The Explanatory Power of Causal Effects

James Stratton and Nicolaj Thor*

November 22, 2025

Abstract

We propose a causal analogue to the predictive R^2: a measure of the share of variation in an outcome *causally* explained by a variable. This "causal R^2" (CR^2) can be interpreted as the goodness-of-fit of a causal model. CR^2 is identified by combining observational and experimental data. We illustrate the measure in several applications. Spring protection causally explains 30% of water quality in Kenyan data. Class size predicts 8% of reading test scores but causally explains only about 3%. Institutions causally explain one-fifth of the variation in national income. Salt intake raises blood pressure similarly for men and women, but explains much less variation among women.

*Emails: jstratton@g.harvard.edu and thor@brown.edu. We are grateful to have received comments from Adamson Bryant, Raj Chetty, Cole Davis, John Friedman, Peter Hull, Kosuke Imai, Larry Katz, Toru Kitagawa, Soonwoo Kwon, Sendhil Mullainathan, Florian Mudekereza, Jonathan Roth, Jesse Shapiro, Elie Tamer, Keyon Vafa, 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

How much of the variation in an outcome Y does a variable X explain? Questions of this form are central to the social sciences: Across workers, what share of wage dispersion is explained by differences in education (Mincer 1974; Card 1999)? Across cities, what share of variation in growth is explained by zoning (Saiz 2010; Glaeser 2011)? Across countries, what share of variation in income is explained by institutions (Hall and Jones 1999; Acemoglu et al. 2001)?

If by "explain" we mean "predict", the R^2 answers these questions. Yet R^2 receives little emphasis in economics, precisely because we typically seek *causal* explanations: if X forecasts Y, but does not causally affect it, then X does not truly "explain" variation.Evaluating these causal explanations requires assessing how much variation in an outcome is attributable to a given cause. For instance, to assess the view that institutions are the fundamental cause of differences in national income, we need a measure of the share of variation in national income caused by differences in institutions. Existing fit statistics such as the R^2 depend only on the joint distribution of (Y, X), and hence cannot untangle correlation from causation.

By contrast, the causal inference literature estimates the effect of *changes* in X on *changes* in Y (*e.g.*, how much an extra year of school raises wages), but not whether *observed variation* in Y is explained by *observed variation* in X. To see this distinction in an extreme case, note that if all workers had identical schooling, education would explain no variation in wages, no matter its causal effect.[1] More generally, even when X has a large causal effect on Y *and* a large variance, it may account for little of the *observed cross-sectional* variance in Y: a highly-effective tutoring program targeted at struggling students may have large causal effects, but fail to explain variation in test scores; indeed, if assignment is sufficiently negatively-selected on potential outcomes, the program may even *reduce* variance in scores. Thus, existing causal tools tell us the causal effect of X on Y, but not how much of Y's variance X explains. To our knowledge, there is no measure of the variation in an outcome causally explained by a variable.

This paper proposes such a measure. Recall that the (non-parametric) population R^2 is the proportional reduction in population mean squared error (MSE) achieved by residualizing Y with respect to its conditional expectation $\mathbb{E}[Y \mid X]$. We view $\mathbb{E}[Y \mid X]$ as a *predictive model* for Y, and R^2 as its goodness-of-fit. Our corresponding *causal model* is the *interventional expectation* $\mathbb{E}[Y \mid\mid X]$, defined as the expected value of Y *conditional on intervening to set* X (Imbens 2014). We define the (non-parametric) population causal R^2 (CR^2) as the proportional reduction in population MSE achieved by residualizing Y with respect to this interventional expectation.

We interpret CR^2 as the share of variance causally explained. Intuitively, for each unit i, we compare the realized outcome Y_i to the outcome one would expect if only X caused variation, $\mathbb{E}[Y \mid\mid X_i]$. We treat the squared residual $(Y_i - \mathbb{E}[Y \mid\mid X_i])^2$ as *causally unexplained variation*.

[1] Likewise, a pre-market drug that has not been released explains no variation in health outcomes, no matter its effects in a clinical trial. Many economics RCTs also involve treatments to which the population of interest has no previous exposure (*e.g.*, Miguel and Kremer 2004; Kleven et al. 2011; Breza and Chandrasekhar 2019).

Averaging over units gives expected unexplained variation; dividing by the outcome variance gives the share of variance that is unexplained. Our CR^2 is the complement of this value, which we interpret as the share of variance causally explained.

We view our approach as the natural extension of R^2 to causal models. Nonetheless, we are not aware of any prior work proposing this approach. When describing the measure's properties, we show that CR^2 is intuitive: it equals 0 if X has no causal effect, and 1 if X fully determines it; it is unitless; and it is bounded above by the predictive R^2, with equality if observables are independent of unobservables. While the standard predictive R^2 is weakly positive, the causal R^2 can be positive, when the variables of interest contribute to variation in the outcome, or negative, when they tend to *suppress* variation in the outcome.

We then turn to identification and estimation. The causal R^2 combines (i) the interventional expectation, which represents the causal effect of X on Y, and (ii) the joint distribution of (Y, X), which determines the interventional expectation's goodness-of-fit. Accordingly, identification requires (i) an *observational dataset*, containing realizations of (Y, X) from the population, and (ii) an *experimental dataset*, containing realizations of (Y', X') in which X' is randomly assigned, potentially affecting Y'.[2] A plug-in estimator is consistent, and we assess its performance in simulations. Inference follows by the boostrap or the Delta method.

We demonstrate the practical relevance of CR^2 in five applications. First, we apply the measure to data from Kremer et al. (2011), who randomized spring protection in Kenya. Spring protection both *predicts* and *causally explains* about a third of variation in water quality. By contrast, in Project STAR, variation in class size *predicts* 8% (5%) of variation in reading (math) scores, but *causally explains* only 3% (2%): the predictive power of class size mostly reflects omitted variables, not causal effects. Applying CR^2 to the settler mortality instrument in Acemoglu et al. (2001), around one fifth of cross-country income variation is causally explained by differences in institutions' extractiveness. Similarly, in Fluegge (2025), around one-fifth of variation in U.S. city growth over the last century is causally explained by exposure to the 1918 Influenza Pandemic. Last, we assess the share of variation in blood pressure explained by sodium intake: sodium causally explains 7% of variation in men's blood pressure, but less than 1% in women, despite similar causal effects. The gender difference arises because women's blood pressure varies more for reasons unrelated to sodium.

These applications illustrate the usefulness of CR^2. *Descriptively*, the measure's value is twofold. First, it evaluates the extent to which a theory provides a complete explanation of variation in the outcome. Second, it indicates the value of future research: if the explained share is low, then we do not yet have a complete theory of the outcome.

Normatively, a policy-maker may care about X, even though it explains little variation in Y. Goldberger (1979), writing in the context of genetic heritability, gives the example of eyesight: population variance in eyesight is largely genetic, but there is still great value in prescribing

[2] Our baseline setting is a randomized experiment, but the approach extends to quasi-experimental variation.

glasses.[3] For this reason, CR^2 is more relevant to a scientist seeking to understand the sources of naturally-occurring variation in Y, than to a policy-maker seeking to affect the value of Y.[4]

Of course, there are 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. We return to these advantages throughout the paper.

Related literature. An existing literature extends R^2 to settings other than causal models.[5] 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 as an input towards assessing causal effects; we instead develop a causal analogue of the R^2.[6]

We also contribute to the literature on causal attribution and the causes of effects. The causal attribution literature asks: given two observed variables (X_1, X_2), and a known potential outcome function $Y(X_1, X_2)$, how much of the joint effect of (X_1, X_2) should be attributed to X_1 *vs.* X_2 (*e.g.*, Datta et al. 2016; Heskes et al. 2020; Jung et al. 2022; Weitze 2025)? By contrast, we compare the variation explained by *observed* variables to the *total* variation in Y. The causes of effects literature asks whether a given unit's outcome would have differed under a different treatment (*e.g.*, Pearl 1999; Halpern and Pearl 2005; Yamamoto 2012; Dawid and Musio 2022);[7] we instead ask what share of *population* variance can be attributed to 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:

> "[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." (p. 3)

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

We also connect to the economics literature on completeness, which compares the predictive power of theory-constrained *vs.* unconstrained models (*e.g.*, Peysakhovich and Naecker 2017;

[3] Manski (2011) echoes this point. Relatedly, the share of variance causally explained is not policy-invariant: for instance, the share of variance in eyesight explained by genetics will be smaller in a population with glasses.

[4] Indeed, many economic experiments—such as those estimating the effect of immutable characteristics (Bertrand and Mullainathan 2004; Neumark et al. 2019)— cannot be interpreted as evaluating policy counterfactuals, and can *only* be understood as explaining differences in outcomes.

[5] These 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" (Wang et al. 2017) and "out-of-sample" (Hawinkel et al. 2024) R^2.

[6] Similarly, "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 et al. 2008), and a risk factor for the attributable fraction (Porta 2014). Both measures assess predictive, not causal, relationships.

[7] This literature has received limited application in economics (*e.g.*, Ganong and Noel 2023).

Apesteguia and Ballester 2021; Fudenberg et al. 2022). We study causal, not predictive models, though our metric can be viewed as the performance of a model constrained to be causal.

Finally, our work relates to the external validity literature. In our baseline setting, an analyst samples observations from the population, and assigns units to either an observational or an experimental setting. Even absent selection into the experiment, the joint distribution of treatment and unobservables in the experiment differs from the population distribution *by construction*;[8] in consequence, the share of variation causally explained *within the experiment*—which we call the "experimental R^2"—differs from the share explained *in the population* (the causal R^2). Under some assumptions, we can extend our results to the case in which the experiment is conducted on a non-random subpopulation, following the transportability literature (Imbens 2010; Angrist and Fernández-Val 2013; Mogstad and Torgovitsky 2018).

Outline. Section 2 presents an illustrative example which introduces some of the main ideas. Section 3 defines CR^2. Section 4 discusses properties. Section 5 covers identification, estimation, and inference. Sections 6–7 present simulations and applications. Proofs are in the appendix; some extra results are in the Online Appendix.

2. An illustrative example

We first illustrate our approach in a simple example. Say we study the relation between student test scores and class sizes. Let Y_i denote student i's test score, C_i class size, and I_i family income. To give the intuition, we begin by imposing a simple constant effects model:

$$Y_i(C_i, I_i) = \alpha + \beta C_i + \gamma I_i, \tag{1}$$

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

where (1) is the potential outcome function for test scores, and (2) is the joint distribution of class size and income, which jointly determine the distribution of test scores.[9] We observe scores and class size, but not income.[10] In particular, we sample observational data—draws of (Y_i, C_i)—and also run an experiment that samples students from the population, randomly assigns class sizes (leaving family income unchanged), and records resulting test scores.

What share of variation in test scores is explained by class size? We first consider this question in the *observational data*; Figure 1(A) presents example realizations. The corresponding line of best fit converges, as the sample grows, to the conditional expectation function $Y^P(C_i) \coloneqq$

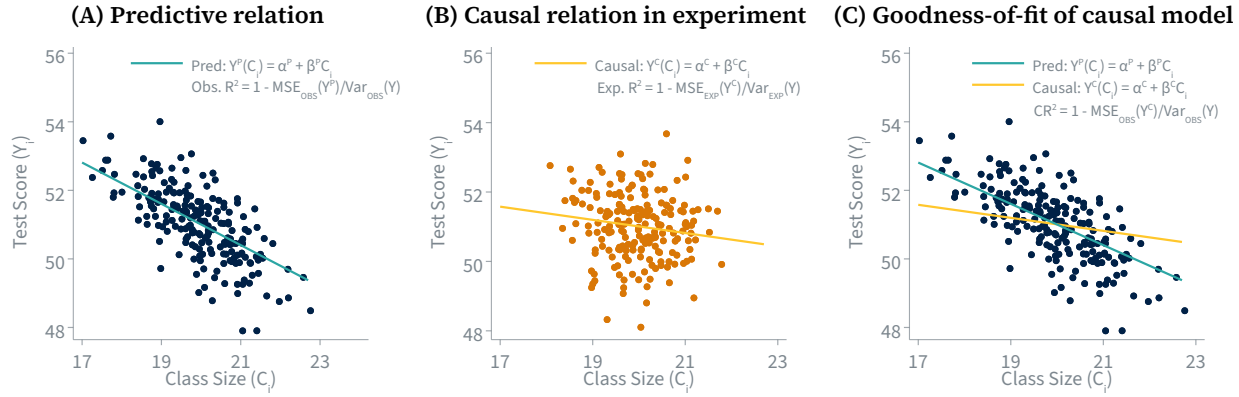
[8] Since treatment is assigned independent of unobservables, the distribution of the outcome is different than in observational data, in which treatment and unobservables covary.

[9] That is, $Y \sim \mathcal{N}(\alpha + \beta\mu_C + \gamma\mu_I, \beta^2 + \gamma^2 + 2\beta\gamma\rho)$, with $\text{Cov}[Y, C] = \beta + \gamma\rho$. For convenience, the example treats all variables as though they are continuous.

[10] If we observed (Y_i, C_i, I_i), we could recover (1) and fully explain variation in test scores. The problem arises only when some determinants are unobserved.

$\mathbb{E}[Y_i \mid C_i] = \alpha^P + \beta^P C_i$, for $(\alpha^P, \beta^P) \coloneqq (\alpha + \gamma\mu_I - \gamma\rho\mu_C, \beta + \gamma\rho)$. This is the best predictive model: it minimises population squared prediction error. Its fit—the standard R^2—measures the share of variation in scores that class size *predicts*, but not how much it *causally explains*: a high R^2 can arise even if class size has no causal effect ($\beta = 0$). More generally, since the causal effect of C on Y is not identified by the joint distribution of (Y, C), the observational dataset does not contain all the information necessary to compute variance *causally* explained.

