difference between the standard causal effects exercise—determining the effect of a change in sodium intake on blood pressure—and our exercise: measuring the share of *observed* blood pressure variation that is causally explained by sodium intake.

The causal effects question is relevant to a doctor advising patients with high salt intake; such a doctor should give reasonably similar advice to men and women, since the causal effects of sodium on blood pressure are similar. However, the question about causally explained variation is more relevant if the doctor is trying to ascertain *why* a patient has high blood pressure. High sodium intake is less likely to be the cause for women than for men. The share of observed blood pressure variation that is causally explained by sodium intake also informs research priorities: the low share, especially among women, suggests that sodium is not a primary source of variation in blood pressure and that other determinants warrant further study.

7 Conclusion

Social scientists are often interested not just in the causal effect of a variable on an outcome, but in how much of the outcome's variance the variable causally explains. Does the variable play a large or a small role in explaining observed differences? This question is related to the causal effect and to traditional goodness-of-fit measures, but not answered by them. We propose the causal R^2 (CR^2) as a measure to answer this question and quantify the importance of the variable in determining the outcome. The CR^2 takes the causal effects from an experiment and evaluates the fit of the implied best causal model in the population. We describe three perspectives that lead us to interpret the causal R^2 as the share of observed variance causally explained by a variable.

We conclude with some limitations and directions for future research. First, a line of work in statistics criticizes even the predictive R^2 . Part of this critique (*e.g.*, King 1986, 1991) is that the R^2 fails to capture causal relations; the CR^2 tackles this concern. Other objections are not resolved by our measure. For instance, Draper (1984) questions R^2 as a proportional goodness-of-fit measure; Healy (1984) objects to proportional measures more broadly.

Second, goodness-of-fit is only one desirable property of a causal model: we may also value parsimony, interpretability, robustness, or portability (Fudenberg et al. 2022). Nonetheless, the extent to which a model causally explains variation speaks to our understanding of the outcome, and to the value of future research.

We are excited about future work that formalises and assesses this trade-off between a causal model's explanatory power and other de-

sirable properties. In our view, empirical research that applies CR^2 to various settings, including using existing causal effect estimates, is also particularly valuable. How complete are our explanations for different questions, and which hypothesised causes have most explanatory power? Lastly, open econometric questions include, for example, how CR^2 may be estimated most efficiently in specific data and identification settings.

References

- Abadie, Alberto. 2003. "Semiparametric instrumental variable estimation of treatment response models." *Journal of Econometrics* 113, no. 2 (April): 231–263. ISSN: 0304-4076. [https://doi.org/10.1016/S0304-4076\(02\)00201-4](https://doi.org/10.1016/S0304-4076(02)00201-4).
- Acemoglu, Daron, Simon Johnson, and James A. Robinson. 2001. "The Colonial Origins of Comparative Development: An Empirical Investigation." *American Economic Review* 91, no. 5 (December): 1369–1401. ISSN: 0002-8282. <https://doi.org/10.1257/aer.91.5.1369>.
- Achilles, C. M., Helen Pate Bain, Fred Bellott, Jayne Boyd-Zaharias, Jeremy Finn, John Folger, John Johnston, and Elizabeth Word. 2008. (Tennessee's Student Teacher Achievement Ratio (STAR) project. V. 1. Dataverse: Project STAR (<https://dataverse.harvard.edu/dataverse/star>). UNF:3:Ji2Q+9HCCZ-Abw3csOdMNdA==. License: CCo 1.0. Accessed January 26, 2026). <https://doi.org/10.7910/DVN/SIWH9F>.
- Albouy, David Y. 2012. "The Colonial Origins of Comparative Development: An Empirical Investigation: Comment." *American Economic Review* 102, no. 6 (October): 3059–3076. ISSN: 0002-8282. <https://doi.org/10.1257/aer.102.6.3059>.
- Andrews, Isaiah, and Emily Oster. 2019. "A simple approximation for evaluating external validity bias." *Economics Letters* 178 (May): 58–62. ISSN: 0165-1765. <https://doi.org/10.1016/j.econlet.2019.02.020>.
- Angrist, Joshua D. 2004. "Treatment effect heterogeneity in theory and practice." *The economic journal* 114 (494): C52–C83.
- Angrist, Joshua D., and Iván Fernández-Val. 2013. "ExtrapoLATE-ing: External Validity and Overidentification in the LATE Framework." In *Advances in Economics and Econometrics: Tenth World Congress*, edited by Daron Acemoglu, Manuel Arellano, and Eddie Dekel, 401–434. Econometric Society Monographs. Cambridge University Press.
- Angrist, Joshua D., Guido Imbens, and Donald Rubin. 1996. "Identification of causal effects using instrumental variables." *Journal of the American statistical Association* 91 (434): 444–455.
- Apestequia, Jose, and Miguel A Ballester. 2021. "Separating predicted randomness from residual behavior." *Journal of the European Economic Association* 19 (2): 1041–1076.
- Athey, Susan, Raj Chetty, and Guido Imbens. 2025. *The Experimental Selection Correction Estimator: Using Experiments to Remove Biases in Observational Estimates*. Technical report. National Bureau of Economic Research.

