An Anti-IV Approach for Pricing Residential Amenities: Applications to Flood Risk*

Alex Bell † Stephen Billings‡ Sophie Calder-Wang§ Shusheng Zhong¶

March 19, 2025

Abstract

Understanding how markets price housing amenities is essential to addressing residential socioeconomic disparities. Traditional methods have struggled with unobserved quality, often producing wrong-signed estimates. This paper introduces a novel "anti-instrument" approach to estimating amenity prices amid unobserved quality. Our identification strategy relies on an "anti-instrumental" variable that is relevant to the unobserved quality, but conditionally independent from the amenity of interest, which are properties opposite of typical instruments for amenities. We apply this method to estimate the implicit price of flood risk using detailed housing transaction records and house-level flood risk measures. Household income, used as an anti-instrument, successfully recovers the negative price of flood risk. We find that being located in a FEMA-designated floodplain reduces home prices by 2.8% on average. Drawing on more granular measures of flood risk, we find increasing price discounts for homes with higher flood probabilities, with the most risky locations suffering a discount of nearly 5%. These price estimates are crucial for evaluating returns on investment in local public goods and environmental policies.

Keywords: Hedonic Pricing, Unobserved Quality, Amenities, Flood Risk
JEL Code: R32, C39, Q51, Q54, R21

*This paper subsumes a previous working paper titled "Pricing Neighborhood Amenities: A Proxy-Based Approach." Calder-Wang thanks the Zell/Lurie Real Estate Center at Wharton for financial support. We thank Joshua Boyd for excellent research assistance. Sample data and code to implement the methods proposed in the paper is available as a Stata package under the command *aivreg* on GitHub.

†Department of Economics, Georgia State University *ambell@gsu.edu*
‡Leeds School of Business, University of Colorado Boulder *stephen.billings@colorado.edu*
§The Wharton School, University of Pennsylvania *sophiecw@wharton.upenn.edu*
¶Kellogg School of Management, Northwestern University *shusheng.zhong@kellogg.northwestern.edu*

1 Introduction

Understanding the prices of housing and neighborhood amenities is key to answering several important economic questions in local public finance, housing inequality, and geographic mobility. Housing can be thought of as a bundle of amenities: besides the usual house-level attributes such as the number of bedrooms, number of bathrooms, and lot size, we are also concerned about the valuation of local amenities such as crime, congestion, school quality, and environmental (dis)amenities such as air pollution, fire, and flood risk. The prices of these neighborhood amenities provide us with guidance on how private and public investments should be allocated and help policymakers evaluate the returns to mitigating environmental externalities. Furthermore, given that households have heterogeneous preferences for these amenities, credible estimates of the trade-offs imposed by the housing market would help researchers unpack how much observed housing inequality is due to different preferences as opposed to other factors such as discrimination or income.

Despite the broad economic relevance and significance of housing consumption, traditional methods based on observational studies are often unsatisfactory in providing credible estimates of the prices of amenities. The challenges are so extreme that the price of neighborhood amenities based on regression methods is prone to the "perverse-signed" problem (Chay and Greenstone 2005). In Figure 2, we show the average home price by flood risk in the U.S., where we find a mostly upward-sloping pattern: places with greater flood risk have higher prices, which generates a wrong-signed positive price estimate of flood risk. Given that people do not desire flooding and often purchase insurance to cover flood risk, it is logical to conclude that there must be some sort of desirability that is not presently quantified but also characterizes these high-priced but flood-prone areas. To address such bias, the usual approach is to control for the omitted variables that may have contributed to or be correlated with that unobserved quality.

Indeed, with more granular geographic controls, such as county or census tract fixed effects, Figure 2 shows that the magnitude of the wrong-signed coefficients has been greatly reduced. However, even with the most granular census block fixed effect, combined with an extensive vector of house-level characteristics, the hedonic regression still finds a positive coefficient for being in higher flood risk categories. Thus, the inherent challenge with the hedonic regression framework resides in the presence of unobserved and highly localized characteristics being correlated with flood risk. The problem is particularly pernicious in the housing context, likely because "all houses are, by nature, a unique combination of aspects of location and structure" (Giglio, Kelly and Stroebel, 2021). The findings here also reflect the broader challenge that the empirical literature has grappled with, where a negative

price for flood risk is not easily found using standard hedonic regressions and research finding wrong-signed coefficients is counter-intuitive and thus unpublishable.[1]

The key contribution of this paper is to develop a novel identification strategy, as well as the corresponding estimation procedure, that allows us to obtain implicit amenity prices in the hedonic framework. Our method yields estimates of how home prices vary by amenity levels while holding all else equal using observational, non-experimental data. Rather than needing to fully observe all aspects of housing characteristics, our approach simply requires us to find at least one relevant "anti-instrument" for the unobserved housing quality that satisfies an appropriate conditional independence assumption with respect to the focal amenity. We termed the variable "anti" instrument in the sense that it has the exact opposite property compared with the traditional instrumental variable strategy: We seek a variable that is *correlated* with the unobserved error, but *conditionally uncorrelated* with the endogenous variable of interest.

We start with the classic sorting models of Rosen (1974) where the "Rosen-frontier" captures the trade-off between price and observed amenity. Importantly, motivated by the empirical concern that unobserved error is likely endogenous in the hedonic literature (Kuminoff, Parmeter and Pope, 2010), we focus on an equilibrium hedonic pricing function that allows for unobserved vertical quality, whose existence is guaranteed by Bajari and Benkard (2005).[2] In other words, we are interested in identifying the price-amenity trade-offs on a surface of such Rosen-frontiers by allowing for unobserved quality. By applying our approach to the trade-off homebuyers face for flood risk, we illustrate how—contrary to popular wisdom—observational data on the housing market *can* yield sensible estimates of such environmental amenities. We use property-level household income as the candidate anti-instrument for unobserved housing quality, because we believe it is likely conditionally independent from flood risk preferences, especially after controlling for granular geographic fixed effects and a wide vector of observed housing attributes. Our approach in this paper builds on one of the authors' work in a companion paper on labor market amenities (Bell, 2022), which has already proven its usefulness for

[1] For example, a number of recent papers (Bernstein, Gustafson and Lewis, 2019; Baldauf, Garlappi and Yannelis, 2020; Murfin and Spiegel, 2020) estimating the capitalized home price of flood risk from sea level rise find a mix of results using similar datasets.

[2] Although we also share with Epple and Sieg (1999); Epple, Quintero and Sieg (2020) in terms of allowing for a vertical dimension of differentiation in the housing market, our primary goal is to control for such vertical differences in the estimation of the amenity prices, which are distinct from their emphasis on concatenating out the vertical segment out as a focal subject of inquiry.

estimating wage-amenity tradeoffs in labor markets (Folke and Rickne, 2022; Burbano et al., 2022).[3]

Concretely, here is how the anti-instrumental variable framework works: In a market where (i) prices are determined by both observed characteristics and *a single index* of unobserved quality, if there exists a variable H that is both (ii) *relevant* for the unobserved quality and satisfies (iii) *conditional independence* with respect to the focal amenity, then the price of the focal amenity is identified. Constructively, the amenity price can be identified by taking the ratio of the rate at which the anti-instrument on average varies with amenity, to the rate at which the anti-instrument varies with price.

The Single Index Assumption is an economic assumption about how prices are determined in a given market. Formally, the single index assumption states the equilibrium price function, $P = P(Z, \Phi)$, exists, and is determined by the combination of observable amenities Z plus a *single* index of unobserved quality Φ. Consequently, for any product in the market, there exists a unique quality index Φ that explains its price P and amenity Z.

The Conditional Independence Assumption, on the other hand, is an econometric assumption about the property of a particular anti-instrumental variable for unobserved quality. Formally, conditional independence requires that, conditioning on the unobserved quality Φ, the proposed anti-instrumental variable H is conditionally independent from the focal amenity, $H \perp Z \,|\, \Phi$. In other words, we need a variable that is *not* correlated with the amount of amenities chosen, conditioning on the unobserved quality and other non-focal amenities.