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



Note: *This figure presents an example of assessing the predictive and causal fit of models relating test scores (Y_i) to class size (C_i). Panel (A) presents example realizations of observational data. Panel (B) presents example realizations of experimental data. Panel (C) contrasts the fit of the best predictive and causal models in the observational data.*

Motivated by this fact, we turn to the experimental data (Figure 1(B)). Since class size is randomized, the best fit line converges to the interventional expectation $Y^C(C_i) \coloneqq \mathbb{E}[Y_i \mid\mid C_i] = Y^C(C_i) = \alpha^C + \beta C_i$, for $\alpha^C \coloneqq \alpha + \gamma\mu_I$, where $\mid\mid$ denotes "conditioning by intervention" (Imbens 2014, p. 10).[11] We call Y^C the *best causal model*. 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, as expected if family income increases test scores ($\gamma > 0$) and correlates with smaller classes ($\rho < 0$).

The best causal model recovers the true causal effect β. Because of this, there is some temptation 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 because the experimental treatment distribution generally differs from the population distribution, for two reasons. First, the *marginal* distributions differ: to return to an example in the introduction, a trial of a pre-market drug can yield a positive R^2 even though the drug has not yet been released to the population, and hence explains no existing health outcomes. Second, even if the marginal distributions coincide—for instance, because the experiment was designed *ex ante* to mimic the population distribution, or because it has been reweighted *ex post*—the *joint* distribution of the treatment and unobservables in the experiment will differ from the population joint distribution. This is by construction:

[11] We define this notation formally in Section 3.

randomization in the experiment ensures that C is independent of unobserved family income I. As a result, the distribution of the outcome will differ from that in the population. Since we aim to explain the outcome in the population, the experimental R^2 never suffices.

For this reason, assessing the share of variance causally explained requires assessing the fit of the causal model in the *observational* data (Panel C). One tempting approach is to compute the variance of the best causal model in the observational data, and to define the ratio of this variance to the variance of Y as the share of variance explained (that is, to define the share of variance explained as $\frac{\text{Var}(Y^C)}{\text{Var}(Y)} = \frac{\beta^2}{\beta^2+\gamma^2+2\beta\gamma\rho}$). We view this as a reasonable measure of the *movement created* by class size, but not of the share of variance *causally explained* by class size. One way to see this is to impose the logical requirement that the share of variance causally explained not exceed 100%: there is no such guarantee for the expression above, and indeed if class size and parent income are sufficiently negatively correlated, then the measure will exceed 100%.[12] More generally, the expression above accounts for the *effects* of class size on test scores, but not whether class size has *explanatory power*, in the sense that the effects of class size account for the observed variation in test scores.

Instead, we take the following approach. Take a student with score y_i and class size c_i. Class size causally explains a score of $\alpha^C + \beta c_i$. The residual $y_i - \alpha^C - \beta c_i$ (the vertical gap between the causal model and a point i) is unexplained and must come from another cause. Summing these squared residuals and dividing by the outcome variance gives the share of variance not causally explained; the complement is the share of variation explained. This parallels the usual R^2, but here the variation is explained causally, not just predictively. We call this measure the causal R^2 (CR^2). Throughout the paper, we argue that this measure is an appealing and natural way to adapt the R^2 to causal models.

3. Defining the goodness-of-fit of a causal model

This section develops the causal R^2 as the fit of a causal model. We formalize the heuristic approach described in the last section, relaxing the simplifying assumptions of a linear, constant-effects model. We begin by describing our setting of interest. We then define causal models and model fit. Finally, we define non-parametric and linear CR^2.

3.1. Setting

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

[12] This will be the case if $\rho < -\frac{\gamma}{2\beta}$, assuming that $\gamma, \beta > 0$.

We interpret $Y(x_i)$ as the outcome for a *unit i* with feature realizations x_i. That is, we treat the potential outcome function $Y(\cdot)$ as a *population* object, rather than defining a unit-specific $Y_i(\cdot)$. This notation is without loss of generality, since X includes all sources of variation in Y.[13] The effects of features may differ across units, but the sources of heterogeneity are included in X. We adopt this notation to emphasize that the goal of interest is understanding the determinants of variation in Y, which comes fully from variation in X.[14]

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

$$F_{Y,X}(y, x_1, \dots, x_K) = \int_{\mathcal{X}} \mathbb{1}_{\{Y(v_1,\dots,v_K)\le y\}} \mathbb{1}_{\{v_1\le x_1,\dots,v_K\le x_K\}} \, \mathrm{d}P_X(v_1, \dots, v_K).$$

Features may be observed or unobserved. Write the *observed features* as $X^O := (X_1, \dots, X_O)$ for $O \le K$, and the *unobserved features* as $X^U := (X_{O+1}, \dots, X_K)$. Denote by P_{Y,X^O} the marginal distribution of (Y, X^O) induced by $P_{Y,X}$. We assume throughout that the outcome and features have finite second moments. 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 function from $\mathcal{X}^O$ to $\mathcal{Y}$. If we observe all features ($O = K$), a natural model is the potential outcome function $Y(\cdot)$, which perfectly predicts and causally explains Y. We think of $Y(\cdot)$ as the *true model*. If we only observe some features ($O < K$), models of predictive *vs* causal relations may differ: a feature might predict, but not cause an outcome.

Below, we formally define the best predictive model and the best causal model of Y given observed features.[15] Since unobserved features may affect the outcome, these models may incur error. We evaluate this error using quadratic loss; other loss functions yield similar results. All expectations and variances are with regard to P_X unless otherwise specified.

Definition 1. Given observed features X^O and an outcome Y:

[13] Similar notation appears in, *e.g.*, Vytlacil (2002), Hernán and Robins (2025), and Pearl (2009). To see formally that this is without loss of generality, adopt the notation of Vytlacil (2002): let $(\Omega, \mathcal{F}, \mathbb{P})$ denote the probability space, with ω an element of Ω, and for each $x \in \mathcal{X}$, denote by $Y_x(\omega)$ the random variable corresponding to the potential outcome under x. Instead of this unit-specific potential outcome, we redefine the feature space X to include any latent factors or indeed the unit index itself: $\tilde{X}(\omega) = (X(\omega), \omega)$. In that case, we can define the population-level $Y(\cdot)$ by $Y(\tilde{X}(\omega)) = Y_{X(\omega)}(\omega)$ for all ω.

[14] An alternative approach would be to define a unit-level potential outcome function $Y_i = \hat{Y}(X_i^O, X_i^U) + \epsilon_i$, for some noise term ϵ_i that is mean-zero and independent of the observed and unobserved features, and then assess the share of variation in $\hat{Y}(X_i^O, X_i^U)$ that is causally explained by variation in X_i^O, ignoring the role of ϵ_i. This approach is in the spirit of Oster (2019), who defines a maximum plausible value of R-squared, called $\mathrm{R}_{\max}$ and compares the observed R^2 to $\mathrm{R}_{\max}$. This raises the difficulty of establishing a plausible value for $\mathrm{R}_{\max}$, which Oster (2019) considers in detail. A similar issue is discussed in the literature on risk adjustment in health insurance plans (see section 3.2.6 Wynand et al. 2000). We do not pursue such an approach here.

[15] When all features are observed, these models coincide with the true model (the potential outcome function).

(i) the ***best predictive model*** is the conditional expectation:

$$Y^P_{X^O}(x^O) \coloneqq \arg\min_z \mathbb{E}\big[(z-Y)^2 \mid X^O = x^O\big] = \mathbb{E}\big[Y \mid X^O = x^O\big].$$

(ii) the ***best causal model*** is the interventional expectation, or average potential outcome:[16]

$$Y^C_{X^O}(x^O) \coloneqq \arg\min_z \mathbb{E}\big[(z-Y)^2 \mid\mid X^O = x^O\big] = \mathbb{E}\big[Y \mid\mid X^O = x^O\big],$$

where $\mid\mid$ "condition[s] by intervention" (Imbens 2014, p. 10):[17] $\mathbb{E}\big[Y \mid\mid X^O = x^O\big] \coloneqq \mathbb{E}_{P_{X^U}}\big[Y(x^O, X^U)\big]$.

To give more intuition on causal models, we offer three perspectives. First, causal and predictive models differ only in the distribution of unobserved features:

$$\text{Predictive: } Y^P_{X^O}(x^O) = \mathbb{E}_{P_{X^U \mid X^O = x^O}}\big[Y(x^O, X^U)\big], \qquad \text{Causal: } Y^C_{X^O}(x^O) \;\; = \mathbb{E}_{P_{X^U}}\big[Y(x^O, X^U)\big],$$

The best predictive model traces the expected value of Y as X^O varies, *accounting for the covariance of X^O and X^U*. The best causal model traces the expected value of Y as X^O is manipulated *independently of X^U*. Intuitively, causal analysis severs the covariance between observed and unobserved features (Holland 1986; Angrist and Pischke 2009).

Second, the best predictive model is the conditional expectation in the population, whereas the best causal model is the conditional expectation in an experiment that randomly assigns X^O. This view connects the best predictive and causal models to Figure 1.

Third, the best causal model equals the average potential outcome that only depends on observables. Let $Y_{X^O}(x^O) = Y(x^O, X^U)$ denote that potential outcome function. Then, $\mathbb{E}\big[Y_{X^O}(x^O)\big] = Y^C_{X^O}(x^O)$. This view relates the causal model to causal effects. The average treatment effect of changing X^O from one realization to another (say from 0 to 1) is the difference in the best causal model evaluated at those points:

$$\text{ATE} = \mathbb{E}[Y_{X^O}(1) - Y_{X^O}(0)] = \mathbb{E}[Y_{X^O}(1)] - \mathbb{E}[Y_{X^O}(0)] = Y^C_{X^O}(1) - Y^C_{X^O}(0).$$

Similarly, the association, or predictive "effect", is $Y^P_{X^O}(1) - Y^P_{X^O}(0)$. In that way, causal and predictive models 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.[18]

3.3. Risk and fit

We now define goodness-of-fit, and apply it to causal models.

[16] We make the standard stable unit treatment value assumption (SUTVA).

[17] Other notation exists for this interventional expectation. Neyman (1923) calls it the *best estimate* of potential yields; see Imbens and Rubin (2015). It is $\mathbb{E}\big[Y(x^O)\big]$ in Hernán and Robins (2006), and $\mathbb{E}\big[Y \mid \text{do}(x^O)\big]$ in Pearl (1995). We believe the $\mid\mid$ notation originated in Lauritzen and Richardson (2002).

[18] We have so far equated observable and manipulable features. In practice, some variables may be observable but not manipulable. We return to this distinction in section 5.5.

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

When all features are observed, the true model $Y(\cdot)$ perfectly explains the outcome, and so achieves zero risk: $\mathcal{R}(Y(\cdot)) = 0$. Conversely, we may consider a *baseline model*, $\mathbb{E}[Y]$, which is the best causal and predictive model when no features are observed, with risk $\text{Var}[Y]$.

Definition 3. The ***fit*** of model M is its proportional reduction in risk relative to the baseline model:

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

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 R^2: $\text{R}^2(X^O) \coloneqq G(Y^P_{X^O})$.[19] We define the ***non-parametric causal*** R^2 as the fit of the best causal model: $\text{CR}^2(X^O) \coloneqq G(Y^C_{X^O})$.

3.4. Parametric causal R^2

So far, models minimize error non-parametrically. 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. Let $\mathcal{F}$ be our function class of interest, *e.g.*, all linear models, $\mathcal{F}_{\text{lin}}$; $\mathcal{F}$ may be motivated by prior knowledge about the relation between Y and X^O, or by a need for tractability. [20]

Definition 4. Given observed features X^O, an outcome Y, and a function class $\mathcal{F}$:

(i) the ***best predictive model under*** $\mathcal{F}$ is $Y^P_{\mathcal{F},X^O} \coloneqq \arg\min_{M\in\mathcal{F}} \mathbb{E}\big[\mathbb{E}\big[(M(X^O) - Y)^2 \mid X^O = x^O\big]\big]$, with fit $\text{R}^2_{\mathcal{F}}(X^O) \coloneqq G(Y^P_{\mathcal{F},X^O})$;

(ii) the ***best causal model under*** $\mathcal{F}$ is $Y^C_{\mathcal{F},X^O} \coloneqq \arg\min_{M\in\mathcal{F}} \mathbb{E}\big[\mathbb{E}\big[(M(X^O) - Y)^2 \mid\mid X^O = x^O\big]\big]$, with fit $\text{CR}^2_{\mathcal{F}}(X^O) \coloneqq G(Y^C_{\mathcal{F},X^O})$.

We call $\mathcal{F}$ ***well-specified*** if it includes the best causal model $Y^C_{X^O}$. Otherwise, $\mathcal{F}$ is ***misspecified***. The non-parametric CR^2 is nested as the case in which $\mathcal{F}$ is 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 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, $\text{CR}^2_{\text{lin}}(X^O)$, as the fit of this model. Applying our definition of goodness-of-fit, it is simple to show that this linear causal R^2 simply replaces the observational OLS coefficients $(\alpha^{\text{OBS}}, \beta^{\text{OBS}})$ in the definition of the predictive R^2 with corresponding experimental coefficients $(\alpha^{\text{EXP}}, \beta^{\text{EXP}})$:

$$\text{R}^2_{\text{lin}}(X^O) = 1 - \frac{\mathbb{E}\big[(Y - \alpha^{\text{OBS}} - (\beta^{\text{OBS}})^\top X^O)^2\big]}{\text{Var}[Y]}, \qquad \text{CR}^2_{\text{lin}}(X^O) = 1 - \frac{\mathbb{E}\big[(Y - \alpha^{\text{EXP}} - (\beta^{\text{EXP}})^\top X^O)^2\big]}{\text{Var}[Y]}.$$

[19] The non-parametric R^2 can be traced back to Rényi (1959), and is defined formally in Doksum and Samarov (1995). For recent discussions, see Li and Wang (2019) and Fudenberg et al. (2022).