- Athey, Susan, Raj Chetty, Guido Imbens, and Hyunseung Kang. 2025. “The Surrogate Index: Combining Short-Term Proxies to Estimate Long-Term Treatment Effects More Rapidly and Precisely.” Forthcoming, *Review of Economic Studies*.
- Athey, Susan, and Guido W. Imbens. 2017. “The Econometrics of Randomized Experiments.” In *Handbook of Field Experiments*, edited by Abhijit Vinayak Banerjee and Esther Duflo, 1:73–140. Handbook of Economic Field Experiments. Elsevier. <https://doi.org/10.1016/bs.hefe.2016.10.003>.
- Bareinboim, Elias, and Judea Pearl. 2016. “Causal inference and the data-fusion problem.” *Proceedings of the National Academy of Sciences* 113 (27): 7345–7352.
- Blau, Francine D, and Lawrence M Kahn. 2017. “The gender wage gap: Extent, trends, and explanations.” *Journal of Economic Literature* 55 (3): 789–865.
- Bleakley, Hoyt. 2010. “Malaria Eradication in the Americas: A Retrospective Analysis of Childhood Exposure.” *American Economic Journal: Applied Economics* 2, no. 2 (April): 1–45. ISSN: 1945-7782. <https://doi.org/10.1257/app.2.2.1>.
- Breza, Emily, and Arun G Chandrasekhar. 2019. “Social networks, reputation, and commitment: evidence from a savings monitors experiment.” *Econometrica* 87 (1): 175–216.
- Chen, Xiaohong, Han Hong, and Alessandro Tarozzi. 2008. “Semiparametric Efficiency in GMM Models with Auxiliary Data.” *The Annals of Statistics* 36 (2): 808–843. ISSN: 00905364.
- Chetty, Raj, John N Friedman, Nathaniel Hilger, Emmanuel Saez, Diane Whitmore Schanzenbach, and Danny Yagan. 2011. “How does your kindergarten classroom affect your earnings? Evidence from Project STAR.” *The Quarterly Journal of Economics* 126 (4): 1593–1660.
- Cinelli, Carlos, and Chad Hazlett. 2020. “Making sense of sensitivity: Extending omitted variable bias.” *Journal of the Royal Statistical Society Series B: Statistical Methodology* 82 (1): 39–67.
- . 2025. “An omitted variable bias framework for sensitivity analysis of instrumental variables.” *Biometrika* 112 (2): asaf004.
- Colnet, Bénédicte, Imke Mayer, Guanhua Chen, Awa Dieng, Ruohong Li, Gaël Varoquaux, Jean-Philippe Vert, Julie Josse, and Shu Yang. 2024. “Causal inference methods for combining randomized trials and observational studies: a review.” *Statistical Science* 39 (1): 165–191.
- Conley, Timothy G, and Morgan Kelly. 2025. “The standard errors of persistence.” *Journal of International Economics* 153:104027.
- Datta, Anupam, Shayak Sen, and Yair Zick. 2016. “Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems.” In *2016 IEEE symposium on security and privacy (SP)*, 598–617. IEEE.
- Dawid, A. Philip, and Monica Musio. 2022. “Effects of Causes and Causes of Effects.” First published as a Review in Advance on November 4, 2021, *Annual Review of Statistics and Its Application* 9:261–287. <https://doi.org/10.1146/annurev-statistics-070121-061120>.
- Diamond, Jared. 1997. *Guns, Germs, and Steel: The Fates of Human Societies*. Hardcover (cloth). Paperback ISBN: 0-393-31755-2. New York: W. W. Norton & Company. ISBN: 0-393-03891-2.