The Relevance Condition is a technical, econometric assumption that requires the candidate anti-instrumental variable to be at least on average related to the unobserved quality. We formalize it as local monotonicity between the predicted anti-instrument $\mathbb{E}[H|\Phi]$ and Φ. Conceptually, while there may exist deviations in how well the actual anti-instrument represents the true unobserved quality, the predicted anti-instrument $\mathbb{E}[H|P, Z]$ averages out the deviations and has to change in tandem with true underlying quality.

With these three assumptions, we construct the estimator for an amenity price based on the ratio

[3]Besides the contextual difference in the empirical application (housing market vs. labor market), the present paper has made notable conceptual improvements to Bell (2022), including an alternative definition of the Single Index assumption (Section 2) and additional constructive proofs for identification in both non-linear and linear cases (Appendix A). In addition, we also provide a different derivation of the bias formula more specific to missing amenities (Appendix C, as opposed to Appendix I of Bell (2022)) and an analysis of a simple market equilibrium from Cobb-Douglas utility to micro-found the use of household income as a candidate anti-IV (Appendix B).

of how the predicted anti-IV varies by the level of amenity, compared to how it varies by price,

$$\left.\frac{\partial P(Z,\Phi)}{\partial Z}\right|_{Z=z,\Phi=\Phi(p,z)} = -\frac{\partial \hat{H}(P,Z)}{\partial Z_1} \Big/ \left.\frac{\partial \hat{H}(P,Z)}{\partial P}\right|_{P=p,Z=z} \tag{1.1}$$

where the predicted anti-IV function is simply defined by the expectation operator $\hat{H}(p,z) \equiv \mathbb{E}[H|P = p, Z = z]$.

Intuitively, the implicit price of an amenity is the rate of change in P with respect to a change in Z controlling for the unobserved quality Φ. It turns out that when H satisfies conditional independence, the predicted anti-instrument $\hat{H}(p,z)$ generates the identical level sets as $\Phi(p,z)$.[4] As a result, when it comes to estimate the implicit price of the amenity, controlling for the predicted anti-instrument $\hat{H}$ is equivalent to controlling for the unobserved quality Φ, thus allowing us to recover the implicit price without needing to control for the underlying quality which can never be fully observed.

In a model where price is a linear function of amenity and quality, then, given an anti-IV that satisfies linear conditional independence, the price of the amenity can be represented neatly as a simple ratio or regression coefficients $\hat{\beta}_z = -\delta_z/\delta_p$ where δ_z and δ_p are the corresponding coefficients of a linear regression of the anti-instrumental variable H on Z and P.

We believe that these assumptions could be reasonably satisfied within typical empirical settings. In terms of the Single Index Assumption, while seemingly restrictive, the existence of such hedonic pricing function $P(Z,\Phi)$ and its inevitability in the unobserved component Φ is provided in Theorem 1 in Bajari and Benkard (2005), which prove that minimal restrictions on the underlying utilities are required other than that the unobserved quality has to enter utility positively. Bajari and Benkard (2005) also guarantees that no specific restrictions on the supply side are required.

In the housing context, this assumption requires that the utility-relevant characteristics of a house have to be well approximated by observable housing and neighborhood amenities, plus a single index of residual unobserved quality. Given the prevalence of hedonic regressions used in the housing context, including the valuation of environmental amenities (Bishop et al., 2020), the existence of a hedonic pricing function seems a reasonable approximation. Moreover, in practice, allowing for a single index of unobserved quality is also widely used in models of housing demand, typically labeled as ξ_j, such as those in Bayer, Ferreira and McMillan (2007); Galiani, Murphy and Pantano (2015); Calder-Wang

[4] In Section 2, we illustrate the intuition behind the estimator, which is also shown graphically in Figure 4, through the lens of the implicit function theorem. The gradient of the predicted anti-IV function $\hat{H}$ represents the direction of the steepest ascent in the predicted anti-IV, shown by the blue arrow. The amenity's price in dollar terms is the rate of change in P with respect to a change in Z within the level set of $\hat{H}$, which is represented by the slope of the tangent vector of the level set, shown by the green arrow. The tangent vector of the level set is, by definition, perpendicular to the gradient of the function.

(2021), and virtually all structural models of differentiated goods following Berry, Levinsohn and Pakes (1995, 2004).[5]

In terms of the Conditional Independence Assumption, finding a variable that satisfies "conditional independence" essentially boils down to finding a variable that has to be a "good" control for quality in the typical OLS: it is not otherwise correlated with the focal amenity once true quality is held constant.[6] While OLS requires finding all the controls, the key benefit of the anti-instrumental variable method is that identification can be obtained as long as the researcher can find just *one* good control.

Empirically, we apply the anti-instrumental variable method to estimate the price of flood risk in the housing context. Understanding the implicit price of flood risk in the housing market is an economically relevant yet econometrically challenging problem. While it is intuitive to expect a negative price on flood risk, as we will show later, even a simple measure of flood risk based on floodplains generates the wrong sign problem, which is not resolved with extensive geographic fixed effects and controls for property attributes. Such wrong-signed problems perhaps explain why much of the recent literature on the valuation of flood risk has shifted towards estimating changes in home prices after flood events, under the assumption that property level amenities are unchanged by a flood event (Ortega and Taşpınar, 2018; Bin and Landry, 2013; Gibson and Mullins, 2020; Zivin, Liao and Panassie, 2023; Beltrán, Maddison and Elliott, 2019). Moreover, such first-difference approach describes changes in the shapes of price functions, rather than movements along a price function (Kuminoff and Pope, 2014), whereas our anti-instrument approach focuses on the latter, allowing estimation using data from a single price function.

We come back to the conventional hedonic regression using cross-sectional variations in flood risk but with a new identification strategy via anti-instruments. We use household income as the anti-instrument for the unobserved housing quality, which we obtain from a fuzzy merge with the Home Mortgage Disclosure Act (HMDA) Data. While the naïve OLS and OLS with extensive controls such as census block fixed effects and a wide vector of housing characteristics still perform poorly in terms of recovering even the correct sign for the price of flood risk, we find that the anti-IV approach consistently estimates a significant and negative coefficient.[7] Moreover, the anti-IV estimates of the price of the

[5] It is worth clarifying that, as with the assumptions in Bajari and Benkard (2005), the hedonic model here typically precludes utility specifications that incorporate logit errors such that dominated products can still have positive demand. In this sense, we are closer to the pure characteristics models by Berry and Pakes (2007). That said, a model with logit error does not preclude the existence of an equilibrium hedonic pricing function, but its existence is simply not guaranteed.

[6] See Section 2.E of Bell (2022) for more on the relation of the conditional independence assumption and good controls.

[7] Since our application is based on variation in property level amenities and include controls for items like elevation, home attributes and census block fixed effects, we limit concerns about missing amenities since amenities (e.g. schools, crime, access to jobs) most valued by homebuyers are larger in spatial scale than our fixed effects.

flood are also monotonically increasing in the level of projected risk. In terms of magnitude, we find that belonging to the highest category of flood risk (Flood Factor = 10) leads to a 4.9% decline in home price. We find the estimates to be robust to further controls such as elevation and distance to the coast among properties in the coastal states. We also find the estimates using the anti-instrumental variable framework to be robust when evaluating the implicit price of being located in a FEMA Flood Zone, leading to an average 2.8% discount in prices.