[20] We assume $\mathcal{F}$ includes at least all constant models M, and that it contains two models M and M' which differ, in the sense that $\mathbb{E}[((M(X^O_i) - M'(X^O_i))^2] > 0$.

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.[21] Sometimes we specify the outcome and write $CR^2_{\mathcal{F}}(Y \to X^O)$ for the causal R^2 of observed features X^O, outcome Y, and function class $\mathcal{F}$.

Proposition 1 (basic properties). *For any function class $\mathcal{F}$, distribution of features P_X, and potential outcome function $Y(\cdot)$:*

(i) The causal R^2 is 0 if the observable features have no effect on the outcome: $\left(\forall x^O, x^{O'}, x^U, Y(x^O, x^U) = Y(x^{O'}, x^U)\right) \implies \mathrm{CR}^2_{\mathcal{F}}(X^O) = 0$.

(ii) If $\mathcal{F}$ is well-specified, the causal R^2 is 1 if the observable features fully determine the outcome: $\left(\forall x^O, x^U, x^{U'} \quad Y(x^O, x^U) = Y(x^O, x^{U'})\right) \implies \mathrm{CR}^2_{\mathcal{F}}(X^O) = 1$.

(iii) The causal R^2 is strictly less than 1 if Y varies even after intervening on X^O: $\mathbb{E}\left[\mathrm{Var}\left[Y \mid\mid X^O = x^O\right]\right] > 0 \implies \mathrm{CR}^2_{\mathcal{F}}(X^O) < 1$.

(iv) The causal R^2 is 0 if the observable features do not vary: $\mathrm{Var}\left[X^O\right] = 0 \implies \mathrm{CR}^2_{\mathcal{F}}(X^O) = 0$.

(v) If $\mathcal{F}$ is well-specified, the causal R^2 is 1 if the unobservable features do not vary: $\mathrm{Var}\left[X^U\right] = 0 \implies \mathrm{CR}^2_{\mathcal{F}}(X^O) = 1$.

(vi) Define $Y' = Y + \varepsilon$, for ε unobserved mean-zero random noise independent of, and causally unaffected by, X. Then the causal R^2 for Y is larger in magnitude than the causal R^2 for Y': $|\mathrm{CR}^2_{\mathcal{F}}(Y \to X^O)| \geq |\mathrm{CR}^2_{\mathcal{F}}(Y' \to X^O)|$.

(vii) CR^2 is not symmetric. Let there be one observable. In general, $\mathrm{CR}^2_{\mathcal{F}}(Y \to X^O) \neq \mathrm{CR}^2_{\mathcal{F}}(X^O \to Y)$.

Parts (i) and (ii) are limiting cases. If the observed features X^O do not affect the outcome, they explain no variation: $\mathrm{CR}^2_{\mathcal{F}}(X^O) = 0$. If instead X^O fully determines the outcome—such that no variation in the outcome remains after setting X^O—then $\mathrm{CR}^2_{\mathcal{F}}(X^O) = 1$, if $\mathcal{F}$ is well-specified.[22]

Part (iii) states that, if the observed features X^O do not fully determine the outcome variable, $\mathrm{CR}^2_{\mathcal{F}}(X^O) < 1$. Consider a single, discrete observable feature X^O, and an experiment with a treatment arm for each possible level of X^O. If the outcome varies in at least one arm, an unobservable must cause that variation: the observable does not explain all variation.

Parts (iv) and (v) also describe limiting cases. If observables do not vary in the population, they explain no variation in the outcome. If only observables vary, they explain all variation.

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

[21] These properties would also hold if we had defined risk from another loss function.

[22] The "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 relation between Y and X is non-linear (Anscombe 1973).

Part (vii) says that CR^2 is not symmetric. Y may explain a certain share of X^O, but X^O might explain a different share of Y. This result 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 causally. Some alternatives do not satisfy them. For instance, the predictive R^2 violates (i), (iii), and (vii). Online Appendix A discusses other seemingly intuitive measures, and shows they lack one or more of the basic properties.

What is more, the CR^2 is invariant to some transformations. In our view, this is conceptually important as the share of variation explained by a variable should not change if the variable is measures in different units. Education explains the same share of variation in wages, no matter if wages are measured in dollars or cents, or if education is measured in years or days.

Proposition 2 (invariance to transformations). *(i) The* CR^2 *is invariant to affine transformations of the outcome, and monotone transformations of the features. That is, consider a sequence of functions* $\{g_Y, g_1, \dots, g_O\}$*, where* g_Y *is affine and strictly monotone, and each* $\{g_k\}_{k=1}^O$ *is strictly monotone. Let* $Y' = g_Y(Y)$ *and, for each* k, $X'_k = g_k(X_k)$*. Then* $CR^2(Y' \to X^{O'}) = CR^2(Y \to X^O)$.

(ii) The CR^2_{lin} *is invariant to affine transformations of the outcome and features. That is, consider a sequence of affine and strictly monotone functions* $\{g_Y, g_1, \dots, g_O\}$*. Let* $Y' = g_Y(Y)$ *and, for each observable* k, $X'_k = g_k(X_k)$*. Then,* $CR^2_{\text{lin}}(Y' \to X^{O'}) = CR^2_{\text{lin}}(Y \to X^O)$.

4.2. Comparison of predictive and causal R^2

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

Proposition 3 (relation between predictive and causal R^2**).** *For any function class* $\mathcal{F}$*, distribution of features* P_X*, and potential outcome function* $Y(\cdot)$*:*

(i) The causal R^2 *is bounded above by the predictive* R^2*:* $CR^2_{\mathcal{F}}(X^O) \leq R^2_{\mathcal{F}}(X^O)$.

(ii) If observables are independent of unobservables, the causal and predictive R^2 *coincide:* $X^O \perp\!\!\!\perp X^U \implies CR^2_{\mathcal{F}}(X^O) = R^2_{\mathcal{F}}(X^O)$.

Part (i) says the predictive R^2 always 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 in the CR^2. This result has practical use. Even without knowing the causal effect of variable, if it has little predictive power, it also has little causal explanatory power. Part (ii) says that the predictive and causal R^2 coincide when observables are independent of unobservables.[23] In that case, the best predictive model equals the best causal model as there is no confounding. Their fit must also equal.

[23] For the linear CR^2, orthogonality suffices.

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). *(i) For any $\mathcal{F}$, $R^2_{\mathcal{F}}$ is bounded between 0 and 1, whereas $CR^2_{\mathcal{F}}$ is bounded above by 1, but may be negative.*
(ii) For any $\mathcal{F}$, $R^2_{\mathcal{F}}$ increases as more features are observed, whereas $CR^2_{\mathcal{F}}$ may fall.

Part (i) says that CR^2 is bounded above by 1 and may be negative. This happens when the observables *suppress*, rather than create, variance in the outcome. Consider the introductory example. Suppose larger classes lower test scores while family income raises them. However, say poorer kids attend smaller classes, *e.g*, as a policy to reduce inequality. With large enough correlation between family income and class size, test scores and class size may be negatively correlated. However, causally, class size lowers test scores. The negative causal effect fits the positive association of test scores and class size worse than the baseline of no effect. Since fit is the proportional reduction in risk relative to a baseline of no effect, the CR^2 is negative.[24]

A negative CR^2 indicates that observed features suppress variation in the outcome: there "should" be more variation in the outcome than we actually 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.[25] 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.

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.[26]

4.4. Properties of CR^2_{lin}

The linear CR^2 is of applied interest because researchers often use linear models. It has a simple closed-form expression.

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

$$CR^2_{\text{lin}}(X^O) = R^2_{\text{lin}}(X^O) - (\tilde{\beta}^P_{X^O} - \tilde{\beta}^C_{X^O})^\top \rho_{X^O} (\tilde{\beta}^P_{X^O} - \tilde{\beta}^C_{X^O}),$$

where $\tilde{\beta}^P_{X^O}$ are the coefficients from an OLS regression of standardized Y on standardized X^O in observational data, $\tilde{\beta}^C_{X^O}$ is the corresponding vector from an OLS regression in an experiment that randomly assigns X^O is randomly assigned in proportion to its marginal distribution in the population

[24] Online Appendix Online Appendix B gives a detailed example.
[25] The decomposition originated in Kitagawa (1955), Blinder (1973), and Oaxaca (1973).
[26] Online Appendix E shows there are no other monotone measures of the share of variation causally explained.

(with the standardization again with respect to the observational mean and variance), and ρ_{X^O} the correlation matrix of observed features in observational data.

Hence, if the model is well-specified, the $\mathrm{CR}^2_{\mathrm{lin}}$ does not require microdata: it can be computed from standard summary statistics. A researcher who studies a given treatment and wishes to compare the explanatory power of that treatment to the explanatory power of another treatment studied in the literature may do so using the reported causal effects of that other treatment. She does not need the underlying microdata of the other study.

5. Identification, estimation, and inference

We now turn to estimation from finite data. We begin by describing the data available to the analyst, before discussing identification and estimation, and then turning to inference.

5.1. Data setting

The CR^2 is defined using the joint distribution of the outcome and observable features, and the potential outcome function, which pins down the best causal model. The former is easy to obtain from observational data, but the latter requires additional information: the interventional expectation is not identified from observational data without further assumptions. For simplicity, we consider estimating the interventional expectation through a randomized experiment, but our method extends to other identification strategies.

The analyst has access to two samples: an observational sample (O) and an experimental sample (E).[27] Following Athey et al. (2025b), 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.

The observational sample consists of N_O realizations (Y_i, X_i^O) drawn i.i.d from the joint distribution of the outcome and observable features, P_{Y,X^O}. This sample identifies this joint distribution, but not the interventional expectation. For the latter, the analyst relies on an experimental dataset, constructed according to a known experiment (Rubin 1978).

Definition 5. An ***experiment*** $E \coloneqq (x_t^O, \Pr(x_t^O))_{t=1}^T$ 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 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 a sample of size N_E, independently assigns each unit to a treatment arm via

[27] This is an example of a "data fusion" setting (Rässler 2012; Bareinboim and Pearl 2016), also called "auxiliary data" (Hellerstein and Imbens 1999; Chen et al. 2008) "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. 2025b), and correct for biases in observational data (*e.g.*, Kallus et al. 2018; Athey et al. 2025a).

the assignment mechanism. She sets the unit's observed features according to their treatment arm, and observes the resulting outcome. Given treatment x_t^O, the outcome $Y_i(x_t^O, X^U)$ depends on the random variable X^U and so is random; denote its distribution by $P^E_{Y|x_t^O}$.

Formally, the analyst draws a sample (S_i, X_i^O, Y_i) of size N 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 sources of data. The last two columns indicate the advantages and drawbacks of each sample. The observational sample is drawn directly from the population, and hence its feature distribution match the population's, but the observed features may not be independent of the unobservables, preventing identification of the best causal model. The experimental sample has the opposite structure: randomization ensures $X_i^O \perp\!\!\!\perp X_i^U \mid S_i = E$, allowing identification of causal effects, but its feature distribution generally differs from the population's. Indeed, even its marginal feature distribution will generally differ: the assignment mechanism is typically chosen to maximize precision, not to match the population distribution (Neyman 1934; Duflo et al. 2007; Athey and Imbens 2017).

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?	Drawn from pop. dist. P?	Random assignment? ($X_i^O \perp\!\!\!\perp X_i^U$)
Obs. sample	O	√	√	×	√	×
Exp. sample	E	√	√	×	×	√

Note: *This table summarizes the observational and experimental samples. The first row describes the observational sample; the second row describes the experimental sample.*

When might this data combination arise? One possibility is that the analyst conducts an experiment on a given population, and separately draws an observational sample from that population, *e.g.*, an initial observational study informs a subsequent randomized intervention. This is common: Epanomeritakis and Viviano (2025) report that over 30% of experimental papers in AEA journals over the last decade also include observational evidence. A second possibility is that the experiment includes a "no-treatment" or "status quo" arm, which may be treated as observational data. Finally, if causal effects are identified from quasi-random variation in observational data, the random variation identifies the causal model, and the observational data more generally can be used to assess its goodness-of-fit.

5.2. Identifying CR^2

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

Proposition 6. *Fix an outcome Y, a vector of observable features X^O.*

(i) For any $\mathcal{F}$, $\mathrm{CR}^2_{\mathcal{F}}(X^O)$ is not identified by an observational sample or an experimental sample.

(ii) The non-parametric causal R^2, $\mathrm{CR}^2(X^O)$, 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, $\mathrm{CR}^2_{lin}(X^O)$, 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 disadvantages 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 insufficient.

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 limited. For instance, with a single, binary observed feature, any non-trivial experiment suffices; with several observed features, each with finite support, a factorial experiment would be required (see, *e.g.*, Mukerjee and Wu 2007).

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 $\mathrm{CR}^2_{\mathrm{lin}}$. On the other hand, if she simply uses linearity as a convenient functional form, then she can understand $\mathrm{CR}^2_{\mathrm{lin}}$ 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 6. 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*** $\widehat{\mathrm{CR}}^2_{\mathcal{F}}(X^O)$ as:

$$\widehat{\mathrm{CR}}^2_{\mathcal{F}}(X^O) \coloneqq 1 - \frac{\widehat{\mathcal{R}}(\hat{Y}^C_{\mathcal{F},X^O})}{\widehat{\mathrm{Var}}[Y]}, \qquad \text{where } \hat{Y}^C_{\mathcal{F},X^O} \coloneqq \arg\min_{M\in\mathcal{F}} \frac{1}{N_E} \sum_{i:s_i=E} (y_i - M(x_i^O))^2,$$

$$\hat{\mathcal{R}}(M) \coloneqq \frac{1}{N_O} \sum_{i:s_i=O} (y_i - M(x_i^O))^2, \qquad \widehat{\mathrm{Var}}[Y] \coloneqq \frac{1}{N_O} \sum_{i:s_i=O} (y_i - \bar{y}_O)^2, \qquad \bar{y}_O \coloneqq \frac{1}{N_O} \sum_{i:s_i=O} y_i.$$

The plug-in estimator replaces each quantity in the definition of $\mathrm{CR}^2_{\mathcal{F}}$ with its finite-sample counterpart.[28] Since each component of $\widehat{CR}^2_{\mathcal{F}}(X^O)$ converges to its population counterpart, applying Slutsky's Theorem shows Proposition 7.