- Draper, Norman R. 1984. "The Box-Wetzel Criterion Versus R^2 ." *Journal of the Royal Statistical Society: Series A (General)* 147 (1): 100–103.
- Duflo, Esther, Rachel Glennerster, and Michael Kremer. 2007. "Using randomization in development economics research: A toolkit." In *Handbook of Development Economics*, 4:3895–3962. Elsevier.
- Dynarski, Susan, Joshua Hyman, and Diane Whitmore Schanzenbach. 2013. "Experimental evidence on the effect of childhood investments on postsecondary attainment and degree completion." *Journal of Policy Analysis and Management* 32 (4): 692–717.
- Dynarski, Susan M., and Katherine Micheltore. 2017. "The gap within the gap." Brookings Institution. Evidence Speaks, April 13, 2017. Accessed January 26, 2026. <https://www.brookings.edu/articles/the-gap-within-the-gap/>.
- Epanomeritakis, Aristotelis, and Davide Viviano. 2025. *Choosing What to Learn: Experimental Design when Combining Experimental with Observational Evidence*. arXiv: 2510.23434 [econ. EM].
- Fudenberg, Drew, Jon Kleinberg, Annie Liang, and Sendhil Mullainathan. 2022. "Measuring the completeness of economic models." *Journal of Political Economy* 130 (4): 956–990.
- Gallup, John Luke, and Jeffrey D. Sachs. 2001. "The Economic Burden of Malaria." *The American Journal of Tropical Medicine and Hygiene* 64 (1-2 Suppl): 85–96. ISSN: 0002-9637. <https://doi.org/10.4269/ajtmh.2001.64.85>.
- Gelman, Andrew, Ben Goodrich, Jonah Gabry, and Aki Vehtari. 2019. "R-squared for Bayesian Regression Models." *The American Statistician* 73 (3): 307–309. ISSN: 0003-1305. <https://doi.org/10.1080/00031305.2018.1549100>.
- Gelman, Andrew, and Guido Imbens. 2013. *Why ask why? Forward causal inference and reverse causal questions*. Technical report. National Bureau of Economic Research.
- Gelman, Andrew, and Iain Pardoe. 2006. "Bayesian measures of explained variance and pooling in multilevel (hierarchical) models." *Technometrics* 48 (2): 241–251.
- GiveWell. 2023. "Malaria income effect size, April 2023 (public)." Google Sheets spreadsheet; linked from GiveWell's Seasonal Malaria Chemoprevention page ("Long-term income increases") as the spreadsheet containing malaria-income calculations. April. Accessed January 24, 2026. https://docs.google.com/spreadsheets/d/1zeVV1nDmM6CnvPK36fRGpujYxIMvPJUHruHwNA_KnXQ/edit?gid=547345468.
- Glaeser, Edward L, Rafael La Porta, Florencio Lopez-de-Silanes, and Andrei Shleifer. 2004. "Do institutions cause growth?" *Journal of Economic Growth* 9 (3): 271–303.
- Glaeser, Edward L. 2011. *Triumph of the City: How Our Greatest Invention Makes Us Richer, Smarter, Greener, Healthier, and Happier*. 352. First edition. New York: Penguin Press, February. ISBN: 978-1-59420-277-3.
- Goldberger, Arthur S. 1979. "Heritability." *Economica* 46 (184): 327–347.
- Griliches, Zvi. 1974. "Errors in variables and other unobservables." *Econometrica: Journal of the Econometric Society*, 971–998.
- Guiso, Luigi, Paola Sapienza, and Luigi Zingales. 2006. "Does Culture Affect Economic Outcomes?" Spring issue, *Journal of Economic Perspectives* 20 (2): 23–48. ISSN: 0895-3309. <https://doi.org/10.1257/jep.20.2.23>.

- Halpern, Joseph Y., and Judea Pearl. 2005. "Causes and Explanations: A Structural-Model Approach. Part I: Causes." *The British Journal for the Philosophy of Science* 56, no. 4 (December 1, 2005): 843–887. ISSN: 0007-0882. <https://doi.org/10.1093/bjps/axi147>.
- Harrell, Frank E, Robert M Califf, David B Pryor, Kerry L Lee, and Robert A Rosati. 1982. "Evaluating the yield of medical tests." *Journal of the American Medical Association* 247 (18): 2543–2546.
- Hartman, Erin, Richard Grieve, Roland Ramsahai, and Jasjeet S Sekhon. 2015. "From sample average treatment effect to population average treatment effect on the treated: combining experimental with observational studies to estimate population treatment effects." *Journal of the Royal Statistical Society Series A: Statistics in Society* 178 (3): 757–778.
- Hawinkel, Stijn, Willem Waegeman, and Steven Maere. 2024. "Out-of-sample R²: estimation and inference." *The American Statistician* 78 (1): 15–25.
- Hay, Simon I., Carlos A. Guerra, Andrew J. Tatem, Abdisalan M. Noor, and Robert W. Snow. 2004. "The Global Distribution and Population at Risk of Malaria: Past, Present, and Future." *The Lancet Infectious Diseases* 4, no. 6 (June): 327–336. [https://doi.org/10.1016/S1473-3099\(04\)01043-6](https://doi.org/10.1016/S1473-3099(04)01043-6).
- Healy, M. J. R. 1984. "The Use of R² as a Measure of Goodness of Fit." *Royal Statistical Society. Journal. Series A: General* 147, no. 4 (December): 608–609. ISSN: 0035-9238. <https://doi.org/10.2307/2981848>. eprint: https://academic.oup.com/jrssa/article-pdf/147/4/608/49757327/jrssa_147_4_608.pdf.
- Heckman, James, Justin L Tobias, and Edward Vytlačil. 2001. "Four parameters of interest in the evaluation of social programs." *Southern Economic Journal* 68 (2): 210–223.
- . 2003. "Simple estimators for treatment parameters in a latent-variable framework." *Review of Economics and Statistics* 85 (3): 748–755.
- Heckman, James J. 2010. "Building bridges between structural and program evaluation approaches to evaluating policy." *Journal of Economic Literature* 48 (2): 356–398.
- Hellerstein, Judith, and Guido Imbens. 1999. "Imposing moment restrictions from auxiliary data by weighting." *Review of Economics and Statistics* 81 (1): 1–14.
- Hernán, Miguel A., and James M. Robins. 2025. *Causal Inference: What If*. Boca Raton: Chapman & Hall/CRC.
- Heskes, Tom, Evi Sijben, Ioan Gabriel Bucur, and Tom Claassen. 2020. "Causal Shapley values: Exploiting causal knowledge to explain individual predictions of complex models." *Advances in Neural Information Processing Systems* 33:4778–4789.
- Hull, Peter. 2025. "'One Weird Trick' to Characterize Effective Populations in Design-Based Specifications." Metrics note (3-page unpublished note). Listed on Peter Hull's "Metrics Notes" page. September 3, 2025. Accessed January 26, 2026. <https://www.dropbox.com/scl/fi/lgba6bh15jm72i3rxfvt1/EffectivePop.pdf>.
- Imbens, Guido, and Joshua D. Angrist. 1994. "Identification and Estimation of Local Average Treatment Effects." *Econometrica* 62 (2): 467–475. ISSN: 00129682, 14680262.