We believe that accurately measured household income, such as those verified through mortgage applications in the HMDA data, provide a good candidate for the anti-instrument for unobserved housing quality. It is likely that higher-income home buyers, who likely have a more lax household budget, would purchase a home with higher unobserved quality, controlling for other observable housing characteristics, thus providing us with a relevant anti-instrument.[8] It also seems plausible that, conditioning on the true unobserved quality, the extent to which household income mismeasures unobserved quality, which may be thought of as driven by unobserved household wealth, is unlikely systematically correlated with preferences for flood risk. Importantly, we do *not* require household income to be unconditionally independent from flood risk preferences, but rather, we simply require that, conditional on the true unobserved quality (or household budget), household income becomes conditionally independent from flood risk preferences.[9]

Overall, compared to the existing literature, our anti-instrumental variable approach has several major advantages. First, a crucial advantage of our anti-IV estimation procedure is that researchers are no longer required to come up with a complete list of controls for all components of unobserved quality. Rather, the focus is shifted towards finding a single appropriate anti-IV that can act as a "shifter" for unobserved quality. Moreover, because the method is based on observational data alone, it can be widely applied to a variety of settings without being constrained by the availability of specific natural experiments. Furthermore, the anti-instrumental variable method can produce more nuanced prices that can vary by the quality segments of the market, which allows for amenity prices to differ for different household income levels. Lastly, the implementation of the anti-IV-based method is also computationally inexpensive.

While the anti-instrumental variable approach offers a clear advantage over existing observational methods for analyzing amenity prices, limitations remain. Although the hedonic pricing function required by the Single Index Assumption may approximate the housing market, it might not apply to

[8] An f-test supports this relevance assumption.

[9] In other words, if higher-income people are more (or less) sensitive to flood risk, this is not a threat to identification so long as it is also the case that people who buy higher-quality homes are also more (or less) sensitive to flood risk.

other markets where a significant portion of prices are driven by idiosyncratic variation. In addition, despite the strength of our approach in holding fixed a single index of housing quality, we cannot control for, or price, additional amenities that the researcher does not observe. The extent to which this is a limitation depends on the nature of the correlation structure between the observed amenity, missing amenity, and unobserved quality. However, as long as the candidate anti-instrument remains conditionally independent from the focal amenity with an appropriately defined unobserved quality that also includes the missing amenity, then our anti-IV framework still produces a consistent estimate.

Econometrically, our estimation framework is conceptually connected to the selection-correction approach of Altonji, Elder and Taber (2005) and Oster (2019) due to the shared focus on the unobservables, but we operate within totally different contexts: The selection-correction approach emphasizes drawing a random subset of otherwise unobserved controls to estimate treatment effects, whereas the anti-instrumental method emphasizes finding a single variable that satisfies conditional independence within an equilibrium hedonic framework. In addition, our identification result is general in that we can flexibly allow for non-linearity and non-separability in the hedonic pricing function, which are shared features emphasized in Ekeland, Heckman and Nesheim (2004); Heckman, Matzkin and Nesheim (2010). However, our core innovation is to provide a new strategy to identify the price-amenity trade-off, rather than to uncover the full underlying demand preferences. That said, with additional assumptions pertaining to distributions or functional forms, one could extend this method for estimating prices of amenities to also provide estimates of demand for amenities as well.

The rest of the paper is organized as follows. Section 2 formalizes the theory behind the anti-instrumental variable approach to estimating amenity prices. We first prove that amenity prices can be identified in the general case with a relevant anti-instrumental variable that satisfies conditional independence. We then proceed with the special linear case, where we show that the amenity price can be estimated as the ratio of two regression coefficients. We also describe a set of relevant extensions to multiple amenities and multiple plausible anti-instruments. In Section 3, we provide an empirical application of the anti-instrumental variable method in estimating the price of flood risk in the residential housing market. Section 4 concludes.

2 Model and Methodology

2.1 Motivation for the Anti-IV Method: "Alice and Bob"

In this section, we describe an illustrative example highlighting the role played by unobservables in driving the wrong-sided problems pervasive in estimating amenity prices. First, consider the naive

approach towards estimating the amenity price using observational data by simply regressing prices on amenities. Figure 3 illustrates this approach. In this example, there are two individuals: Alice and Bob. Alice lives in an expensive oceanfront single-family home, with substantial flood risk. Bob lives in a similarly sized single-family home that is just down the block but not abutting the shore. While Bob's home lacks beach access and thus is commonly regarded as less desirable — and therefore commands a lower price — it does hold the advantage of a lower flood risk. However, a naïve regression of home price on amenities would find that the higher price of Alice's house compared to Bob's house must be justified by the difference in flood risk, suggesting a positive price for flood risk, which is not the expected sign.

So, why does this naive regression of price on amenity fail? To the extent that there are other unobserved differences between Alice and Bob's houses, it would all be erroneously attributed to the price of flood risk. In other words, Alice's house is expensive, probably because it is nicer than Bob's house in many uncontrolled ways, such as better access to the beach, local neighborhood appeal, and the quality of the interior of the house. Figure 3 illustrates the stylized "quality segment" lines in Figure 3, where Alice's house sits on a higher quality segment than Bob's house.

Naturally, a sensible approach to amend this problem of missing quality is to find other homes with comparable quality to Alice's house. For example, Ashley's house on the same beach and also at the oceanfront might be a suitable candidate: it sits on the same quality segment with a similar ocean view and neighborhood appeal, but it sits on a more elevated part of the land and thus has lower flood risk than Alice's house. As a result, by comparing the price difference between Ashley's house and Alice's house, where both sit on the *same* quality segment, we would be able to infer the (negative) price of flood risk.

Conventional approaches would attempt to address this problem by including additional control variables in the OLS to address the unobserved quality in the regression, such as granular geographic fixed effects and detailed house-level characteristics. However, because it is never possible to control for all potential amenities, it is still possible that the remaining unobservable is correlated with the amenity. In this particular example, it could be that Ashley's home, while located at the same distance to the coast, faces a less appealing and more rocky part of the beach yet with lower flood risk. In this case, Ashley's home would sit at a lower quality tier level than Alice's but with lower flood risk, still contributing to the wrong-signed problem. The problem with unobservables is a particularly salient feature of the housing market. Every house is inherently unique in its location and characteristics, making it impossible to control for all *all* components that contribute to its price.

The core innovation of our paper is that we can eschew the need to include all quality-relevant controls by using an "anti-instrumental" variable. All we need is *one* relevant anti-IV that satisfies the "conditional independence" assumption, namely, the proposed anti-IV is relevant in terms of being informative of the unobserved quality, but *not correlated* with preferences for the focal amenity conditioning on the quality and other observables. Then, instead of having to condition on the unobserved quality itself, we can instead condition on the component of the anti-IV that co-varies with the unobserved quality, allowing us to estimate the amenity price consistently. The next section provides a more formal account of the theory.

2.2 Theory of the Anti-Instrumental Variable Method

In this section, we state the the assumptions underlying the anti-instrumental variable approach and the main identification result. We first develop the main theorem in a general framework where the price function $P = P(Z, \Phi)$ is a non-parametric, non-separable function. Then, we develop the main theorem when the price function is linear in both Z and Φ as a special case. We provide the corresponding estimation procedures in both cases and illustrate their uses in our empirical application in Section 3.

2.2.1 Theory of the Anti-Instrumental Variable Method (General Case)

Assumption 2.1. ***(Single Index Assumption)*** *The price of a home P is determined by its amenity Z and a* ***single*** *index of housing quality Φ, where the pricing function $P = P(Z, \Phi)$ exists and is invertible in Φ*

$$\Phi = \Phi(P, Z) = P^{-1}(P, Z). \tag{2.1}$$

Conceptually, the *Single Index Assumption* first rests on the existence of a pricing function $P(Z, \Phi)$ that maps observed housing amenity Z and unobserved quality Φ into price $\Big($i.e, $(Z_j, \Phi_j) = (Z_{j'}, \Phi_{j'}) \Rightarrow P_j = P_{j'}\Big)$. More substantively, the *Single Index Assumption* ensures the invertibility in the unobserved Φ, such that one can always solve for the unobserved quality for a given price and observed amenity.

Note that *Single Index Assumption* is an economic assumption about the underlying structure of how prices are determined. Next, we outline the two econometric assumptions about a candidate anti-instrumental variable H that provide us with identification.