Proposition 7. *Suppose the conditions for identification in parts (ii) and (iii) of Proposition 6 are*

[28] When the analyst's sample is a combination of observational and experimental data, each of these expressions is well-defined with probability one as the sample grows large. In particular, for $p \in (0,1)$ $N_E > 0$ and $N_O > 0$ with probability one for N large, so the fractions are well-defined. The arg min is unique in the case of a non-parametric model with a full-support experiment, or a linear model with a full-rank experiment.

satisfied. Then the plug-in estimator is a consistent estimator for $\mathrm{CR}^2_{\mathcal{F}}(X^O)$.

Nonetheless, $\widehat{\mathrm{CR}}^2_{\mathcal{F}}$ will generally be downward-biased in small samples, since estimation error in the best causal model will tend to inflate the model's risk. In Section 6, we present several simulations in which this bias vanishes reasonably quickly as the sample size grows.

One can conduct inference for $\widehat{\mathrm{CR}}^2_{\mathcal{F}}$ using 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 C. The Delta method approach is convenient in the case of $\widehat{\mathrm{CR}}^2_{\mathrm{lin}}$, and begins by writing the plug-in estimator in terms of other sample quantities:

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

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 $\widehat{\mathrm{CR}}^2_{\mathrm{lin}}$.

Outside the linear case, it is more convenient to construct bootstrapped standard errors, randomly sampling with replacement from both the observational and experimental samples. For details on this approach and the Delta method, see Online Appendix C.

5.4. Measurement error

In practice, the analyst may measure the features or outcome with error. We now turn to how measurement error affects the CR^2. We focus on the linear CR^2 in the case of a single observed feature, and restrict attention to classical measurement error: we say there is ***classical measurement error*** in $Z \in \{X^O, Y\}$ if the observed value is $Z' = Z + \varepsilon$ with ε uncorrelated with (X^O, Y) and $\mathbb{E}[\varepsilon] = 0$. There are four cases to consider: whether noise appears in the outcome or feature; and whether noise appears in the experimental or observational data.[29]

Proposition 8 (measurement error)**.** *Say there is a single observable feature (*$O = 1$*). Denote by* $\mathrm{CR}^2_{\mathrm{CME}}$ *the linear* CR^2 *when there is classical measurement error (CME), and by* CR^2 *the true value.*

(i) CME in the outcome in the observational data attenuates the estimated CR^2*: in particular,* $\mathrm{CR}^2_{\mathrm{CME}} = \frac{\mathrm{Var}[Y]}{\mathrm{Var}[Y]+\mathrm{Var}[\varepsilon]} \times \mathrm{CR}^2$.

(ii) CME in the feature in the observational data reduces the estimated CR^2*:* $\mathrm{CR}^2_{\mathrm{CME}} = \mathrm{CR}^2 - \beta^2 \frac{\mathrm{Var}[\varepsilon]}{\mathrm{Var}[Y]}$*, where* β *is the slope of the true linear interventional expectation.*

(iii) CME in the outcome in the experiment does not affect the estimated CR^2*:* $\mathrm{CR}^2_{\mathrm{CME}} = \mathrm{CR}^2$.

(iv) CME in the feature in the experiment can increase or reduce the estimated causal R^2*, though it is still bounded above by the* R^2 *in observational data.*

[29] It may be surprising to consider measurement error in the experimentally-assigned feature. This would occur, *e.g.*, if treatment status is sometimes recorded incorrectly, or if there is non-compliance.

Proposition 8 can be used to sign the likely bias from measurement error. The proposition also motivates correcting CR^2 for measurement error, if the analyst is willing to impose further structure. For instance, suppose the analyst observes two noisy measurements of each unit's outcome, $Y'_{i,1} = Y_i + \epsilon_{i,1}$ and $Y'_{i,2} = Y_i + \epsilon_{i,2}$, where $(\epsilon_{i,1}, \epsilon_{i,2})$ are mean-zero noise terms, independent of one another, the outcome, and the features. In that case, it is well-known (Spearman 1904; Griliches 1974) that the "raw predictive R^2" (R^2_{raw}) between Y' and X^O is attenuated relative to the "signal predictive R^2" (R^2_{signal}) between Y and X^O, and that the the analyst can recover R^2_{signal} by dividing the observed R^2_{raw} by the reliability $r := \text{Corr}[Y'_{i,1}, Y'_{i,2}]$.[30] Following this logic, in the well-specified case, we can write $CR^2_{\text{signal}} = CR^2_{\text{raw}}/r$, and hence correct for measurement error in the outcome.[31]

5.5. Limitations of our data setting

We conclude with two limitations of our data setting. The first concerns external validity. Our definition of an experiment requires (i) random sampling from the population ("external validity") and (ii) random assignment to treatment arms, conditional on inclusion in the experiment ("internal validity"). Some randomized controlled trials plausibly satisfy both conditions. However, many experiments satisfy (ii) but not (i). For instance, in a standard instrumental variables design with a binary treatment, the instrument randomizes treatment only among "compliers", identifying a local average treatment effect (LATE) rather than the population average treatment effect (ATE) (Imbens and Angrist 1994; Angrist et al. 1996).[32] Even in some randomized trials, the experiment is conducted on a subpopulation. Hence, our baseline setting excludes some situations of practical interest.

Without further assumptions, such an experiment will not generally identify the CR^2, even when combined with observational data. To see why, suppose for example that the feature of interest has a large causal effect for one subpopulation, but no causal effect for another subpopulation, and that only variation in the second subpopulation is used in the experiment. In that case, the true causal CR^2 in the general population will differ from zero, whereas the experiment would suggest that the feature does not explain any variation.

This difficulty in generalizing causal effects from subpopulations is well-known. When the experimental subpopulation is itself of interest, the analyst can simply report the share of variation explained in that subpopulation. Alternatively, she can compare the observable characteristics of the experimental and target populations to assess external validity. In a randomized experiment, the experimental subpopulation is observed directly; in instrumental

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

[31] Similarly, in some applications, the outcome is an estimate computed by the analyst *with known error*—for example, when explaining variation in neighborhoods' upward mobility or teacher value-added, which are estimated with noise (*e.g.*, Kane and Staiger 2008; Chetty et al. 2014; Angrist et al. 2017; Chetty et al., forthcoming).

[32] Similarly, when assignment is randomized within strata Z, the coefficient on X in a regression of Y on X and Z identifies a variance-weighted average of stratum-specific ATEs, with weights proportional to $\text{Var}(X \mid Z)$ (see, *e.g.*, Angrist 1998).

variables designs, Abadie's (2003) weighting theorem identifies the distribution of complier covariates. If these distributions are similar, the analyst may have more confidence in generalizing her effects; if they differ, the analyst can compute covariate-specific LATEs and reweight them to recover the population ATE, under the assumption that effect heterogeneity is fully captured by observables (Angrist and Fernández-Val 2013; Hartman et al. 2015). Finally, structural approaches based on parametric latent-index models can extrapolate causal effects beyond compliers (see, *e.g.*, Heckman et al. 2001, 2003; Angrist 2004; Heckman 2010). Our main contribution is to show the usefulness of CR^2 for an externally-valid experiment, although these approaches can extend our method to other settings.

A second limitation that we have assumed features are either (i) "observable" in the sense that they are observed in the observational dataset and manipulated in the experiment, or (ii) neither observed nor manipulated. In practice, there is a third class, which we call "covariates": features which are observed, but not manipulated—either because the features are inherently difficult to manipulate (*e.g.*, they involve sex or race) or because they are not the main focus of the experiment. In the main text, we consider the simple case without covariates, but we show how they can be incorporated into the analysis in Online Appendix D.

6. Simulations

We now present several simulations. We have two goals in doing so: first, to gain familiarity with the measure by displaying it in some simple settings; and, second, to examine the performance of the plug-in estimator.

Simulation 1: independent features in a well-specified linear model. We begin with the simplest non-trivial setting, in which the potential outcome function is linear and the feature variables are independent of one another. Say there are two features, only one of which is observed ($K = 2, O = 1$), with true data-generating process:

$$Y(X_1, X_2) = \beta_1 X_1 + X_2 \tag{3}$$

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim \mathcal{N}(\mathbf{0}, \Sigma), \quad \Sigma = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \tag{4}$$

where (3) is the (linear) potential outcome function, and (4) is the joint distribution of (X_1, X_2), which jointly pin down the distribution of Y. Our simulation varies the causal effect of the observed feature (β_1).

Our results are summarized in Panel A of Figure 2. We begin by examining the true $CR^2(X_1)$, illustrated by the purple line. Since the data-generating process is linear, this coincides with $CR^2_{\text{lin}}(X_1)$. Moreover, since the observed and unobserved features are independent, this also coincides with the standard (predictive) R^2. When X_1 has no effect on the outcome ($\beta_1 = 0$), the causal R^2 is equal to 0; as the causal effect increases (in magnitude), the share of variance

explained grows. When X_1 has the same causal effect in magnitude as the unobserved feature ($|\beta_1 = 1$), X_1 explains half of the total variance in the outcome, and so $\text{CR}^2(X_1) = 1/2$. As the causal effect of X_1 grows, the share of variation it explains approaches 1.

To examine the plug-in estimator's performance, we specify several additional parameters. We imagine the analyst collects a sample of size N, of which three-fifths of units are in the observational sample, and the remaining two-fifths in the experimental sample ($p = 0.6$). The experiment consists of two equally-probable treatment arms, one assigning units to a treatment value of $X_1 = 0$, and the other to $X_1 = 1$. The navy dots show the performance of the estimator when the full sample size is $N = 200$; we perform $1,000$ simulations and present the mean estimate. The estimator displays a downward bias which falls fairly quickly as the number of units increases, as the light blue ($N = 500$) and orange ($N = 2,000$) series show.

Simulation 2: correlated features in a well-specified linear model. For our second simulation, we maintain the linear model, but allow the features to be correlated. As before, there are two features, one of which is observed, with true data-generating process:

$$Y(X_1, X_2) = X_2 \tag{5}$$

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim \mathcal{N}(\mathbf{0}, \Sigma), \quad \Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}, \tag{6}$$

where (5) is the (linear) potential outcome function, and (6) is the joint distribution of (X_1, X_2). Relative to the first simulation, we fix the first feature to have no causal effect ($\beta_1 = 0$), but vary the correlation between features (ρ).

Panel B summarizes our results. As before, we begin by showing the true $\text{CR}^2(X_1)$ (which again coincides with $\text{CR}^2_{\text{lin}}(X_1)$). The CR^2 is always equal to zero, since X_1 does not have any effect on Y, and so causally explains none of the variation. By contrast, the predictive R^2 is strictly positive when $\rho \neq 0$: X_1 does have some predictive power due to its correlation with X_2. We use the same parameters as in the previous simulation to assess the performance of our plug-in estimator. As before, the plug-in estimator shows a finite-sample downward bias that vanishes reasonably quickly in the number of observations.

Simulation 3: correlated features in a misspecified model. Finally, we introduce misspecification. As before, there are two features, one of which is observed; but now we introduce a non-linearity into the true potential outcome function:

$$Y(X_1, X_2) = 0.2X_1 + \gamma X_1^2 + X_2 \tag{7}$$

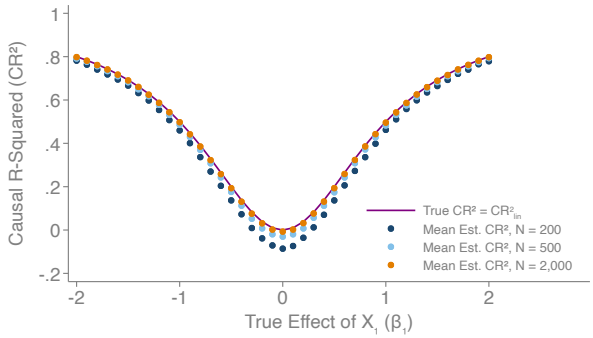
$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim \mathcal{N}(\mu, \Sigma), \quad \mu = \begin{pmatrix} 5 \\ 0 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}. \tag{8}$$

As part of the simulation, we vary the value of the quadratic term in the effect of X_1 on Y (γ).

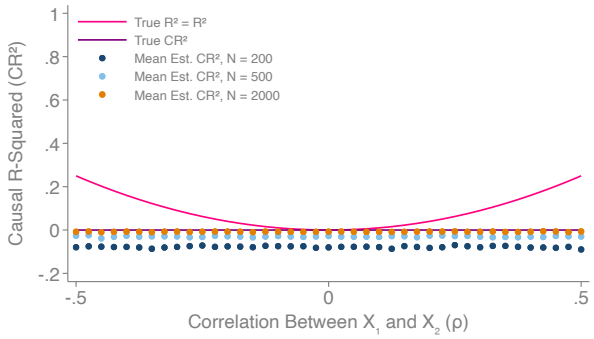
The purple line in Panel C shows the true $\text{CR}^2(X_1)$. Since the true potential outcome function is quadratic, this coincides with $\text{CR}^2_{\text{quad}}(X_1)$, where quad denotes the class of quadratic models ($x_1 \to \mu + \nu x_1 + \pi x_1^2$). As the quadratic term (γ) increases, the true share of variation in Y explained by X_1 also increases, and approaches 1 for large values of γ.