- Institute of Medicine. 2010. *A Population-Based Policy and Systems Change Approach to Prevent and Control Hypertension*. Washington, DC: The National Academies Press. ISBN: 978-0-309-14809-2. <https://doi.org/10.17226/12819>.
- Joint National Committee on Prevention Detection Evaluation and Treatment of High Blood Pressure. 1997. "The Sixth Report of the Joint National Committee on Prevention, Detection, Evaluation, and Treatment of High Blood Pressure." *Archives of Internal Medicine* 157, no. 21 (November): 2413. ISSN: 0003-9926. <https://doi.org/10.1001/archinte.1997.00440420033005>.
- Jones, Nicholas R, Terry McCormack, Margaret Constanti, and Richard J McManus. 2020. "Diagnosis and management of hypertension in adults: NICE guideline update 2019." *The British Journal of General Practice* 70 (691): 90.
- Jung, Yonghan, Shiva Kasiviswanathan, Jin Tian, Dominik Janzing, Patrick Blöbaum, and Elias Bareinboim. 2022. "On measuring causal contributions via do-interventions." In *International Conference on Machine Learning*, 10476–10501. PMLR.
- Kallus, Nathan, Aahlad Manas Puli, and Uri Shalit. 2018. "Removing hidden confounding by experimental grounding." In *Advances in neural information processing systems*, vol. 31.
- King, Gary. 1986. "How not to lie with statistics: Avoiding common mistakes in quantitative political science." *American Journal of Political Science*, 666–687.
- . 1991. "' Truth' Is Stranger than Prediction, More Questionable than Causal Inference." *American Journal of Political Science* 35 (4): 1047–1053.
- Kleven, Henrik Jacobsen, Martin B Knudsen, Claus Thustrup Kreiner, Søren Pedersen, and Emmanuel Saez. 2011. "Unwilling or unable to cheat? Evidence from a tax audit experiment in Denmark." *Econometrica* 79 (3): 651–692.
- Kremer, Michael, Jessica Leino, Edward Miguel, and Alix Peterson Zwane. 2011. "Spring cleaning: Rural water impacts, valuation, and property rights institutions." *The Quarterly Journal of Economics* 126 (1): 145–205.
- Krueger, Alan B. 1999. "Experimental estimates of education production functions." *The Quarterly Journal of Economics* 114 (2): 497–532.
- Li, Gang, and Xiaoyan Wang. 2019. "Prediction Accuracy Measures for a Non-linear Model and for Right-Censored Time-to-Event Data." First published online 2019-03-11. *Journal of the American Statistical Association* 114 (528): 1815–1825. ISSN: 0162-1459. <https://doi.org/10.1080/01621459.2018.1515079>.
- Malaria Atlas Project. 2026. "Malaria Atlas Project Data Platform." Accessed January 24, 2026. <https://data.malariaatlas.org/>.
- Maxcy, Kenneth F. 1923. "The Distribution of Malaria in the United States as Indicated by Mortality Reports." *Public Health Reports* 38, no. 21 (May 25, 1923): 1125–1138.
- McArthur, John W., and Jeffrey D. Sachs. 2001. *Institutions and Geography: Comment on Acemoglu, Johnson and Robinson (2000)*. NBER Working Paper 8114. National Bureau of Economic Research, February. <https://doi.org/10.3386/w8114>.
- McFadden, Daniel. 1973. "Conditional Logit Analysis of Qualitative Choice Behavior." In *Frontiers in Econometrics*, edited by Paul Zarembka. New York: Wiley.