Assumption 2.2. ***(Conditional Independence)*** A candidate anti-instrumental variable H is con-

ditionally independent of Z if

$$H \perp Z \,|\, \Phi. \tag{2.2}$$

In other words, a candidate variable H satisfies conditional independence if it is *not* informative about the level of amenity Z once conditioned on a particular level of unobserved quality Φ.

Assumption 2.3. ***(Monotonicity)***

$$\mathbb{E}[H|\Phi] \text{ is strictly monotone in } \Phi. \tag{2.3}$$

The monotonicity assumption stipulates that the anti-instrumental variable under consideration H is relevant: it is, on average, informative about the quality index Φ.

Before proceeding, the following lemma forms the basis of our main theorem.

Lemma 2.4. *In a market that satisfies the Single Index Assumption, if a candidate anti-instrumental variable H satisfies Conditional Independence, then*

$$\mathbb{E}[H \,|\, P = p, Z = z] = \mathbb{E}[H \,|\, \Phi = \phi]. \tag{2.4}$$

where $\phi = \Phi(p, z)$.

Proof.

$$\begin{aligned}
&\mathbb{E}[H|P = p, Z = z] && && (2.5)\\
&= \mathbb{E}[H|P = p, Z = z, \Phi = \phi] && \text{by Single Index Assumption (invertibility in } \phi = \Phi(p, z)) && (2.6)\\
&= \mathbb{E}[H|Z = z, \Phi = \phi] && \text{by Single Index Assumption (existence of } P(z, \phi)) && (2.7)\\
&= \mathbb{E}[H|\Phi = \phi]. && \text{by Conditional Independence Assumption} && (2.8)
\end{aligned}$$

□

We define the predicted anti-IV function as $\hat{H}(P = p, Z = z) \equiv \mathbb{E}[H \,|\, P = p, Z = z]$ and $\hat{H}(\phi) \equiv \mathbb{E}[H \,|\, \Phi = \phi]$, then, Lemma 2.4 can be written as $\hat{H}(P = p, Z = z) = \hat{H}(\phi)$, where $\phi = \Phi(p, z)$.

Intuitively, the lemma means that the predicted anti-IV function $\hat{H}(P = p, Z = z)$ captures the component of H that covaries with the unobserved quality Φ.

Theorem 2.5. ***(The price of the amenity is identified using an anti-instrumental variable.)*** *Consider a market that satisfies the single index assumption as described in Assumption 2.1. If there*

exists an anti-IV H that satisfies conditional independence (Assumption 2.2) and strict monotonicity locally at (p, z) (Assumption 2.3), then the price of the amenity is identified and can be consistently estimated as

$$\left.\frac{\partial P(Z,\Phi)}{\partial Z}\right|_{(z,\phi)} = -\frac{\partial \hat{H}(P,Z)}{\partial Z} \Big/ \left.\frac{\partial \hat{H}(P,Z)}{\partial P}\right|_{(p,z)} \tag{2.9}$$

where $\phi = \Phi(p, z)$, and $\hat{H}(\cdot, \cdot)$ refers to the predicted anti-IV

$$\hat{H}(p, z) \equiv \mathbb{E}[H \mid P = p, Z = z]. \tag{2.10}$$

Proof. See Appendix A.1. □

The intuition behind our main theorem is as follows. By definition, the price of the amenity at any given observed (p_0, z_0) is determined by how prices change in response to a change in amenity, holding its unobservable quality $\phi_0 = \Phi(p_0, z_0)$ fixed. Therefore, as illustrated in Figure 4, the amenity price can be represented by the tangent direction of the level set of $\Phi(p, z)$ valued at ϕ_0. When there exists an anti-instrumental variable H that satisfies Conditional Independence, Lemma 2.4 asserts that the predicted anti-IV function satisfies $\hat{H}(p, z) = \hat{H}(\phi)$. This means that the level set of the function $\hat{H}(\cdot, \cdot)$ valued at $\hat{H}(\phi_0)$ is *identical* to the level set of $\Phi(\cdot, \cdot)$ when valued at ϕ_0:

$$\Big\{(p, z) : \Phi(p, z) = \phi_0\Big\} = \Big\{(p, z) : \hat{H}(p, z) = \hat{H}(\phi_0)\Big\}, \tag{2.11}$$

thus producing the same tangent vector. Therefore, the amenity price can be estimated as the tangent vector of the level set of the predicted anti-IV function, namely,

$$\left.\frac{\partial P}{\partial Z}\right|_{z_0,\phi_0} = -\frac{\partial \Phi}{\partial Z} \Big/ \left.\frac{\partial \Phi}{\partial P}\right|_{p_0,z_0} = -\frac{\partial \hat{H}}{\partial z} \Big/ \left.\frac{\partial \hat{H}}{\partial p}\right|_{p_0,z_0}. \tag{2.12}$$

Moreover, Figure 4 also shows that the tangent direction is by definition perpendicular to the gradient of the predicted anti-IV function $\nabla \hat{H} = [\partial \hat{H}/\partial Z, \partial \hat{H}/\partial P]$, where the gradient represents the direction of increasing unobserved quality.

Therefore, even though we cannot fully observe the true quality ϕ, the fact that $\hat{H}(P, Z)$ and $\Phi(P, Z)$ share identical level sets allows us to identify the correct amenity price as long as we estimate $\hat{H}(P, Z)$. In other words, instead of directly controlling for the unobserved quality Φ, it suffices to control for the predicted anti-IV $\hat{H}$ when computing the trade-off between price and amenity. This is the key

intuition for why the anti-IV method can be effective in estimating implicit prices despite the presence of unobserved quality.

2.3 A Discussion of the Assumptions

Given that the applicability of the anti-instrumental variable method depends on the plausibility of various assumptions, in this section, we discuss some of the intuitions behind the key assumptions.

2.3.1 The Single Index Assumption

Underlying the anti-IV approach is a structural assumption that there is a *single* index of unobserved quality for which the researcher wants to control. Is that a reasonable assumption? What kind of restrictions does the *Single Index Assumption* place on the underlying household demand and the supply?

A direct application of Theorem 1 of Bajari and Benkard (2005) produces both the existence of the equilibrium price function $P(Z, \Phi)$,[10] and the invertability in $\Phi = \Phi(P, Z)$. Bajari and Benkard (2005) provides a set of sufficient conditions that guarantee them, which are minimal in both utilities, and none in terms of the supply-side assumptions.[11] The only substantive restriction is that the single, unobserved component Φ enters positively into all household utilities. It is in this sense that the unobserved error is a *vertical* component, where all household agrees on the sign in which it enters their utility. Intuitively, if the unobserved error were horizontal, then invertability may not hold, because there may be multiple values of the observed error that could justify the observed price and amenity.

While it may seem restrictive to concatenate all unobservables into a single quality index, we believe that this may be a practical empirical approximation: It represents a combination of all hard-to-observe quality of a house, representing the quality of its built and its maintenance (e.g., whether plumbing was properly done, whether the basement was properly sealed, whether the fence was aesthetically erected, whether pests was effectively controlled etc.) Moreover, it is not more onerous than the typical regression setting, where one automatically assumes a single dimension of unobserved error. To the

[10] Strictly speaking, Theorem 1 of Bajari and Benkard (2005) applies to a hedonic model of individual utility maximization whereby individuals maximizes their utility $u_i(z_j, \phi_j, c)$ subject to some budget constraint y_i. The utility, however, does not have an error term that is specific to both the choice and the individual, namely, $\epsilon_{i,j}$. However, the mere presence of $\epsilon_{i,j}$ also does not necessarily preclude the existence of an equilibrium pricing function, but, in that case, it would typically require more restrictions on the supply side. However, in the housing context, given the prevalence of hedonic models in both research and practice, we believe that the assumption that an equilibrium hedonic pricing function exists is also not restrictive.