Figure 2. Results of simulations

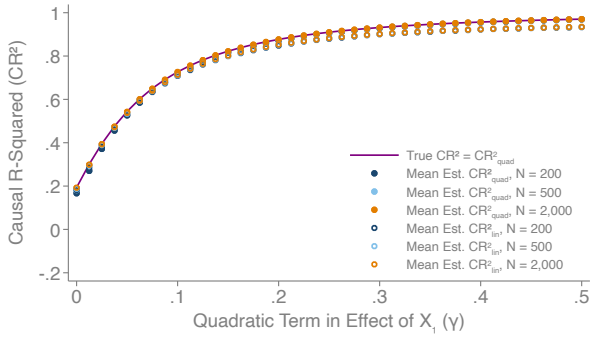
(A) Results of simulation 1



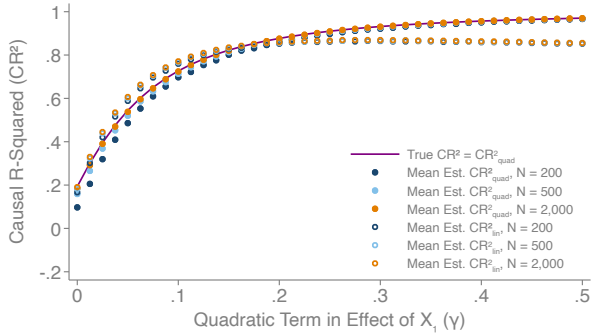
(B) Results of simulation 2



(C) Results of simulation 3: experiment (a)



(D) Results of simulation 3: experiment (b)



Note: *This figure presents results from our simulations. The purple line in Panel (A) displays the true* CR^2 *in simulation 1; since the data generating processing is linear, this coincides with* CR^2_{lin}. *The navy dots show the mean estimated* CR^2 *among 2,000 simulations, when the sample size is 200. The light blue and orange dots replicate the navy dots, for sample sizes of 500 and 2,000, respectively. Panel (B) replicates Panel (A) for Simulation 2. Panel (C) replicates Panel (A) for the first experiment in Simulation 3, and Panel (D) for the second experiment.*

We then turn to estimating $\text{CR}^2(X_1)$ using a combination of observational and experimental data. We consider two experiments; in both cases, to allow for the possibility of estimating a quadratic model, the experiments have three treatment arms. In experiment (a) (Panel C), we suppose the experimental sample is evenly divided between being assigned the mean value of X_1, or one standard deviation above or below. The solid circles in Panel C display the mean estimates if the analyst specifies a quadratic model; as before, there is a small, downward finite-sample bias that vanishes reasonably quickly as the number of observations grows.

Now suppose the analyst estimates a misspecified linear model (the hollow circles in Panel C). In this case, the plug-in estimator may not converge to the true CR^2. In practice, the degree of

divergence in Panel C is reasonably small. Intuitively, although the true model is quadratic, a linear analysis of the experiment around the mean value of X_1 recovers an average treatment effect that approximates the true quadratic effect.

In Panel D, we consider an alternative experimental design in which the experimental sample is evenly divided between being two, three, or four standard deviations above the mean value of X_1. As before, the estimates produced from the quadratic model (in solid circles) converge to the true CR^2. In this case, however, the estimates produced from the linear model (in hollow circles) are quite different from the true CR^2. This is because the linear model estimates a local average treatment effect among the treatment arms in the experiment; as a result, the fit of the causal model estimated from the experiment is worse when the treatment arms are further away from the mass of the distribution of the observables.

We draw three conclusions from the simulations. First, in each case, the CR^2 captures an intuitive notion of the share of variation explained. Second, the downward bias of $\widehat{CR}^2_{\mathcal{F}}$ vanishes quickly. Third, misspecification can cause $\widehat{CR}^2_{\mathcal{F}}$ not to converge to the non-parametric $\widehat{CR}^2$, though the degree of difference seems reasonably small at least in the simulations.

7. Applications of the CR^2

We illustrate our measure in five settings, summarized in Table 2. We choose the settings both for substantive interest, and to show extensions to the baseline data setting in section 5.

Table 2. Summary of applications

	Population	Outcome (Y)	Observed cause of interest (X^O)	Extensions relative to baseline data setting
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	IV design to estimate causal effect
3.	Former colonies (Acemoglu et al., 2001)	GDP per capita	Expropriation risk	IV design to estimate causal effect; no separation between experimental and observational samples
4.	U.S. cities (Fluegge, 2025)	City population	Exposure to influenza	IV design to estimate causal effect; no separation between experimental and observational sample
5.	Adults in the U.S. and U.K. (DASH, NDNS)	Blood pressure	Sodium intake	Experimental and observational samples drawn from different populations

Note: *This table summarizes the five applications described in this section.*

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

We begin with an application that closely resembles our baseline data setting. Kremer et al. (2011) conduct a randomized controlled trial evaluating the effects of spring protection on water quality. The study takes place in the rural Busia and Butere-Mumias districts of Kenya's Western Province. 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). The primary measure of water quality is the *E. coli* level. We ask what share of variation in water quality across springs is causally explained by variation in spring protection.

The experimental sample consists of springs involved in the authors' experiment. The observational sample consists of nearby springs that were not involved in the experiment.[33] Protection status is binary. We assess the predictive relationship between water quality and spring protection through an OLS regression in the observational data of the form

$$\ln \text{E Coli}_s = \alpha^P + \beta^P \text{Protection}_s + \epsilon_s^P \tag{9}$$

where $\ln \text{E Coli}_s$ 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 $(\hat{\alpha}^P, \hat{\beta}^P)$ define our best predictive model, $\widehat{\ln \text{E Coli}}_s^P(\text{Protection}) = \hat{\alpha}^P + \hat{\beta}^P \text{Protection}$. We assess the causal relationship through a corresponding OLS regression in the experimental data:

$$\ln \text{E Coli}_s = \alpha^C + \beta^C \text{Protection}_s + \epsilon_s^C. \tag{10}$$

The estimated coefficients $(\hat{\alpha}^C, \hat{\beta}^C)$ define our best causal model, $\widehat{\ln \text{E Coli}}_s^C(\text{Protection}) = \hat{\alpha}^C + \hat{\beta}^C \text{Protection}_s$. Panel (A) of Table 2 presents the estimated values $(\hat{\alpha}^P, \hat{\beta}^P)$ and $(\hat{\alpha}^C, \hat{\beta}^C)$, 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 the best predictive and causal 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 3 show.

Kremer et al. (2011) note that there is substantial error in measuring the *E. coli* level. Proposition 8 suggests that this measurement error in the outcome (in both the observational and experimental samples) will tend to attenuate the true share of variation explained. Under the assumption that this measurement error is independent of both spring protection and the true water quality, we can correct for this attenuation by dividing the raw 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 3(A) display this signal CR^2, suggesting

[33] In Appendix Table 1, we show that the observational and experimental samples are similar on pre-treatment characteristics.

that variation in spring protection explains around a third of variation in water quality, in both causal and predictive senses.

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		
Const.	4.82	3.64
	(0.144)	(0.089)
Spring protection	-1.98	-1.47
	(0.273)	(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 exp. in pop. (obs. R^2)	16.42%	–
% of var. causally exp. in exp. (exp. R^2)	–	10.71%
% of var. causally exp. in pop. (CR^2)	15.38%	–
D. Hypothesis tests		
$CR^2 = 0$	0.000	

Note: *This table presents our analysis of the share of variation in water quality (as measured by E. Coli level) causally explained by spring protection. Panel (A) presents estimates of equations* (9) *and* (10)*, with corresponding standard errors. Panel (B) presents the mean squared error of the models estimated in (A). Panel (C) presents the share of variance explained.*

7.2. 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 et al. 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. (2008). The data include information on school IDs, students' class sizes in grades K–3, and math and reading test scores in grades 1–3. For ease of interpretation, we express scores in each grade in percentage terms, and then compute mean scores over grades 1–3, restricting the sample to students observed in each grade. Following previous literature, we

assess the causal effect of class size in a two-stage least-squares regression of the form:

$$\text{ClassSize}_i = \pi^C + \rho^C \text{Small}_i + \nu_i^C, \tag{11}$$

$$\text{Score}_{i,s} = \alpha_s^C + \beta_s^C \widehat{\text{ClassSize}}_i + \varepsilon_{i,s}^C, \tag{12}$$

where $\text{Score}_{i,s}$ is student i's score in subject s (reading or math), ClassSize_i is her class size, and Small_i is an indicator for being assigned to a small class. The coefficients from the second-stage regression, $(\hat{\alpha}_s^C, \hat{\beta}_s^C)$, define our estimated best causal model for test scores:

$$\widehat{\text{Score}}^C(\text{ClassSize}) = \hat{\alpha}_s^C + \hat{\beta}_s^C \text{ClassSize}. \tag{13}$$

Achilles et al. (2008) 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 treat these schools as our observational sample. We assess the predictive relation between students' test scores and class size using the regression:

$$\text{Score}_{i,s} = \alpha_s^P + \beta_s^P \text{ClassSize}_i + \epsilon_{i,s}^P. \tag{14}$$

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

$$\widehat{\text{Score}}_s^P(\text{ClassSize}) = \hat{\alpha}_s^P + \hat{\beta}_s^P \text{ClassSize}. \tag{15}$$

Finally, we assess the goodness-of-fit of the best predictive and causal models. Panel B of Figure 3 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 explained in a causal sense. The causal and predictive R^2 differ due to the difference in the experimental and observational regression coefficients, *i.e.,* omitted variable bias.

The estimates of goodness-of-fit are reasonably precise, which reflects the large number of students in both observational and experimental samples. For both reading and math scores, we reject that the predictive and causal R^2 are equal ($p < 0.001$ in both cases). In the case of 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.

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

Our next application studies the determinants of national income. Acemoglu et al. (2001) examine the effects of differences in institutions between countries on differences in national income.[34] The authors' main measure of institutional quality is an index of expropriation risk.

[34]Indeed, the authors explicitly describe their purpose as being about understanding the sources of naturally-occurring variation in incomes, rather than seeking to inform policy: the article's opening sentence asks, "What are the fundamental causes of the large differences in income per capita across countries?"

Instrumenting for expropriation risk using variation in the mortality rates of early European settlers, the authors show that expropriation has large effects on GDP per capita.[35]

We build on this analysis by asking what share of variation in national income is causally explained by differences in expropriation risk. We begin by estimating our best causal model. Replicating Acemoglu et al. (2001), we assess the causal effect of expropriation risk on national incomes in a two-stage least-squares regression of the form:

$$\text{ExpropriationRisk}_c = \pi^C + \rho^C \text{SettlerMortality}_c + \nu_c^C, \tag{16}$$

$$\text{GDP}_c = \alpha^C + \beta^C \widehat{\text{ExpropriationRisk}}_c + \varepsilon_c^C, \tag{17}$$

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 GDP_c is GDP per capita in c. The coefficients from the second-stage regression, (α^C, β^C) define our estimated best causal model:

$$\widehat{\text{GDP}}^C(\text{ExpropriationRisk}) = \alpha^C + \beta^C \text{ExpropriationRisk}.$$

We then evaluate the model's goodness-of-fit. Panel C of Figure 3 shows a scatterplot of countries' log GDP per capita *vs.* their average expropriation risk. The navy line shows the best causal model. Computing the mean squared error from the best causal model gives a causal R^2 of 0.19: that is, the results in Acemoglu et al. (2001) suggest that around one-fifth of variation in national income is causally explained by differences in institutions' expropriation risk between countries.[36] 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. 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.

7.4. Application 4: Share of variation in city growth explained by rainfall during the 1918 influenza pandemic

Our next application studies the determinants of city growth. Fluegge (2025) examines the long-run effects of the 1918 influenza pandemic on the population of American cities. Using a rainfall instrument for influenza exposure, Fluegge (2025) finds that the pandemic had effects on city population and GDP that persist until the present day.[37]

We build on this analysis by examining the share of variation in city growth that is explained

[35] This 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 examine only its implications for the share of variance in national incomes that is causally explained by expropriation risk.

[36] As Acemoglu et al. (2001) note, this only reflects one aspect of differences between institutions: ". In reality the set of institutions that matter for economic performance is very complex, and any single measure is bound to capture only part of the "true institutions"" (pp. 1385–1386).

[37] The idea behind the instrument is that rain drives people indoors, where the flu spreads more easily.

by differential exposure to rainfall in the early days of the influenza pandemic.[38] We begin by estimating our best causal model. Fluegge (2025) argues that rainfall in the early days of the pandemic is randomly assigned, conditional on rainfall in the corresponding calendar days over the period 1910-1927. For that reason, for each year $y \in \{1920, 1930, \dots, 2010\}$ we estimate the causal effect of rainfall in an OLS regression of the form:[39]

$$\Delta\text{Pop}_{c,y} = \alpha_y^C + \beta_y^C \text{RainDays}_c + \lambda_y' X_c + \epsilon_{c,y},$$

where $\Delta\text{Pop}_c, y$ is the change in city c's log population between 1910 and year y, RainDays_c is the number of days with at least 0.01 inches of rainfall over the first 30 days after the first recorded influenza case in city c, and X_c is a vector of pre-treatment controls, including the mean number of rainy days in the corresponding calendar days over the period 1910-1927.[40] We estimate the model on a sample of 43 large American cities described in Fluegge (2025).[41] The coefficients from this regression define our estimated best causal model:

$$\widehat{\Delta\text{Pop}}_{c,y}^C = \hat{\alpha}_y^C + \hat{\beta}_y^C \text{RainDays}_c.$$

We then evaluate the model's goodness-of-fit (Panel D). Our point estimate suggests almost 20% of city growth between 1910 and 2010 is causally explained by rainfall exposure in the early days of the influenza pandemic. The estimated share has been roughly constant in recent decades.[42] We conclude that the point estimate suggests that a large share of city growth can be explained by the historical fact of rainfall in the early days of the pandemic.

The graph also shows bootstrapped standard errors. Of course, given the small sample size, there is considerable uncertainty around this point estimate. This is unsurprising, given that the causal effect of rainfall itself is estimated reasonably imprecisely in Fluegge (2025).[43]

[38] In the case of an instrumental variables analysis, one natural "causal model" relates the instrument to the outcome using the "reduced form" relation between the instrument and outcome; another relates the endogenous regressor to the outcome, using a two-state least-squares regression to estimate the effect of the regressor on the outcome. We take the former route in this section, and the latter route in the preceding section.