- Miguel, Edward, and Michael Kremer. 2004. "Worms: identifying impacts on education and health in the presence of treatment externalities." *Econometrica* 72 (1): 159–217.
- Mincer, Jacob. 1958. "Investment in Human Capital and Personal Income Distribution." *Journal of Political Economy* 66, no. 4 (August): 281–302. ISSN: 0022-3808. <https://doi.org/10.1086/258055>.
- Mukerjee, Rahul, and C. F. Jeff Wu. 2006. *A Modern Theory of Factorial Design*. Springer Series in Statistics. New York, NY: Springer, June 21, 2006. ISBN: 978-0-387-31991-9. <https://doi.org/10.1007/0-387-37344-6>.
- Nagelkerke, N. J. D. 1991. "A Note on a General Definition of the Coefficient of Determination." *Biometrika* 78 (3): 691–692. ISSN: 0006-3444, 1464-3510. <https://doi.org/10.1093/biomet/78.3.691>.
- Neyman, Jerzy. 1934. "On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection." *Journal of the Royal Statistical Society* 97 (4): 558–625. ISSN: 09528385.
- Nunn, Nathan, and Diego Puga. 2012. "Ruggedness: The Blessing of Bad Geography in Africa." *The Review of Economics and Statistics* 94, no. 1 (February): 20–36. ISSN: 0034-6535. https://doi.org/10.1162/REST_a_00161.
- Oster, Emily. 2019. "Unobservable selection and coefficient stability: Theory and evidence." *Journal of Business & Economic Statistics* 37 (2): 187–204.
- Pearl, Judea. 1999. "Probabilities of Causation: Three Counterfactual Interpretations and Their Identification." *Synthese* 121 (1-2): 93–149. <https://doi.org/10.1023/A:1005233831499>.
- . 2009. "Causal Inference in Statistics: An Overview." *Statistics Surveys* 3:96–146. <https://doi.org/10.1214/09-SS057>.
- Pearl, Judea, and Elias Bareinboim. 2022. "External Validity: From Do-Calculus to Transportability Across Populations." In *Probabilistic and Causal Inference: The Works of Judea Pearl*, 1st ed., edited by Hector Geffner, Rina Dechter, and Joseph Y. Halpern, 36:451–482. ACM Books. Association for Computing Machinery. <https://doi.org/10.1145/3501714.3501741>.
- Peysakhovich, Alexander, and Jeffrey Naecker. 2017. "Using methods from machine learning to evaluate behavioral models of choice under risk and ambiguity." *Journal of Economic Behavior & Organization* 133:373–384.
- Porta, Miquel, ed. 2014. "Attributable Fraction for the Population." In *A Dictionary of Epidemiology*, 6th ed. Oxford University Press. ISBN: 9780199976720.
- Rässler, Susanne. 2012. *Statistical matching: A frequentist theory, practical applications, and alternative Bayesian approaches*. Vol. 168. Springer Science & Business Media.
- Ridder, Geert, and Robert Moffitt. 2007. "The Econometrics of Data Combination." In *Handbook of Econometrics*, edited by James J. Heckman and Edward E. Leamer, vol. 6, B, 5469–5547. Elsevier. [https://doi.org/10.1016/S1573-4412\(07\)06075-8](https://doi.org/10.1016/S1573-4412(07)06075-8).
- Rosenman, Evan TR, Guillaume Basse, Art B Owen, and Mike Baiocchi. 2023. "Combining observational and experimental datasets using shrinkage estimators." *Biometrics* 79 (4): 2961–2973.
- Rubin, Donald B. 1978. "Bayesian Inference for Causal Effects: The Role of Randomization." *The Annals of Statistics* 6 (1): 34–58. <https://doi.org/10.1214/aos/1176344064>.

- Sacks, Frank M., Laura P. Svetkey, William M. Vollmer, Lawrence J. Appel, George A. Bray, David Harsha, Eva Obarzanek, et al. 2001. "Effects on Blood Pressure of Reduced Dietary Sodium and the Dietary Approaches to Stop Hypertension (DASH) Diet." *New England Journal of Medicine* 344, no. 1 (January): 3–10. ISSN: 0028-4793, 1533-4406. <https://doi.org/10.1056/NEJM200101043440101>.
- Saiz, Albert. 2010. "The geographic determinants of housing supply." *The Quarterly Journal of Economics* 125 (3): 1253–1296.
- Scientific Advisory Committee on Nutrition. 2003. *Salt and Health*. Report. Published for the Food Standards Agency and the Department of Health. © Crown Copyright 2003. Scientific Advisory Committee on Nutrition, April.
- Spearman, C. 1904. "The Proof and Measurement of Association between Two Things." *The American Journal of Psychology* 15 (1): 72–101. ISSN: 00029556.
- Tabellini, Guido. 2010. "Culture and Institutions: Economic Development in the Regions of Europe." *Journal of the European Economic Association* 8, no. 4 (June 1, 2010): 677–716. ISSN: 1542-4766. <https://doi.org/10.1111/j.1542-4774.2010.tb00537.x>.
- University Of Cambridge, MRC Epidemiology Unit and NatCen Social Research. 2021. (National Diet and Nutrition Survey Years 1-11, 2008-2019 [SN 6533]. V. 19th Edition. [data collection]; accessed January 26, 2026). <https://doi.org/10.5255/UKDA-SN-6533-19>.
- Visscher, Peter M, William G Hill, and Naomi R Wray. 2008. "Heritability in the genomics era—concepts and misconceptions." *Nature reviews genetics* 9 (4): 255–266.
- Vytlacil, Edward. 2002. "Independence, monotonicity, and latent index models: An equivalence result." *Econometrica* 70 (1): 331–341.
- Wang, Xufei, Bo Jiang, and Jun S Liu. 2017. "Generalized R-squared for detecting dependence." *Biometrika* 104 (1): 129–139.
- Weitze, Jason. 2025. "Causal Attribution Bounds: Decomposing the Effects of Multiple Causes." Job market paper. "Link to latest version" on author website. November 7, 2025.
- Whelton, Paul K., Robert M. Carey, Wilbert S. Aronow, Donald E. Casey, Karen J. Collins, Cheryl Dennison Himmelfarb, Sondra M. DePalma, et al. 2018. "2017 ACC/AHA/AAPA/ABC/ACPM/AGS/APhA/ASH/ASPC/NMA/PCNA Guideline for the Prevention, Detection, Evaluation, and Management of High Blood Pressure in Adults: A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines." *Hypertension* 71, no. 6 (June). ISSN: 0194-911X, 1524-4563. <https://doi.org/10.1161/HYP.000000000000065>.
- Yamamoto, Teppei. 2012. "Understanding the past: Statistical analysis of causal attribution." *American Journal of Political Science* 56 (1): 237–256.