[11] The assumptions stated by Bajari and Benkard (2005), include (i) individual utility $u_i(z_j, \phi_j, c)$ is continuous differentiable in the composite good c, and strictly increasing in c with $\partial u_i / \partial c > \epsilon$ for some $\epsilon > 0$ and $c \in (0, y_i]$, (ii) u_i is Lipschitz continuous in z_j and ϕ_j, (iii) u_i is strictly increasing in ϕ_j.

extent that we are mostly concerned about the unobserved quality being correlated with the focal amenity, leading to biased estimation, it is a single-dimensional, vertically differentiated error is what we are most interested in capturing.

2.3.2 The Conditional Independence Assumption

Compared with traditional IV methods, which attempt to find a variable that shifts with the endogenous treatment but otherwise uncorrelated with the unobservable error, the anti-IV method attempts to find a variable that shifts with the unobserved error, as required by Monotonicity but is otherwise uncorrelated with the observable amenity, as required by Conditional Independence, and hence the name "anti" instrument. In other words, while the traditional IV strategy focuses on instrumenting for the endogenous *observable* variable, the anti-IV method focuses on instrumenting for an endogenous *unobservable* variable.

While conceptually distinct from traditional instrumental variable methods, econometrically, one may interpret the Conditional Independence assumption as an alternative form of exclusion restriction that provides identification. In this sense, the anti-IV method shifts the burden of identification from finding the most comprehensive set of controls to finding an appropriate anti-instrument, much the same way that traditional IV methods shift the burden of identification from finding such controls towards finding an appropriate instrument that satisfies the canonical IV exclusion restriction. Just as there is no direct way to "prove" IV validity, there is also no direct way to "prove" anti-IV validity. Instead, establishing the plausibility of validity would require contextual knowledge and relevant robustness and/or over-identification tests.

When conditional independence is violated, it would lead to a biased estimate of the amenity price. Figure 5 illustrates the intuition behind the direction of the bias. First, the top panel shows that when the anti-IV and the focal amenity are conditionally independent, variations in the anti-IV are not correlated with the focal amenity within each quality segment. However, in the bottom panel, when the anti-IV and the focal amenity are positively correlated conditional on unobserved quality Φ, then, within each quality segment Φ, higher H are more likely to appear at higher levels of Z. Consequently, the level set of H, namely, the set of P, Z that would generate the same H, would appear steeper than the true level set of Φ, leading to an upward biased estimate.

2.3.3 Theory of the Anti-Instrumental Variable Method (Linear Case)

In this section, we provide the formulation of the main theorems when amenity Z and the unobserved quality Φ each enter linearly into the price.

Because the unobserved quality Φ has no inherent units, without loss of generality, let the equilibrium price be determined by Z and Φ in a linear model as $P = \beta_0 + \beta_z Z + \Phi$. Notice that it implies the quality is easily inverted as $\Phi(P, Z) = P - \beta_0 - \beta_z Z$. In addition, the true price of the amenity is β_z.

Definition 2.6. Define the linear projection operator $\mathbb{E}^*[Y|X]$ as follows:

$$\mathbb{E}^*[Y|X] = \beta_0 + \beta_x X \tag{2.13}$$

where β_x denotes the OLS coefficients.

Theorem 2.7. ***(The price of the amenity is identified by the "ratio of regression coefficients" in a linear model.)*** *Consider a market that satisfies (i) the linear single index assumption as follows*

$$P = \beta_0 + \beta_z Z + \Phi. \tag{2.14}$$

If there exists an anti-instrumental variable H that satisfies

$$\textit{(ii) Linear conditional independence:} \quad cov(H, Z^{\perp\Phi}) = 0 \tag{2.15}$$

$$\textit{(iii) Relevance:} \quad cov(H, \Phi) \neq 0 \tag{2.16}$$

where $Z^{\perp\Phi}$ denotes the residual of the linear projection from Z to Φ, then the price of the amenity is identified and can be consistently estimated as the ratio of two regression coefficients

$$\beta_z = -\frac{\delta_z}{\delta_p} \tag{2.17}$$

where δ_z and δ_p are OLS regression coefficients of H on Z and P respectively:

$$\mathbb{E}^*[H|Z, P] = \delta_0 + \delta_z Z + \delta_p P. \tag{2.18}$$

Proof. See Appendix A.2 □

Intuitively, the "ratio of regression coefficients" estimator for the linear estimator is simply a special case of applying the general theorem, where the general expectation operator $\mathbb{E}[H|P, Z]$ is replaced with the linear expectation operator $\mathbb{E}^*[H|P, Z]$. Nonetheless, it is important to note that the price of the amenity in the general case is defined locally for every triplet (p, z, ϕ), whereas the price of the

amenity in the linear case is defined globally across the entire domain due to the linear single index assumption.

2.4 Estimation Procedure

Building on the results of Theorem 2.5 and Theorem 2.7, one can formalize an estimator for both the general case and for when the single latent index enters linearly.

2.4.1 Estimation Procedure for the General Case

The main Theorem 2.5 above not only provides us with the appropriate assumptions for when the amenity price is identified with the anti-IV method, but it also suggests a natural method to obtain an estimate of the price in practice. Concretely, to estimate the price of the amenity in the neighborhood, one can first divide the observations into quantiles Q_i^z of z (indexed by i) and quantiles Q_j^p of p (indexed by j). Next, we can compute the predicted anti-IV by taking the average of the anti-IV for all the observations of (p, z) in each of the cell among the Cartesian-product of quantiles: $\hat{H}_{Q_j^p, Q_i^z} = \mathbb{E}[H | p \in Q_j^p, z \in Q_i^z]$.[12] Then, one may approximate the partial derivatives by their discrete analogues:

$$\Delta\hat{H}/\Delta Z\Big|_{(p,z)} = (\hat{H}_{Q_j^p, Q_{i+1}^z} - \hat{H}_{Q_j^p, Q_i^z})/(\hat{z}_{Q_{i+1}^z} - \hat{z}_{Q_i^z}) \tag{2.19}$$

$$\Delta\hat{H}/\Delta P\Big|_{(p,z)} = (\hat{H}_{Q_{j+1}^p, Q_i^z} - \hat{H}_{Q_j^p, Q_i^z})/(\hat{p}_{Q_{j+1}^p} - \hat{p}_{Q_j^p}) \tag{2.20}$$

where $(p \in Q_j^p, z \in Q_i^z)$ and $\hat{p}$ and $\hat{z}$ denote the means in the corresponding quantile. Hence, by Theorem 2.5, the price of the amenity can be estimated as

$$-\frac{\Delta\hat{H}/\Delta Z}{\Delta\hat{H}/\Delta P}\Big|_{(p,z)}. \tag{2.21}$$

2.4.2 Estimation Procedure for the Linear Case

By Theorem 2.7, the price of the amenity in the linear model is estimated by the ratio of two regression coefficients

$$-\frac{\delta_z}{\delta_p} \tag{2.22}$$

[12] Our approach of binning by quantiles is essentially the three-dimensional analogue of the binscatter non-parametric approach made popular by work such as Chetty, Friedman and Rockoff (2014) and formalized by Stepner (2013).

where δ_z and δ_p denote the coefficients from a linear regression of H on P and Z:

$$\mathbb{E}^*[H|P,Z] = \delta_0 + \delta_p P + \delta_z Z. \tag{2.23}$$

It is also interesting to note that a numerically equivalent estimator is to view the linear regression of H on P and Z in Equation (2.23) above as a "first stage". The predicted values from the first stage, $\hat{H}$, are then included as a control in the original hedonic regression as follows:

$$\mathbb{E}^*[P|Z,\hat{H}]. \tag{2.24}$$

Intuitively, the reason that the "second-stage regression" in (2.24) is equivalent to the "ratio-of-coefficients" method in recovering the correct price could be viewed through the lens of the implicit function theorem: our main results in Theorem 2.7 shows that we can use the tangent vector of the quality level set as the price of the amenity, which motivates the ratio-of-coefficients specification. On the other hand, controlling for $\hat{H}$ is akin to estimating the tangent vector by directly controlling for being on the same quality level set through the inclusion of $\hat{H}$ as part of the control.