[39] Fluegge (2025) studies the population *level*, controlling for the 1910 population. We use the *change vs.* 1910.

[40] Following Fluegge (2025), X_c consists of city population characteristics in 1910 (share white, share urban, share of white residents who are foreign-born, share age 6-20, share age 6-20 in school), long-term weather characteristics (mean precipitation days in September and October in the period 1910–1927, mean January temperature in the period 1900–1940), in addition to the mean number of days of rainfall in the calendar days over the period 1910–1927 that correspond to the first 30 days after the first recorded case of influenza in the city.

[41] Fluegge (2025) also analyzes a larger county-level dataset, where some car needs to be taken due to missing data problems. For simplicity, we restrict attention to the smaller city-level dataset.

[42] A natural placebo test is to evaluate what share of city growth between 1900 and 1910 is causally explained by rainfall; when performing that test, we find that only about 2% of city growth over that period is explained by rainfall, and we cannot reject that 0% of city growth is explained by rainfall.

[43] In light of our Proposition 1, if the 95% confidence interval for the estimated effect of the observed feature includes 0, then the 95% confidence interval for the CR^2 associated with this feature must also include 0. Since the causal effects of rainfall on city growth are marginally significant at the 95% level, it follows that the CR^2 can be no more than marginally significant at the 95% level.

7.5. Application 5: Share of variation in hypertension explained by sodium intake

Finally, we consider the causal R^2 in a stylised health setting.High blood pressure is a major cause of death in high-income countries through cardiovascular disease such as heart attacks and strokes. Excess salt consumption is considered a leading cause of high blood pressure (Institute of Medicine 2010). We investigate what share of variation in blood pressure is causally explained by salt consumption, and how its explanatory power differs by gender.

First, we collect data on the causal effect of salt on blood pressure from the DASH-Sodium experiment (Sacks et al. 2001), a randomized controlled trial which evaluated, separately, the effects of salt intake and a healthy eating plan ("DASH" diet) on blood pressure. The study recruited 412 US participants with normal, high-normal, or high blood pressure.[44] These people were randomised into a control group and six treatment groups. Each treatment was a combination of 1) a typical US diet *vs.* a healthy diet and 2) low, intermediate, or high salt levels (50, 100, and 150 mmol sodium per day respectively). Study staff prepared the food, and participants received all their meals and snacks at an outpatient clinic. After a two-week run-in period during which everyone ate a high-sodium control diet, participants followed their assigned treatment diet for 30 days. At the end of the month, researchers measured participants' blood pressure, which is the main outcome of the study.

In a typical U.S. diet, lowering salt from a high to low level reduces systolic blood pressure by 6.7mm Hg (Figure 1A in Sacks et al. 2001). The effect differs by gender: a 5.6mm Hg reduction among men *vs.* a 7.4mm Hg reduction among women (Figure 2A ibid.).[45]

To assess how well salt intake explains high blood pressure, we combine the causal effects with observational data from the UK National Diet and Nutrition Survey (University Of Cambridge, MRC Epidemiology Unit and NatCen Social Research 2021).[46] This long-running, nationally-representative study assesses the diets and nutritional status of people in the UK, including blood pressure measurements. In years 2008–2012, the study also collected data on urinary sodium, which approximates sodium intake well.[47] We follow a public health literature on sodium intake and blood pressure which essentially treats the U.S. and UK populations as transportable (Jones et al. 2020).

[44]The study required participants to have a diastolic blood pressure of 80–95 mm Hg and a systolic blood pressure of 120–160 mm Hg.

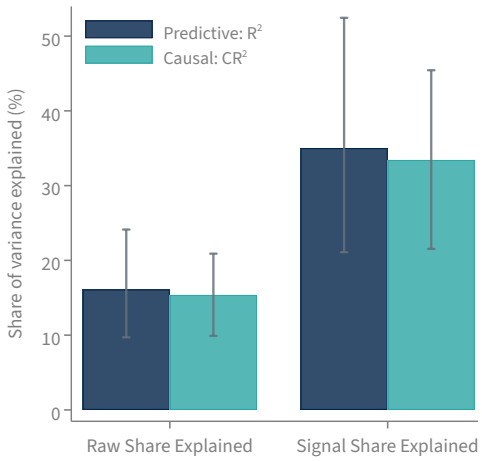
[45]Other studies have also reported a stronger blood pressure response to salt in women *vs.* men: *e.g.*, J. He et al. (2009); Bailey and Dhaun (2024); and the meta-analyses of F. J. He et al. (2013) and Huang et al. (2020).

[46]The data are available via the UK Data Service under study number 6533.

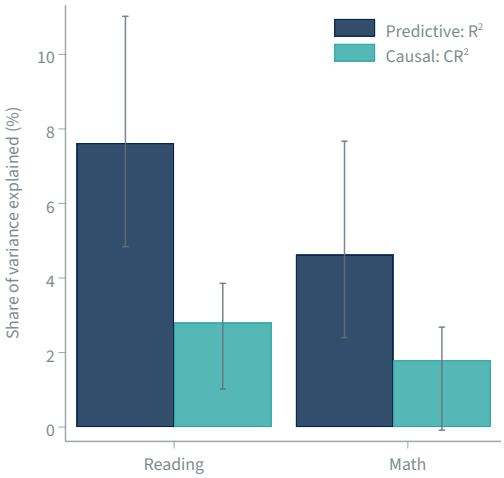
[47]Measuring sodium in urine collected over 24 hours, as done here, is the most accurate method to estimate salt intake (World Health Organization 2021) as about 93% of sodium is excreted through urine (Campbell et al. 2023). Other studies measuring salt consumption often rely on recall: they ask participants what foods they ate and estimate their sodium content. Those methods are substantially noisier (McLean et al. 2018).

Figure 3. Causal R^2 in Applications

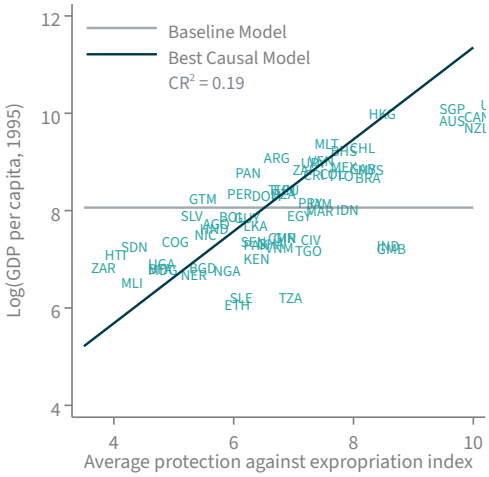
(A) Share of variance in water quality causally explained by spring protection



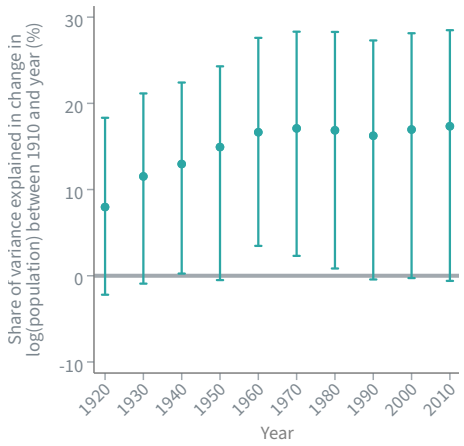
(B) Share of variance in test scores causally explained by class size



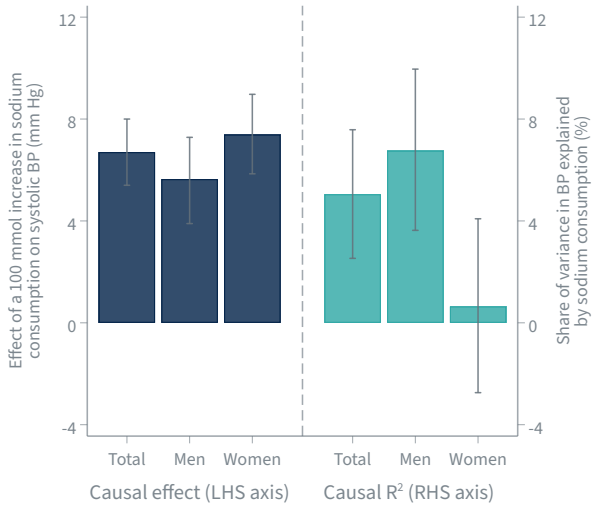
(C) Share of variance in national income causally explained by expropriation risk



(D) Share of variance in city growth causally explained by rainfall during the early days of the influenza pandemic



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



Note: *This figure summarizes the results in our five applications. Panel (A) shows the variance in water quality in Kenyan springs causally explained by spring protection. The set of bars on the right show the* R^2 *and* CR^2 *corrected for measurement error. Panel (B) shows the share of variance in test scores causally explained by class size. The navy bars show the share of variance in test scores predicted by class size (the standard predictive* R^2*), and the turquoise bars show the corresponding causal* R^2*. Panel (C) shows the share of variance in national income causally explained by institutions (expropriation risk). Panel (D) shows the share of variance in city population growth causally explained by variation in rainfall during the early days of the 1918 influenza pandemic. Panel (E) shows the share of variance in systolic blood pressure (mm Hg) causally explained by salt intake (100 mmol Na), as well as the causal effect of salt intake on blood pressure. The navy bars show the estimated causal effects of salt intake on blood pressure, in total, and separately for men and women. The turquoise bars show the share of variance causally explained by salt intake. Panels (A), (B), (D), and (E) show bootstrapped 95% confidence intervals.*

Salt consumption causally explains 5.1% of the variation in systolic blood pressure, and 6.8% of that variation in men. However, it explains substantially less in women: 0.6% (p-value for difference < 0.01). It may be surprising that the drop in explanatory power is higher in women than in men since the effect of sodium on blood pressure is similar, if anything somewhat larger, for women. That is because there is more omitted variable bias among women. One might have thought that once one takes into account the higher salt sensitivity among women, sodium causally explains a similar share of variation among men *vs.* women. However, the causal R^2 tells us that that reasoning is incorrect. In fact, salt explains blood pressure explains less among women than among men.

8. Conclusion

We argue that social scientists are interested not just in the causal effect of a given feature, but in how much of the population variance in the outcome is explained by that feature—a question left unanswered by causal effects alone as well as by traditional goodness-of-fit measures. We recast this question as the goodness-of-fit of the causal model generated by an experiment. The natural measure of goodness-of-fit—the "causal R^2"—has a share of variance explained interpretation that is analogous to the predictive R^2. The CR^2 is identified by a combination of experimental and observational data, and has a simple, consistent plug-in estimator. We illustrate the usefulness of the measure in applications to development, institutions, urban economics, education, and health.

We conclude with three limitations of our approach. First, a line of literature in statistics criticizes the usefulness of the predictive R^2.[48] Part of this criticism (*e.g.*, King 1986, 1991) argues that the R^2 is incorrectly interpreted causally; our alternative measure, CR^2, can be seen as a response to this criticism. There are other lines of criticism which are not resolved by our measure.[49] Nonetheless, since R^2 remains the primary way to assess the share of variation explained, we consider it valuable to address one deficiency. Moreover, if one accepts our claim that the share of variation explained can be recast as the goodness-of-fit of a causal model, then it is natural to develop alternative measures of causal goodness-of-fit.

Second, the relevant causal model may depend on the appropriate "distance" from the outcome. For instance, return to our motivating example of class size. Suppose the class size is a deterministic (decreasing) function of the local government's education budget. If the analyst studies the effect of class size on test scores, she will find the same share of variation in test scores is explained by class size, or by the size of the education budget. As such, even a causal R^2 of 1 does not rule out that there are other causal models that also fully explain the outcome. Nonetheless, the usefulness of the two models will depend on the analyst's goal.[50]

[48] For examples, see Draper (1984), Healy (1984), Kvålseth (1985), King (1986, 1991), and Scott and Wild (1991).

[49] For instance, Draper (1984) objects to the specific use of R^2 as a measure of proportional goodness-of-fit; Healy (1984) objects to the usage of proportional measures more broadly.

[50] Luskin (1991, p. 1038) makes this point well: "A given y can always be explained in a number of equally valid ways—in terms of a larger set of conceptually finer x's or a smaller set of conceptually grosser ones, in terms of

Finally, 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.

References

Abadie, Alberto. 2003. "Semiparametric instrumental variable estimation of treatment response models." *Journal of Econometrics* 113 (2): 231–263.

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, Charles 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." *Harvard Dataverse* 1:2008.

Albouy, David Y. 2012. "The colonial origins of comparative development: an empirical investigation: comment." *American Economic Review* 102 (6): 3059–3076.

Andrews, Isaiah, and Emily Oster. 2019. "A simple approximation for evaluating external validity bias." *Economics Letters* 178:58–62.

Angrist, Joshua D. 1998. "Estimating the Labor Market Impact of Voluntary Military Service Using Social Security Data on Military Applicants." *Econometrica* 66 (2): 249–288. ISSN: 00129682, 14680262.

———. 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., Peter D Hull, Parag A Pathak, and Christopher R Walters. 2017. "Leveraging lotteries for school value-added: Testing and estimation." *The Quarterly Journal of Economics* 132 (2): 871–919.

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.

Angrist, Joshua D., and Jörn-Steffen Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion.* Princeton, NJ: Princeton University Press.

Anscombe, Francis J. 1973. "Graphs in Statistical Analysis." *The American Statistician* 27 (1): 17–21.

Apesteguia, 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. 2025a. *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. 2025b. "The Surrogate Index: Combining Short-Term Proxies to Estimate Long-Term Treatment Effects More Rapidly and Precisely." Forthcoming, *Review of Economic Studies.*

variables that have their effect at close quarters or variables that act from afar."

Athey, Susan, and Guido Imbens. 2017. "The econometrics of randomized experiments." In *Handbook of economic field experiments,* 1:73–140. Elsevier.