Appendix

Structure. The main appendices prove statements in the main text. The online appendices provide some other results and examples.

Appendix A. Proofs for Section 3

Proof of Lemma 1. The first sentence is direct from the definition of a causal model. For the second sentence, the risk of causal model M with constant c is

$\mathcal{R}(M) = \mathbb{E}[(M(X^O) - Y)^2] = \mathbb{E}[(\mathbb{E}[Y(x^O)] + c - Y)^2]$. This is a convex function in c . Taking a first-order condition, the risk-minimizing constant c^* satisfies $2\mathbb{E}[c^* + \mathbb{E}[Y(X^O)] - Y] = 0$, and hence $c^* = \mathbb{E}[Y] - \mathbb{E}[\mathbb{E}[Y(X^O)]]$. \square

Appendix B. Proofs for Section 4

Proof of Proposition 1. For (i), the best causal model is then $Y_{\mathcal{F}, X^O}^C(x^O) = \mathbb{E}[Y]$ for all X^O , and hence $\text{CR}_{\mathcal{F}}^2(X^O) = 0$. For (ii), since $Y(x^O, x^U) = Y(x^O, x^{U'})$ for any $(x^O, x^U, x^{U'})$, write $Y(x^O, x^U) = Y(x^O)$. The model $Y_{X^O}^C(x^O) = Y(x^O)$ clearly maximizes fit (since it achieves a fit of 1), and respects causal effects; since \mathcal{F} is well-specified, it falls within \mathcal{F} . Hence, the best causal model achieves a fit of 1.

For (iii), the best causal model is then constant at $\mathbb{E}[Y]$, and hence achieves risk $\text{Var}[Y]$, which implies $\text{CR}_{\mathcal{F}}^2(X^O) = 0$. The argument for (ii) also shows (iv).

For (v), since ε is causally unaffected by X^O , and is mean-zero, its addition does not affect the best causal model. Denote this best causal model by M^C . Then:

$$\begin{aligned} | \text{CR}_{\mathcal{F}}^2(Y' \rightarrow X^O) | &= | \frac{\text{Var}[Y'] - \mathbb{E}[(Y' - M^C(X^O))^2]}{\text{Var}[Y']} | \\ &= | 1 - \frac{\mathbb{E}[(Y - M^C(X^O))^2] + \text{Var}[\varepsilon]}{\text{Var}[Y] + \text{Var}[\varepsilon]} | \\ &\leq | 1 - \frac{\mathbb{E}[(Y - M^C(X^O))^2]}{\text{Var}[Y]} | = | \text{CR}_{\mathcal{F}}^2(Y \rightarrow X^O) | . \end{aligned}$$

For (vi), an example suffices: the true data-generating process for (Y_i, X_i) is $Y_i(X_i) = X_i, X_i \sim \mathcal{N}(0, 1)$. That is, X causally affects Y , but not *vice versa*. For an analyst studying the determinants of Y , who observes X , the best causal model is $Y_X^C(x) = x$, with $\text{CR}^2(X \rightarrow Y) = 1$. For an analyst studying the determinants of X , who observes Y , the best causal model

is $X_Y^C(y) = 0$, so $\text{CR}^2(Y \rightarrow X) = 0$. \square

Proof of Proposition 2. For (i), since the transformation applied to Y is affine, write the transformed outcome as $y' = \alpha + \beta y$, for $\beta \neq 0$. Write each transformed feature X_k as g_k , for $g_k(\cdot)$ strictly monotone. It is simple to show that the best causal model for Y' is $Y'_{X^O}^C(x^{O'}) = \alpha + \beta Y_{X^O}^C(g^{-1}(x^{O'}))$, where g^{-1} is the element-wise inverse of $(g_k)_{k=1}^O$, and hence risk is β^2 times the risk of the best causal model for Y . Since variance is also inflated by β^2 , the ratio of the risk to variance is unchanged, and hence CR^2 is unchanged. Part (ii) follows similar steps, additionally using the fact g^{-1} is affine. \square