2.4.3 Standard Errors

In general, one can use the delta method to derive the asymptotic standard errors or bootstrap to account for finite-sample biases.[13] However, as discussed in Bell (2022), just like the conventional IV estimation could be susceptible to weak instrument problems when the first-stage coefficient is close to zero, the anti-IV estimator could also be susceptible to a problem analogous to weak inference: If the denominator in the "ratio-of-coefficient" estimator in Eq (A.15) is close to zero, namely, if the first-stage relationship between the anti-IV and home prices is weak, then the resulting distribution of the estimator is not well approximated by a normal distribution, which afflicts both the asymptotic and bootstrapped standard errors. To address this problem, we suggest using Anderson-Rubin confidence intervals (Anderson and Rubin, 1950; Andrews, Stock and Sun, 2019), which are known to be robust to such weak inference problem, when the first stage regression of the anti-IV on price is found to be weak.

[13]The delta method is implemented in Stata by the command *nlcom*.

2.5 Additional Properties of the Anti-IV Estimator

2.5.1 Extensions to Allow for Multiple Amenities

The identification result and the estimation strategy are readily extended to allow for a vector of amenities. Moreover, we allow them to be divided into a subset of focal amenities $Z_1, \ldots, Z_K$ and another subset of non-focal amenities as controls X. To examine market trade-offs between P and the focal amenities $Z_1, \ldots, Z_K$, the conditional independence assumption must hold for each focal amenity as follows

$$\forall Z_k \in \{Z_1, \ldots, Z_K\} : \; H \perp Z_k \,|\, (\Phi, X). \tag{2.25}$$

Intuitively, the anti-instrument H must contain no further information about an individual's choices over any given focal amenity condition on their choice of unobserved quality Φ and other observable controls X. In general, X could represent a wide vector of observable characteristics that may be relevant for price, but they are not the subject of interest, such as geographic fixed effects in the housing context. Including an extensive vector of X could make the single index assumption more palatable when it allows for various horizontal amenities so that the unobserved error can be better approximated by a single index of vertical quality.

Estimation of the multiple-amenity model, again, follows either the ratio-of-coefficients approach or the predicted-values approach as described in Section 2.4.2. In particular, we can estimate the amenity price for each amenity as the ratio of coefficients as follows:

$$\mathbb{E}^*[H|P, Z_1, ..., Z_K, X] = \delta_0 + \delta_P P + \delta_1 Z_1 + ... + \delta_K Z_K + \delta_X X \tag{2.26}$$

$$\forall k = 1, \ldots, K : \hat{\beta}_k = -\frac{\delta_k}{\delta_P} \tag{2.27}$$

The proof of the result follows directly from the existing proof for Theorem 2.7 by including the controls X in the linear projection of H onto P and Z. Because the conditional independence assumption in Eq (2.25) includes both X and Φ in the conditioning set, the implicit price on Z is still identified, even though the implicit price on X is not.

Equivalently, the amenity price can be estimated by performing a second-stage regression of price on amenities controlling for the predicted anti-IV

$$\mathbb{E}^*[P|Z_1, ..., Z_K, \hat{H}] = \beta_0 + \beta_1 Z_1 + ... + \beta_K Z_K + \beta_H \hat{H} \tag{2.28}$$

where the predicted anti-IV is analogously defined as $\hat{H}(P, Z_1, \ldots, Z_K, X) \equiv \mathbb{E}^*[H|P, Z_1, ..., Z_K, X]$.

2.5.2 Extensions to Multiple Anti-Instruments

When multiple valid anti-instruments are thought to exist for unobserved quality, the set of options available to the researcher is comparable to the case when multiple instruments are available for assessing the effect of an endogenous treatment on an outcome of interest.

The extent to which the estimates are similar means that either all anti-instruments satisfy the identification condition, or else all are biased in a similar way, and the likelihood of rejecting this hypothesis can be quantified with a J-statistic. A complementary approach to leveraging over-identification would be to combine the information from all anti-instruments into a single, more precise estimate with an optimal weighting given by generalized methods of moments (GMM) estimator identified by our linearized conditional independence assumptions.

2.5.3 Missing Amenities

A natural concern that arises here is how to address the issue of missing amenities. What if there are multiple utility-relevant amenities, but the researchers are only able to observe a subset of them?[14]

It is important to note that missing amenities would only bias our estimates if conditional independence requires the inclusion of such amenities. In other words, if the candidate anti-IV remains conditionally independent from the focal amenity without controlling for such utility-relevant amenity, then our anti-IV framework still produces a consistent estimate.

If the missing amenity is indeed required by conditional independence, then whether this leads to "bias" depends on the equilibrium quantity the researcher is seeking to estimate. On the one hand, our approach will not pin down a hypothetical price of an amenity holding all other amenities and quality constant.[15] On the other hand, our method does estimate the price a home-buyer would pay to obtain the additional amenity in the current market, factoring in the existing correlation between the observed amenity and unobserved amenities as well as unobserved quality in the market.

[14] In practice, the scale at which we observe our focal amenity makes the concern about missing amenities less relevant since we can measure flood risk at the level of an individual property and include quite granular fixed effects. Thus, we are only concerned with amenities that are quite spatially concentrated, which would eliminate important housing amenities such as schools, crime or job access. Applications to amenities with larger spatial scales will generate more concern about missing amenities and for amenities at similar spatial scales, one can only interpret estimates for a bundle of amenities.

[15] We derive a corresponding bias formula for the anti-IV estimate, which is distinct from the OLS bias formula, detailed in Appendix Section C.

2.6 Relationship with Existing Estimators

2.6.1 Relationship with Treatment Effect Estimators

While our anti-instrumental variable framework is primarily developed to estimate amenity prices in the presence of unobserved quality, to the extent it is a method that addresses endogeneity by leveraging the properties of the *unobservables*, it shares conceptual similarities with treatment effect estimators that rely on assumptions on the selection on unobservables, such as Altonji, Elder and Taber (2005), henceforth AET, and Oster (2019).

Our *single index assumption* (Assumption 2.1) is analogous to setting $R_{max} = 1$ in Oster's correction formula for calculating the true parameter on the treatment effect β^*, as given by

$$\beta^* = \tilde{\beta} - [\mathring{\beta} - \tilde{\beta}]\frac{R_{max} - \tilde{R}}{\tilde{R} - \mathring{R}} \tag{2.29}$$

where $\tilde{\beta}$ and $\mathring{\beta}$ are the coefficients from regressions with and without the observed control, respectively, and likewise $\tilde{R}$ and $\mathring{R}$ are the corresponding R-squared values. The only parameter not given by the data, which must be chosen by the researcher, is R_{max}, which represents the R-squared from a "hypothetical regression of the outcome on treatment and both observed and unobserved controls." In our case, R_{max} could be thought of as the R-squared from a regression of home prices on all observed amenities and the true unobserved quality.

While researchers differ in terms of whether $R_{max} = 1$ is appropriate for their respective settings, in the housing setting, we believe that the assumption that home prices are driven by a wide array of observed amenities and unobserved quality is a reasonable approximation for the market. In other settings, Altonji, Elder and Taber (2005) set $R_{max} = 1$ when estimating returns to Catholic schools. Oster (2019) argues whether this assumption holds is driven by the extent of idiosyncratic variation or measurement error in the outcome.[16] Since home prices are typically accurately measured by transaction records for tax purposes, we are less concerned about measurement error. It also seems plausible that the role of search and match frictions is secondary to observed housing amenities and housing quality in determining prices. Moreover, to the extent that Oster (2019) has suggested setting $R_{max} = 1.3\tilde{R}$ for typical empirical applications, our hedonic regressions on home prices generate $\tilde{R}$ that is over 80%, putting $1.3\tilde{R}$ well above 100%.