Bailey, Matthew A., and Neeraj Dhaun. 2024. "Salt Sensitivity: Causes, Consequences, and Recent Advances." *Hypertension* 81, no. 3 (March): 476–489. ISSN: 0194-911X, 1524-4563. https://doi.org/10.1161/HYPERTENSIONAHA.123.17959.

Bareinboim, Elias, and Judea Pearl. 2016. "Causal inference and the data-fusion problem." *Proceedings of the National Academy of Sciences* 113 (27): 7345–7352.

Bertrand, Marianne, and Sendhil Mullainathan. 2004. "Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination." *American Economic Review* 94 (4): 991–1013.

Blau, Francine D, and Lawrence M Kahn. 2017. "The gender wage gap: Extent, trends, and explanations." *Journal of Economic Literature* 55 (3): 789–865.

Blinder, Alan S. 1973. "Wage discrimination: reduced form and structural estimates." *Journal of Human Resources,* 436–455.

Breza, Emily, and Arun G Chandrasekhar. 2019. "Social networks, reputation, and commitment: evidence from a savings monitors experiment." *Econometrica* 87 (1): 175–216.

Campbell, Norman R. C., Paul K. Whelton, Marcelo Orias, Laura L. Cobb, Erika S. W. Jones, Renu Garg, Bryan Williams, Nadia Khan, Yook-Chin Chia, Tazeen H. Jafar, and Nicole Ide. 2023. "It is strongly recommended to not conduct, fund, or publish research studies that use spot urine samples with estimating equations to assess individuals' sodium (salt) intake in association with health outcomes: a policy statement of the World Hypertension League, International Society of Hypertension and Resolve to Save Lives." Open access under CC BY-NC-ND 4.0, *Journal of Hypertension* 41, no. 5 (May): 683–686. https://doi.org/10.1097/HJH.0000000000003385.

Card, David. 1999. "The causal effect of education on earnings." *Handbook of Labor Economics* 3:1801–1863.

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 Hendren, Maggie R Jones, and Sonya R Porter. Forthcoming. "The opportunity atlas: Mapping the childhood roots of social mobility." *American Economic Review.*

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.

Chetty, Raj, John N Friedman, and Jonah E Rockoff. 2014. "Measuring the impacts of teachers II: Teacher value-added and student outcomes in adulthood." *American Economic Review* 104 (9): 2633–2679.

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.

Doksum, Kjell, and Alexander Samarov. 1995. "Nonparametric estimation of global functionals and a measure of the explanatory power of covariates in regression." *The Annals of Statistics,* 1443–1473.

Draper, Norman R. 1984. "The Box-Wetz Criterion Versus R2." *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." *Handbook of Development Economics* 4:3895–3962.

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.

Epanomeritakis, Aristotelis, and Davide Viviano. 2025. *Choosing What to Learn: Experimental Design when Combining Experimental with Observational Evidence.* arXiv: 2510.23434 [econ.EM].

Fluegge, Robert B. 2025. "Death, Destruction, and Growth in Cities: Entrepreneurial Capital and Economic Geography After the 1918 Influenza." *Available at SSRN 4173228.*

Fudenberg, Drew, Jon Kleinberg, Annie Liang, and Sendhil Mullainathan. 2022. "Measuring the completeness of economic models." *Journal of Political Economy* 130 (4): 956–990.

Ganong, Peter, and Pascal Noel. 2023. "Why do borrowers default on mortgages?" *The Quarterly Journal of Economics* 138 (2): 1001–1065.

Gelman, Andrew, Ben Goodrich, Jonah Gabry, and Aki Vehtari. 2019. "R-squared for Bayesian regression models." *The American Statistician.*

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.

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.

Hall, Robert E, and Charles I Jones. 1999. "Why do some countries produce so much more output per worker than others?" *The Quarterly Journal of Economics* 114 (1): 83–116.

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.*

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 R 2: estimation and inference." *The American Statistician* 78 (1): 15–25.

He, Feng J., Jiafu Li, and Graham A. MacGregor. 2013. "Effect of longer-term modest salt reduction on blood pressure." Systematic review, Article No. CD004937, *Cochrane Database of Systematic Reviews* (4). ISSN: 1465-1858. https://doi.org/10.1002/14651858.CD004937.pub2.

He, Jiang, Dongfeng Gu, Jing Chen, Cashell E Jaquish, Dabeeru C Rao, James E Hixson, Ji-chun Chen, Xiufang Duan, Jian-feng Huang, Chung-Shiuan Chen, Tanika N Kelly, Lydia A Bazzano, and Paul K Whelton. 2009. "Gender Difference in Blood Pressure Responses to Dietary Sodium Intervention in the GenSalt Study." *Journal of Hypertension* 27, no. 1 (January): 48–54. ISSN: 0263-6352. https://doi.org/10.1097/HJH.0b013e328316bb87.

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/jrsssa/article-pdf/147/4/608/49757327/jrsssa_147_4_608.pdf.

Heckman, James, Justin L Tobias, and Edward Vytlacil. 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. 2006. "Instruments for Causal Inference: An Epidemiologist's Dream?" *Epidemiology* 17 (4). ISSN: 1044-3983.

———. 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.

Holland, Paul W. 1986. "Statistics and causal inference." *Journal of the American Statistical Association* 81 (396): 945–960.

Huang, Liping, Kathy Trieu, Sohei Yoshimura, Bruce Neal, Mark Woodward, Norm R C Campbell, Qiang Li, Daniel T Lackland, Alexander A Leung, Cheryl A M Anderson, Graham A MacGregor, and Feng J He. 2020. "Effect of dose and duration of reduction in dietary sodium on blood pressure levels: systematic review and meta-analysis of randomised trials." *BMJ* 368 (February). https://doi.org/10.1136/bmj.m315. eprint: https://www.bmj.com/content/368/bmj.m315.full.pdf.

Hull, Peter. 2025. *"One Weird Trick"' to Characterize Effective Populations in Design-Based Specifications.* Technical report. memo.

Imbens, Guido. 2010. "Better LATE than nothing: Some comments on Deaton (2009) and Heckman and Urzua (2009)." *Journal of Economic Literature* 48 (2): 399–423.

———. 2014. "Instrumental Variables: An Econometrician's Perspective." *Statistical Science* 29 (3): 323–358. ISSN: 08834237, 21688745.

Imbens, Guido, and Joshua D. Angrist. 1994. "Identification and Estimation of Local Average Treatment Effects." *Econometrica* 62 (2): 467–475. ISSN: 00129682, 14680262.

Imbens, Guido, and Donald B. Rubin. 2015. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction.* Cambridge: Cambridge University Press. ISBN: 978-0-521-88588-1.

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.

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." *Advances in neural information processing systems* 31.

Kane, Thomas J, and Douglas O Staiger. 2008. *Estimating teacher impacts on student achievement: An experimental evaluation.* Technical report. National Bureau of Economic Research.

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.

Kitagawa, Evelyn M. 1955. "Components of a difference between two rates." *Journal of the American Statistical Association* 50 (272): 1168–1194.

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.

Kvålseth, Tarald O. 1985. "Cautionary note about R 2." *The American Statistician* 39 (4): 279–285.

Lauritzen, Steffen L., and Thomas S. Richardson. 2002. "Chain Graph Models and their Causal Interpretations." *Journal of the Royal Statistical Society Series B: Statistical Methodology* 64, no. 3 (August): 321–348. ISSN: 1369-7412. https://doi.org/10.1111/1467-9868.00340. eprint: https://academic.oup.com/jrsssb/article-pdf/64/3/321/49721907/jrsssb_64_3_321.pdf.

Li, Gang, and Xiaoyan Wang. 2019. "Prediction accuracy measures for a nonlinear model and for right-censored time-to-event data." *Journal of the American Statistical Association.*

Luskin, Robert C. 1991. "Abusus non tollit usum: standardized coefficients, correlations, and R 2s." *American Journal of Political Science,* 1032–1046.

Manski, Charles F. 2011. "Genes, eyeglasses, and social policy." *Journal of Economic Perspectives* 25 (4): 83–94.

McArthur, John W, and Jeffrey D Sachs. 2001. *Institutions and geography: comment on Acemoglu, Johnson and Robinson (2000).*

McFadden, Daniel. 1973. "Conditional Logit Analysis of Qualitative Choice Behavior." In *Frontiers in Econometrics,* edited by Paul Zarembka. New York: Wiley.

McLean, Rachael M., Victoria L. Farmer, Alice Nettleton, Claire M. Cameron, Nancy R. Cook, Mark Woodward, Norman R. C. Campbell, and TRUE Consortium (in Ternational Consortium for Quality Research on Dietary Sodium/Salt). 2018. "Twenty-Four–Hour Diet Recall and Diet Records Compared with 24-hour Urinary Excretion to Predict an Individual's Sodium Consumption: A Systematic Review." *The Journal of Clinical Hypertension* 20, no. 10 (October): 1360–1376. ISSN: 1524-6175, 1751-7176. https://doi.org/10.1111/jch.13391.

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. 1974. *Schooling, Experience, and Earnings.* National Bureau of Economic Research.

Mogstad, Magne, and Alexander Torgovitsky. 2018. "Identification and extrapolation of causal effects with instrumental variables." *Annual Review of Economics* 10 (1): 577–613.

Mukerjee, Rahul, and CF Jeff Wu. 2007. *A modern theory of factorial design.* Springer Science & Business Media.

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.

Neumark, David, Ian Burn, and Patrick Button. 2019. "Is it harder for older workers to find jobs? New and improved evidence from a field experiment." *Journal of Political Economy* 127 (2): 922–970.

Neyman, Jerzy. 1923. "On the application of probability theory to agricultural experiments. Essay on principles. Section 9." Translated and edited by D. M. Dabrowska and T. P. Speed, 1990, *Statistical Science* 5 (4): 463–480.

———. 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.

Oaxaca, Ronald. 1973. "Male-female wage differentials in urban labor markets." *International Economic Review,* 693–709.

Olkin, Ingram, and John W Pratt. 1958. "Unbiased estimation of certain correlation coefficients." *The annals of mathematical statistics,* 201–211.

Oster, Emily. 2019. "Unobservable selection and coefficient stability: Theory and evidence." *Journal of Business & Economic Statistics* 37 (2): 187–204.

Pearl, Judea. 1995. "Causal Diagrams for Empirical Research." *Biometrika* 82 (4): 669–688. ISSN: 00063444, 14643510.

———. 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,* 451–482.

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.

Rényi, Alfréd. 1959. "On measures of dependence." *Acta Mathematica Hungarica* 10 (3-4): 441–451.

Ridder, Geert, and Robert Moffitt. 2007. "The econometrics of data combination." *Handbook of Econometrics* 6:5469–5547.

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, Paul R. Conlin, Edgar R. Miller, Denise G. Simons-Morton, Njeri Karanja, Pao-Hwa Lin, Mikel Aickin, Marlene M. Most-Windhauser, Thomas J. Moore, Michael A. Proschan, and Jeffrey A. Cutler. 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.

Scott, Alastair, and Chris Wild. 1991. "Transformations and R 2." *The American Statistician* 45 (2): 127–129.

Spearman, C. 1904. "The Proof and Measurement of Association between Two Things." *The American Journal of Psychology* 15 (1): 72–101. ISSN: 00029556.

University Of Cambridge, MRC Epidemiology Unit and NatCen Social Research. 2021. *National Diet and Nutrition Survey Years 1-11, 2008-2019.* 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.* Technical report. working paper.

World Health Organization. 2021. *How to Obtain Measures of Population-Level Sodium Intake in 24-Hour Urine Samples: Protocol.* Protocol WHO/EURO:2021-2333-42088-57949. Licence: CC BY-NC-SA 3.0 IGO. Copenhagen: WHO Regional Office for Europe.

Wynand, PMM, Van De Ven, and Randall P Ellis. 2000. "Risk adjustment in competitive health plan markets." In *Handbook of health economics,* 1:755–845. Elsevier.

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 additional results and examples.

Notation. We use the operators $\mathbb{E}$ and Var to refer to the expectation and variance in the true population, rather than in the hypothetical population from which the experiment is drawn.

A. Proofs for Section 4

Proof of Proposition 1. For (i), since the observable features have no effect on the outcome, the interventional expectation $\mathbb{E}[Y \mid\mid X^O = x^O] = \mathbb{E}[Y]$, and hence $\text{CR}^2(X^O) = 0$. Since $\mathcal{F}$ includes all constant models, $\mathcal{F}$ is well-specified, so $\text{CR}^2_{\mathcal{F}}(X^O) = \text{CR}^2(X^O) = 0$. For (ii), since X^O fully determines Y, $\mathbb{E}[Y \mid\mid X^O = x^O] = Y(x^O, X^U)$, for all X^U, which implies $\text{CR}^2(X^O) = 1$. Since $\mathcal{F}$ is well-specified, $\text{CR}^2_{\mathcal{F}}(X^O) = \text{CR}^2(X^O) = 1$.

For (iii), $\text{CR}^2_{\mathcal{F}}(X^O) \leq 1$, with equality only if $\mathbb{E}\left[(Y - Y^C_{\mathcal{F},X^O}(x^O))^2\right] = 0$, which requires $Y(x^O, x^U) = Y^C_{\mathcal{F},X^O}(x^O)$ almost surely. This is immediately contradicted by $\mathbb{E}[\text{Var}[Y] \mid\mid X^O = x^O] > 0$. For (iv), note that when there is no variance in observables, then X^O is constant in the population. Denote the constant value of X^O by a. $Y^C_{\mathcal{F},X^O}(a) = \mathbb{E}[Y]$ since conditioning any realization by intervention to $X^O = a$ does not change its outcome Y as it was already assigned $X^O = a$. As a result, the expectation will also equal $\mathbb{E}[Y]$. For (v), note that if there is no variance in unobservable features $Y = \mathbb{E}\left[Y_{X^O}(x^O)\right] = Y^C_{X^O}(x^O) = Y^C_{\mathcal{F},X^O}(x^O)$. The first equality says that when unobservables do not vary, then the outcome Y is equal to the expected potential outcome function that only depends on observables. The second equality is the definition of the best causal model. The third equality follows because $\mathcal{F}$ is well-specified. As a result, the risk of the best causal model under $\mathcal{F}$ is zero and the causal R^2 equals 1.