Proof of Proposition 3. Part (i) follows directly from the fact that Definition 3 adds a constraint relative to Definition 2. For (ii), under independence, $\text{ATE}(x_1^O \rightarrow x_2^O) = \mathbb{E}[Y | X^O = x_2^O] - \mathbb{E}[Y | X^O = x_1^O]$, and hence the best causal and predictive models coincide, and so have the same fit. \square

Proof of Proposition 4. In (i), the statement about predictive R^2 is well-known. That CR^2 is bounded above by 1 follows directly from the fact that risk is non-negative. To show that CR^2 may be negative, consider the example discussed in Section 2, with $\alpha = 0$. The best causal model (unrestricted or linear) is $Y^C(c_i) = \beta c_i$, with goodness-of-fit

$$G(Y_{X^O}^C) = 1 - \frac{\gamma^2}{\beta^2 + \gamma^2 + 2\beta\gamma\rho} = \frac{\beta^2 + 2\beta\gamma\rho}{\beta^2 + \gamma^2 + 2\beta\gamma\rho}.$$

Choosing $\rho = -1$, the expression can be rewritten $G(Y_{X^O}^C) = \frac{\beta(\beta-2\gamma)}{(\beta-\gamma)^2}$, which can be made arbitrarily negative by taking $\beta = \gamma + \epsilon$ for ϵ small.

Part (ii) can again be illustrated by example: add a third variable Z to the example above with no causal effect. The causal R^2 of Z is zero; following the logic above, the joint causal R^2 of Z and C will be negative. \square

Before proving Proposition 5, we establish an intermediate step.

Lemma A1. Fix an outcome Y and a vector of observed features X^O , both with finite first and second moments and non-zero variances. Denote by \tilde{Y} the standardized value of Y , by \tilde{X}_k the standardized value of observed feature X_k , and by \tilde{X}^O the corresponding vector of standardized observed features. Denote by ρ_{X^O} the correlation matrix of observed features.

(i) $R_{\text{lin}}^2(X^O) = (\tilde{\beta}^P)' \rho_{X^O} \tilde{\beta}^P$, where $\tilde{\beta}^P$ is the vector of OLS coefficients from a regression of \tilde{Y} on \tilde{X}^O in the population.

(ii) If \mathcal{F}_{lin} is well-specified, $\text{CR}_{\text{lin}}^2(X^O) = 2(\tilde{\beta}_{X^O}^P)' \rho_{X^O} \tilde{\beta}_{X^O}^C - (\tilde{\beta}_{X^O}^C)' \rho_{X^O} \tilde{\beta}_{X^O}^C$, where $\tilde{\beta}^C$ is the vector of OLS coefficients from a regression of \tilde{Y} on \tilde{X}^O in an experiment in which X^O is randomly assigned.

Part (i) is well-known, but we include a proof for completeness.

Proof of Lemma A1. It is well-known that the R_{lin}^2 is invariant to affine transformations, and Proposition 2 shows that this is also true of CR_{lin}^2 . Then we can demonstrate the lemma by considering the standardized versions of all variables. For any linear model $\tilde{M}(\tilde{x}^O) = \tilde{\alpha} + \tilde{\beta}' \tilde{x}^O$ for the outcome \tilde{Y} :

$$\begin{aligned} G(M) &= 1 - \frac{\mathbb{E}[(\tilde{Y} - M(\tilde{X}^O))^2]}{\text{Var}(\tilde{Y})} \\ &= 1 - [1 - 2\text{Cov}[\tilde{Y}, \tilde{M}(\tilde{X}^O)] + \text{Var}[M(\tilde{X}^O)]] \\ &= 2\text{Cov}[\tilde{Y}, \tilde{\beta}' \tilde{X}^O] - \tilde{\beta}' \rho_{X^O} \tilde{\beta}. \end{aligned}$$

For (i), take the best linear predictor, $Y_{\text{lin}, X^O}^P(\tilde{x}^O) = \tilde{\alpha}^P + \tilde{\beta}^P \tilde{x}^O$. Then, $\text{Cov}[\tilde{Y}, (\tilde{\beta}^P)' \tilde{X}^O] = (\tilde{\beta}^P)' \rho_{X^O} \tilde{\beta}^P$, and hence $R_{\text{lin}}^2(X^O) = (\tilde{\beta}^P)' \rho_{X^O} \tilde{\beta}^P$. For (ii), take the best linear causal model, $Y^C(X^O) = \tilde{\alpha}^C + \tilde{\beta}^C \tilde{X}^O$. If the model is well-specified, $\text{CR}_{\text{lin}}^2(X^O) = 2\text{Cov}[\tilde{Y}, (\tilde{\beta}^C)' \tilde{X}^O] - (\tilde{\beta}^C)' \rho_{X^O} \tilde{\beta}^C$. \square