Our key econometric assumption of *conditional independence* (Assumption 2.2) can be viewed as

[16] For example, in the labor context, in his critique of reverse regression for salary discrimination, Goldberger (1984) points out substantial idiosyncrasies in wage determination (of the type later formalized by Mortensen), even when controlling for true productivity, necessitates setting R_{max} to be below 1.

analogous to *equal selection* in Altonji, Elder and Taber (2005). In the treatment-effect setting, Altonji et al. (2024) states this assumption as "selection on the unobservables is the same as selection on the remaining observables." In our anti-IV setting, it means that the portion of unobserved quality that is driven by the observed anti-IV H is equally correlated with the focal amenity as the portion of unobserved quality that is driven by other unobservables. Thus, once the entirety of the unobserved quality is controlled for, the anti-IV will no longer be correlated with the focal amenity, implying conditional independence.[17]

Despite the formulaic equivalence and conceptual connections, our framework remains distinct from the selection literature because it addresses problems that arise in entirely different contexts. The anti-IV method operates within a structural framework where the hedonic price function exists. It then leverages a candidate anti-IV to estimate implicit amenity prices when prices are determined simultaneously by both observed amenity and unobserved quality. On the other hand, the AET method is built on a reduced-form setting where a large set of observed and unobserved controls can be thought of as randomly drawn. It uses the inclusion of additional control to address the likely bias due to unobserved controls.

As a result, they are used for different types of empirical applications and often for distinct purposes. Many papers that cite Oster (2019) use the correction framework to assess coefficient stability in the face of additional controls, arguing the sufficiency of the existing controls. However, in our framework, we are concerned about coefficient *instability*, where we examine how a relevant anti-IV can restore the correct sign from commonly encountered wrong-signed amenity price estimates when unobserved quality is a key confounder.

2.6.2 Relationship with Control Functions Approaches

Despite the seeming similarities in making the unobserved error explicit and the ability to deal with non-linear, non-separable functions, the anti-instrumental variable framework is conceptually distinct from control function methods.

Control function methods, just like traditional instrumental variables, leverage variables that are correlated with the observed endogenous variable, whereas the anti-IV methods leverage variables that are correlated with unobserved error. Recall that, in its simplest format, control function methods provide identical point estimates as 2SLS (Wooldridge, 2007). In this sense, given our framework for instrumenting for the unobserved error is labeled as "anti-IV" in the linear case, our method can also

[17] Consider $\Phi = \gamma H + \Phi^u$ where γ represents the projection of Φ on H. Then, equal selection can be thought of as $\mathbb{E}^*[Z|\gamma H, \Phi^u] = \psi_h \gamma H + \psi_u \Phi^u$, where $\psi_h = \psi_u = \psi$. Hence, the projection $\mathbb{E}^*[Z|\gamma H, \Phi] = \psi(\gamma H + \Phi^u) = \psi\Phi$, implying a zero coefficient in front of H, which represents conditional independence.

be labeled as an "anti-control function" approach in the non-linear case.

3 Empirical Application

Our main empirical application of the anti-instrument is estimating the capitalized value of property level flood risk. Because we want to highlight the ability of our anti-instrument to address concerns about unobservables that are correlated with flood hazard risk, we first present a standard OLS hedonic model that has been the most widely implemented method in this literature (Giglio, Kelly and Stroebel, 2021). These hedonic models often include controls for property attributes as well as quite granular geographical fixed effects to address local unobserved amenities. Then, we implement the anti-instrumental variable strategy using household income as the candidate anti-IV. Additionally, we examine a number of different measures of flood risk: *Flood Factor (1-10)* which provides a measure of future flood risk and *In a Floodplain*, which is an indicator if a home is currently located in a 100 year floodplain.

3.1 Data

In order to study the capitalization of property level flood risk, it requires multiple dimensions of data disaggregated to individual parcels. Here we describe the datasets incorporated into our analysis that covers transactions from 2006-2017.

3.1.1 Home prices

We obtained national property-level data from local tax assessor records collected by a third-party company that contains detailed information on property attributes as well as transacted sales price, mortgage information and the date of the corresponding property sale and mortgage. Variables from the assessor files provide the exact geocoded location of each property, as well as information on a broad set of property information including square footage, number of stories, the number of bed/bathrooms, and build year. We limit our sample of transactions to those that involve an arm's length transaction with a transacted price of at least $20,000 and not to exceed $5M in order to address unusual properties in the tails of the distribution. Since our anti-instrument requires information on homebuyer income for a given transaction, our sample is limited to those transactions which contain a mortgage that can be matched to a transaction based on the methodology described below.

3.1.2 Flood Risk Measures

We incorporate two measures of flood risk into our application. The first measure is forward-looking in nature and incorporates multiple hydrological and climate based datasets to generate a predictive

measure of future flood risk. The second measure is a standard indicator of current flood risk, being located in Federal Emergency Management Agency's (FEMA) Special Flood Hazard Area (SFHA), or commonly know as the 100 year floodplain. This measure of flood risk is the most salient information to homebuyers since mortgage lenders require supplemental flood insurance for properties located in the 100 year floodplain. This fact alone should generate a negative relationship between being in a floodplain and home prices.

Flood Factor

The First Street Foundation has developed and refined a dynamic hydrology model based on the work of leading researchers in climate science and hydrology. The goal of First Street Foundation is to provide property level information on future risk of flooding and subsequently wildfire and heat. This information is designed to help in planning for both residents and businesses. Our focus is on the First Street Foundation Flood Model V2. The key features of this model is that it incorporates simulated natural disaster and flood events over the next 30 years based on meteorological predictions under various climate models with detailed digital evaluation models. The model is forward-looking in nature and incorporates both historic flood risk as well as climate change driven changes in flood risk. The corresponding dataset from the First Street Foundation provides a composite measure of future flood risk based on a score between 1 and 10 with larger values indicating greater future flood risk. Each measure of flood risk is based on granular measures of flood risk assigned to the centroid of the building footprint of a parcel.[18]

FEMA Floodplain

As part of their data distribution, First Street foundation provides information on parcel level location in Federal Emergency Management Agency (FEMA) designated flood zones. We designate a property to be "In a Floodplain" if FEMA has it designated within a Special Flood Hazard Area (SFHA), which is defined to be areas that will be "inundated by the flood event having a 1-percent chance of being equaled or exceeded in any given year", or sometimes referred to as "100-year flood."[19] These areas are part of the National Flood Insurance Program (NFIP) and mandatory purchase of flood insurance is required if the property is purchased using a home mortgage. Some individuals may drop insurance coverage after initial purchase but this would violate the terms of the mortgage and

[18] For more details on the First Street Foundation Flood Model see https://firststreet.org/methodology/flood .

[19] "SFHAs are labeled as Zone A, Zone AO, Zone AH, Zones A1-A30, Zone AE, Zone A99, Zone AR, Zone AR/AE, Zone AR/AO, Zone AR/A1-A30, Zone AR/A, Zone V, Zone VE, and Zones V1-V30." https://www.fema.gov/glossary/flood-zones, accessed July 5, 2024.

individuals without a mortgage are not required to have flood insurance.[20]

We combine the property database to the First Street Foundation dataset using closest latitude and longitudinal coordinates which are both assigned to the centroid of each parcel. This spatial merge provides a unique measure of flood risk and floodplain for each parcel which allows us to implement quite granular controls for neighborhood amenities.