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

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

For (vii), an example suffices: the true data-generating process for (Y_i, X_i) is $Y_i(X_i) = X_i, X_i \sim \mathcal{N}(0,1)$. Consider an analyst studying the determinants of Y_i, who observes X_i. The best causal model is $\mathbb{E}[Y_i \mid\mid X_i = x_i] = x_i$, with $\text{CR}^2(Y \to X) = 1$. For an analyst studying the determinants of X_i, who observes Y_i, the best causal model is $\mathbb{E}[X_i \mid\mid Y_i = y_i] = 0$, so $\text{CR}^2(X \to Y) = 0$. □

Proof of Proposition 2. For (i), since g_Y is affine and strictly monotone, write $g_Y(Y) = \alpha + \beta Y$, for $\beta > 0$. The interventional expectation for Y' is

$$\begin{aligned}\mathbb{E}[Y' \mid\mid X^{O'} = x^{O'}] &= \alpha + \beta \mathbb{E}[Y \mid\mid X^O = g^{-1}(x^{O'})] \\ &= \alpha + \beta \mathbb{E}[Y \mid\mid X^O = x^O]\end{aligned}$$

where the first line defines g^{-1} as the element-wise inverse of $(g_k)_{k=1}^O$, which is well-defined by virtue of each (g_k) being strictly monotone. Then

$$\begin{aligned}\mathrm{CR}^2(Y' \to X^{O'}) &= 1 - \frac{\mathbb{E}\big[(Y' - \alpha - \beta \mathbb{E}[Y \mid\mid X^O = x^O])^2\big]}{\beta^2 \operatorname{Var}[Y]} \\ &= 1 - \frac{\mathbb{E}\big[(\alpha + \beta Y - \alpha - \beta \mathbb{E}[Y \mid\mid X^O = x^O])^2\big]}{\beta^2 \operatorname{Var}[Y]} \\ &= 1 - \frac{\beta^2\, \mathbb{E}\big[(Y - \mathbb{E}[Y \mid\mid X^O = x^O])^2\big]}{\beta^2 \operatorname{Var}[Y]} = \mathrm{CR}^2(Y \to X^O).\end{aligned}$$

Part (ii) follows similar steps, additionally making use of the fact g^{-1} is affine. □

Proof of Proposition 3. For (i), fixing the function class $\mathcal{F}$, distribution of features P_X, and potential outcomes function $Y(\cdot)$, the best predictive model $Y^P_{\mathcal{F},X^O}$ minimizes $\mathcal{R}(\cdot)$ among all models in $\mathcal{F}$, and hence maximizes $G(\cdot)$ among all models in $\mathcal{F}$, so $\mathrm{R}^2(X^O) = G(Y^P_{\mathcal{F},X^O}) \geq \mathrm{CR}^2(X^O) = G(Y^C_{\mathcal{F},X^O})$. For (ii):

$$\begin{aligned}Y^C_{\mathcal{F},X^O} &= \arg\min_{M \in \mathcal{F}} \mathbb{E}\big[\mathbb{E}\big[(M(X^O) - Y)^2 \mid\mid X^O = x^O\big]\big] \\ &= \arg\min_{M \in \mathcal{F}} \mathbb{E}\big[\mathbb{E}\big[(M(X^O) - Y(x^O, X^U))^2\big]\big] \\ &= \arg\min_{M \in \mathcal{F}} \mathbb{E}\big[\mathbb{E}\big[(M(X^O) - Y(x^O, X^U))^2 \mid X^O = x^O\big]\big] \\ &= \arg\min_{M \in \mathcal{F}} \mathbb{E}\big[\mathbb{E}\big[(M(X^O) - Y)^2 \mid X^O = x^O\big]\big] = Y^P_{\mathcal{F},X^O},\end{aligned}$$

where the third line follows from the independence of X^O and X^U. Hence, the best predictive and causal models coincide for each x^O. This implies they have the same risk, and hence goodness-of-fit. □

Proof of Proposition 4. For (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 the statement about causal R^2, consider the following example:

(A1) $$Y_i(x_{1,i}, x_{2,i}) = \beta_1 x_{1,i} + \beta_2 x_{2,i},$$

(A2) $$\begin{pmatrix} X_{1,i} \\ X_{2,i} \end{pmatrix} \sim \mathcal{N}(\mathbf{0}, \Sigma), \quad \Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}.$$

Suppose $X_{1,i}$ is observed, but $X_{2,i}$ is not. For $\mathcal{F}$ unrestricted or linear, the best causal model is ${Y^*_{X^O}}^{\mathcal{F}}(x_{1,i}) = \beta_1 x_{1,i}$. Then $\mathcal{R}({Y^*_{X^O}}^{\mathcal{F}}) = \beta_2^2$, and hence the goodness-of-fit of the causal model is

$$G({Y^*_{X^O}}^{\mathcal{F}}) = \frac{\beta_1^2 + 2\beta_1\beta_2\rho}{\beta_1^2 + \beta_2^2 + 2\beta_1\beta_2\rho}.$$

To see that this expression can be arbitrarily negative, note that, choosing $\rho = -1$, the expression can be rewritten

$$G({Y^*_{X^O}}^{\mathcal{F}}) = \frac{\beta_1(\beta_1 - 2\beta_2)}{(\beta_1 - \beta_2)^2},$$

and that the term inside the brackets can be made arbitrarily close to zero (from below) by taking $\beta_1 = \beta_2 + \epsilon$ for ϵ sufficiently small.

For (ii), that $\mathrm{R}^2_{\mathcal{F}}$ is monotonically increasing is well-known. Example A6 shows that $\mathrm{CR}^2_{\mathcal{F}}$ may not be monotonically increasing. □

Before proving Proposition 5, we establish an intermediate step.

Lemma 1. *Fix an outcome Y and a vector of observed features X^O. Suppose these variables have 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 matrix of correlations between the observed features. Then:*

(i) The $\mathrm{R}^2_{\mathrm{lin}}$ can be written: $\mathrm{R}^2_{\mathrm{lin}}(X^O) = \tilde{\beta}^{P\prime}\rho_{X^O}\tilde{\beta}^P$, where $\tilde{\beta}^P$ is the vector of OLS coefficients from a regression of $\tilde{Y}$ on $\tilde{X}^O$ in the population.

(ii) The $\mathrm{CR}^2_{\mathrm{lin}}$ can be written $\mathrm{CR}^2_{\mathrm{lin}}(X^O) = 2\tilde{\beta}^{\mathrm{P}}_{X^O}{}'\rho_{X^O}\tilde{\beta}^{\mathrm{C}}_{X^O} - \tilde{\beta}^{\mathrm{C}}_{X^O}{}'\rho_{X^O}\tilde{\beta}^{\mathrm{C}}_{X^O}$, where $\tilde{\beta}^C$ is the vector of OLS coefficients from a regression of $\tilde{Y}$ on $\tilde{X}^O$ in a hypothetical experiment in which X^U is distributed according to the population, and X^O is randomly assigned according to the population distribution.

Part (i) is well-known; part (ii) less so.

Proof of Lemma 1. It is well-known that the $\mathrm{R}^2_{\mathrm{lin}}$ is invariant to affine transformations, and Proposition 2 shows that this is also true of $\mathrm{CR}^2_{\mathrm{lin}}$. Then we can demonstrate the lemma by considering the standardized versions of all variables. For any linear model $M = \tilde{\alpha} + \tilde{\beta}X^O$:

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

where $\mathrm{Cov}[\tilde{Y}, \tilde{X}^O] = \mathbb{E}[(\tilde{X}^O - \mathbb{E}[\tilde{X}])(\tilde{Y} - \mathbb{E}[\tilde{Y}])] = \mathbb{E}[\tilde{X}^O\tilde{Y}] \in \mathbb{R}^K$.

For part (i), take the best predictive model (within the linear function class), $Y^P(X^O) = \tilde{\alpha}^P + \tilde{\beta}^P X^O$. Then $\text{Cov}[\tilde{Y}, \tilde{\beta}^P \tilde{X}^O] = \tilde{\beta}^{P\prime}\text{Cov}[\tilde{X}^O]\tilde{\beta}^P = \tilde{\beta}^{P\prime}\rho_{XO}\tilde{\beta}^{P\prime}$, and hence we are left with $\text{R}^2_{\text{lin}}(X^O) = \tilde{\beta}^{P\prime}\rho_{XO}\tilde{\beta}^P$.

For part (ii), take the best causal model (within the linear function class), $Y^C(X^O) = \tilde{\alpha}^C + \tilde{\beta}^C X^O$. Then we have $\text{CR}^2_{\text{lin}}(X^O) = 2\text{Cov}[\tilde{Y}, \tilde{\beta}^{C\prime}\tilde{X}^O] - \tilde{\beta}^{C\prime}\rho_{XO}\tilde{\beta}^C$. □

Now we move to the main statement.

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

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

where the first line follows from the fact ρ_{XO} is a correlation matrix (and so symmetric), and the second line applies both parts of Lemma 1. Rearranging gives the statement. □

B. Proofs for Section 5

Proof of Proposition 6. For part (i), $Y^C_{\mathcal{F},XO}$ cannot be identified from the observational sample. Conversely, for any model M, $\mathcal{R}(M)$ cannot be identified from the experimental sample. Since $\mathcal{CR}^2_{\mathcal{F}}(X^O)$ depends on both

For part (ii), since we have access to the observational sample, for any model M, $\mathcal{R}(M)$ is identified. The question is then whether we can identify Y^C_{XO}. If the experiment is full-support, for each $x^O \in \mathcal{X}^O$, we can identify $Y^C_{XO}(x^O)$ pointwise, and hence identify the function Y^C_{XO}. Otherwise, we cannot identify $Y^C_{XO}(x^O)$ for at least some values of x^O.

For part (iii), as before, $\mathcal{R}(M)$ is identified by the observational data, for any model M. Under our assumption that the linear model is well-specified, the interventional expectation is $Y^C_{XO}(x^O) = \alpha + \sum_{k=1}^O x^O\beta^O$. Regressing Y on X^O in the experimental data recovers (α, β), and hence the interventional expectation Y^C_{XO}. □

Proof of Proposition 7. Follows immediately from noting that (i) $\hat{Y}^C_{\mathcal{F},XO}$ converges to $Y^C_{\mathcal{F},XO}$, (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. □

Proof of Proposition 8. For simplicity, denote $X \coloneqq X^O$.

For part (i), expand the definition of the CR^2 with classical measurement error:

$$\begin{aligned} CR^2_{CME} &= 1 - \frac{\mathbb{E}\left[\beta^2 X^2 - 2\beta XY - 2\beta X\varepsilon + Y^2 + 2Y\varepsilon + \varepsilon^2\right]}{\text{Var}[Y] + \text{Var}[\varepsilon]} \\ &= 1 - \frac{\text{MSE} + \text{Var}[\varepsilon]}{\text{Var}[Y] + \text{Var}[\varepsilon]} \end{aligned}$$

Now, separately multiply the CR^2 without measurement error by $\frac{\text{Var}[Y]}{\text{Var}[Y]+\text{Var}[\varepsilon]}$ and rearrange:

$$\begin{aligned} CR^2 \frac{\text{Var}[Y]}{\text{Var}[Y] + \text{Var}[\varepsilon]} &= \frac{\text{Var}[Y]}{\text{Var}[Y] + \text{Var}[\varepsilon]} - \frac{\text{Var}[Y]\,\text{MSE}}{\text{Var}[Y]\,(\text{Var}[Y] + \text{Var}[\varepsilon])} \\ &= \frac{\text{Var}[Y] + \text{Var}[\varepsilon]}{\text{Var}[Y] + \text{Var}[\varepsilon]} - \frac{\text{MSE} + \text{Var}[\varepsilon]}{\text{Var}[Y] + \text{Var}[\varepsilon]} \\ &= 1 - \frac{\text{MSE} + \text{Var}[\varepsilon]}{\text{Var}[Y] + \text{Var}[\varepsilon]} \\ &= CR^2_{CME} \end{aligned}$$

For part (ii), note that the only difference between the CR^2 with and without classical measurement error is that $\text{MSE}_{CME} \neq \text{MSE}$.

$$\begin{aligned} \text{MSE}_{CME} &= \mathbb{E}\left[(\beta(X + \varepsilon) - Y)^2\right] \\ &= \mathbb{E}\left[\beta^2 X^2 + 2\beta^2 X\varepsilon + \beta^2\varepsilon^2 - 2\beta XY - 2\beta X\varepsilon + Y^2\right] \\ &= \text{MSE} + \mathbb{E}\left[2\beta^2 X\varepsilon + \beta^2\varepsilon^2 - 2\beta X\varepsilon\right] \\ &= \text{MSE} + \beta^2\,\text{Var}[\varepsilon] \end{aligned}$$

It follows that $CR^2_{CME} = CR^2 - \beta^2 \frac{\text{Var}[\varepsilon]}{\text{Var}[Y]}$.

Part (iii) follows from the well-known result that classical measurement error in the dependent variable does not bias OLS coefficients. Since the causal coefficient β is unchanged, the CR^2_{CME} is unchanged.

For part (iv), $\text{Var}_{EXP}[\cdot]$ refers to the variance in the experiment. It is a well-known result that under classical measurement error in the dependent variable, the OLS coefficient will be downward-biased as follows: $\beta_{CME} = \beta \frac{\text{Var}_{EXP}[X]}{\text{Var}_{EXP}[X]+\text{Var}_{EXP}[\varepsilon]}$. The result follows from subtracting CR^2 from CR^2_{CME} and rearranging. □