Proof of Proposition 5. Beginning with the second term on the RHS:

$$\begin{aligned} (\tilde{\beta}_{X^O}^P - \tilde{\beta}_{X^O}^C)' \rho_{X^O} (\tilde{\beta}_{X^O}^P - \tilde{\beta}_{X^O}^C) &= (\tilde{\beta}_{X^O}^P)' \rho_{X^O} \tilde{\beta}_{X^O}^P - 2(\tilde{\beta}_{X^O}^P)' \rho_{X^O} \tilde{\beta}_{X^O}^C + (\tilde{\beta}_{X^O}^C)' \rho_{X^O} \tilde{\beta}_{X^O}^C \\ &= R_{\text{lin}}^2(X^O) - \text{CR}_{\text{lin}}^2(X^O), \end{aligned}$$

where the first line follows from the fact ρ_{X^O} is a correlation matrix (and so symmetric), and the second line applies both parts of Lemma A1. \square

Appendix C. Proofs for Section 5

Proof of Proposition 6. For (i), $Y_{\mathcal{F}, X^O}^C$ cannot be identified from the observational sample. Conversely, for any model M , $\mathcal{R}(M)$ cannot be identified from the experimental sample. Since $\text{CR}_{\mathcal{F}}^2(X^O)$ depends on both $Y_{\mathcal{F}, X^O}^C$ and $\mathcal{R}(M)$, it cannot be identified by either sample alone.

For (ii), for any M , $\mathcal{R}(M)$ is identified by the observational subsample. If the experiment is full-support, for each $x^O \in \mathcal{X}^O$, we can identify $Y_{X^O}^C(x^O)$ pointwise in the experimental subsample, and hence identify the function $Y_{X^O}^C$. Otherwise, we cannot identify $Y_{X^O}^C(x^O)$ for at least some values of x^O .

For (iii), as before, for any model M , $\mathcal{R}(M)$ is identified by the obser-

vational subsample. Under our assumption that the linear model is well-specified, the best causal model is $Y_{X^O}^C(x^O) = \alpha + \sum_{k=1}^O x^O \beta^O$. If the experiment is full-rank, regressing Y on X^O in the experimental subsample recovers (α, β) , and hence $Y_{X^O}^C$. If the experiment is not full-rank, it is easy to see that at least one β^k must not be identified. \square

Proof of Proposition 7. Follows immediately from noting that (i) $\hat{Y}_{\mathcal{F}, X^O}^C$ converges to $Y_{\mathcal{F}, X^O}^C$, (ii) for any M , $\hat{\mathcal{R}}(M)$ converges to $\mathcal{R}(M)$, (iii) $\widehat{\text{Var}}[Y]$ converges to $\text{Var}[Y]$, and then applying Slutsky's Theorem. \square

Proof of Proposition 8. For simplicity, denote $X := X^O$. Denote by Y_{lin}^C the best linear causal model for the relationship between X and Y . For part (i), noting that CME in the observational data does not affect the best causal model:

$$\begin{aligned} \text{CR}_{\text{CME}}^2 &= 1 - \frac{\mathbb{E}[(Y' - Y_{\text{lin}}^C(X))^2]}{\text{Var}[Y']} \\ &= 1 - \frac{\mathbb{E}[(Y - Y_{\text{lin}}^C(X))^2] + \text{Var}[\epsilon]}{\text{Var}[Y] + \text{Var}[\epsilon]} = \text{CR}^2 \frac{\text{Var}[Y]}{\text{Var}[Y] + \text{Var}[\epsilon]}, \end{aligned}$$

where the second line makes use of the independence of ϵ . For part (ii), as before, CME in the observational data does not affect the best causal model. Denoting by β the slope coefficient from a regression of Y on X in the experimental data:

$$\begin{aligned} \text{CR}_{\text{CME}}^2 &= 1 - \frac{\mathbb{E}[(\alpha + \beta(X + \epsilon) - Y)^2]}{\text{Var}[Y]} \\ &= \frac{\text{Var}[Y] - \mathbb{E}[(\alpha + \beta X - Y)^2] - \beta^2 \text{Var}[\epsilon]}{\text{Var}[Y]} = \text{CR}^2 - \beta^2 \frac{\text{Var}[\epsilon]}{\text{Var}[Y]}, \end{aligned}$$

where the first line defines (α, β) as the intercept and slope respectively in the best causal model, and the second line again uses the independence of ϵ .

Part (iii) follows from the well-known result that classical measurement error in the dependent variable does not bias OLS coefficients. For part (iv), denote by $\text{Var}_{\text{EXP}}[\cdot]$ the variance in the experiment. It is well-known that, under CME in the independent variable, the OLS slope coefficient will be attenuated as follows: $\beta_{\text{CME}} = \beta \frac{\text{Var}_{\text{EXP}}[X]}{\text{Var}_{\text{EXP}}[X] + \text{Var}_{\text{EXP}}[\epsilon]}$. The result follows from subtracting CR^2 from CR_{CME}^2 and rearranging. \square