3.1.3 Household income

Our candidate anti-instrument for unobserved housing quality is household income which should capture residual house-level quality even after we control for granular census geographical fixed effects and other observed housing characteristics, but plausibly conditionally independent from flood risk. We operationalize household income using the publicly available loan-level data made available under HMDA (Home Mortgage Disclosure Act). This datasets represents the income of the individual(s) purchasing a home with a mortgage. We focus only on transactions (not refinances) and the HMDA data provides loan amount, lender name, Census tract for home purchase, year of purchase and stated income of borrower, as well as race, age and gender of borrower. Since HMDA does not provide transaction level identifiers, this merge requires fuzzy matching on transaction year, lender name, loan amount, and Census tract. In practice, few transactions are non-unique on these dimensions. Fifty-five percent of our transaction sample merges uniquely with the HMDA data. About 60% of these unmatched transactions are due to property purchases that did not use a mortgage.[21]

3.1.4 Other Data Sources

Finally, we supplement our property level database using spatial information available from the US Geological Society (USGS). We provide parcel centroid-based measures of elevation using digital elevation models from USGS with 10m by 10m resolution. Elevation is intended to capture amenities associated with views from the home. We also provide a measure of distance to the coast for properties in coastal counties. This is based on waterline maps from the USGS and we include a measure of Euclidean distance to the coast. This measure of distance to coast captures the amenities of living near the beach and we typically bin distance to coast into increasing larger distance intervals based on miles, namely, $[0, 0.05), [0.05, 0.1), [0.1, 0.2), [0.2, 0.5), [0.5, 1), [1, 1.5), [1.5, 2)$ and greater than 2 miles.

[20] Vossler and Holladay (2018) survey residents in New York and find that for those in a 100-year floodplain, 90% of households with a mortgage have flood insurance while 60% of households without a mortgage had flood insurance.

[21] See Billings (2019) for technical details on the HMDA matching procedure to property datasets.

3.2 Summary Statistics

Table 1 summarizes our main dataset after merging the public records on housing transactions with property-level measures of flood risk and household income. We have 9,132,510 property records. The median home price (in 2020 dollars) is $165k and the median annual household income is $56k. Because household income requires us to rely on only home buyers who take out a mortgage, it misses cash buyers. However, a priori, it seems unlikely that cash buyers are strongly selected on their valuation of flood risk.

Figure 1 shows the distribution of Flood Factor assigned by First Street at the property level. Flood Factor reports on a 1 to 10 scale based on both the probability and the depth of flooding projected for the next 30 years. 1 represents minimal risk, while 10 represents extreme risk. Figure 1 shows that the vast majority of the homes (82%) are assigned a Flood Factor of 1, representing minimal risk. On the other hand, 1.9%, 1.8%, and 0.7% of homes are categorized at a Flood Factor of 8, 9, and 10, respectively, representing a small but still significant fraction of the overall housing stock, amounting to over 400,000 properties.

House-level characteristics available include the total number of bathrooms, the year built, total square footage, the lot size, and the distance to the coast. In cases where property attributes are missing, we impute the value to be 0 and include a missing indicator for that variable.

3.3 Results

In this section, we illustrate that the anti-instrumental variable method can be used to obtain sensible estimates of the price of flood risk. We demonstrate how it works in both the linear case as well as how to extend it to the more general non-parametric case. We show that the anti-IV-based estimates are able to address the "wrong-signed" problems that seem to be pervasive in the estimation of similar environmental amenities. We also compare that the estimates with other existing reduced-form estimates using extensive hedonic controls.

3.3.1 Estimates of Flood Risk in Linear Models

If we assume that observed housing amenities and unobserved quality enter linearly into prices

$$P = \beta_0 + \beta_z Z + \beta_x X + \Phi \tag{3.1}$$

where Z represents the focal amenity of flood risk and X represents all other observed controls, then we can use the "ratio-of-coefficients" method outlined in Section 2.4.2 and the extension in Section

2.5.1 to obtain a single coefficient estimate of the amenity price.

Table 2 shows the main results. The set of focal amenities Z are indicators of a property's Flood Factor being at various levels from 2 to 10, with 1 being the omitted category. Column (1) shows the results from OLS regression. Even though it seemingly produces some small negative coefficients for intermediate levels of risk, it estimates an insignificant result for Flood Factor 9. Even more troubling, OLS estimates a significant, positive coefficient for the highest risk category of 10. Column (2) attempts to control for the unobserved quality directly by using the log of household income as a control. As the OLS regression, it produces an insignificant coefficient for Flood Factor 9, and a positive, significant coefficient for Flood Factor 10. Notably, though, the coefficient for Flood Factor 10 has moved significantly from 0.932% in OLS to 0.536% when income control is added, suggesting the important roles the unobservable may be playing. Column (3) estimates the prices of various Flood Factors using household income as an anti-instrument. In this case, it finds insignificant coefficients in front of moderate risk categories (2 to 7). However, the estimated coefficients become negative at Flood Factor 8. At Flood Factor 9, the estimated coefficient becomes statistically significant at -0.9%, suggesting that the implicit price of Flood Factor 9 (relative to 1) is -0.9% of the home price. At Flood Factor 10, the estimated coefficient is also statistically significant at -4.86%.

Figure 6 presents our main estimates of the flood risk by categories visually. It is striking to observe that anti-IV methods produce price estimates that are rapidly decreasing when approaching the highest risks, which stands in sharp contrast with the flat lines associated with OLS estimates, even with extensive controls. It is also notable that the estimator had no inherent knowledge of the ordering of the Flood Factor labels but is able to recover the monotonicity of the amenity price by the level of risk through the estimation itself.

In terms of non-focal amenities X, we include a rich set of controls. In particular, we allow for extremely detailed geographic fixed effects at the census block level to control for neighborhood-level characteristics. Census blocks are generally small in area, where "in a city, a census block looks like a city block bounded on all sides by streets," and are the smallest level of geography in the US census.[22] As a result, allowing for census block fixed effects amounts to controlling for a wide variety of neighborhood characteristics ranging from school quality, access to labor markets, transportation infrastructure, healthcare, retail, and entertainment, as well as environmental amenities such as air quality or heat risk that can be thought of as uniform within a small geography. In addition, we also include a vector of housing-level characteristics that include the number of bathrooms, the age of the

[22] What are census blocks? *US Census Bureau*. Retrieved July 4, 2024.

structure (in bins), total square footage, and lot size. In the robustness tests, we also consider including further controls such as distance to the coast and elevation to account for within-block variations.

We use household income as the anti-instrumental variable for the unobserved quality. To the extent that there still exists residual differences in home prices after controlling for the extensive set of fixed effects, housing characteristics, and property-level flood risk, it seems plausible to attribute to the residual unobserved quality. Since quality is a desirable trait across all levels of household budgets, it is conceivable that many models of utility can generate a pattern where those with larger budgets will choose to consume more quality.[23] We think it is reasonable that reported household income at the time of mortgage application co-vary with household budget and, therefore, the level of unobserved quality chosen, making it likely a relevant anti-instrument.

To the extent that reported household income mismeasures the true unobserved quality, conditional independence requires it to be uncorrelated with their preferences for flood risk. So, how would reported income mismeasure the unobserved quality? For example, if household budgets are determined by both income and wealth and household budgets are the true driver of the amount of quality chosen, then reported income may mismeasure household budgets, which in turn mismeasures unobserved quality. In this case, conditional independence requires that, conditioning on the true budgets, how reported income mismeasures household budgets to be uncorrelated with their preferences for flood risk.

To make it more concrete, suppose household preferences for flood risk vary by political affiliation as documented in the literature (Bernstein et al., 2022), then conditional independence would require that the extent that income mismeasures household budget, conditioning at a given level of true household budget, be not different for Democrats and Republicans, which seems plausible.[24]

Another concern regarding conditional independence is that the mismeasurement of true unobserved quality by reported income could be systematically correlated with race. For instance, Blacks might have higher reported income when controlling for true wealth. *If* race is also correlated with preference for flood risk, say if Blacks were more tolerant of flood risk, then, it would *appear* that household income would become positively correlated with the amount of flood risk chosen, condition on true unobserved quality. As shown in Figure 5, such violation of conditional independence with $cov(H, Z|\Phi) > 0$ would lead to an upward biased estimate of the amenity price. Given that HMDA data also collects

[23] To make it more concrete, we provide a simulated market equilibrium using Cobb-Douglas utility in Appendix B to illustrate how unobserved housing quality is determined by household budget, for which reported income is a noisy measure.

[24] If one were to collect such voter registration data, then one could test whether the estimated amenity prices are sensitive to including political affiliation as an additional control.

information about applicant race, we re-run the anti-IV estimate where we also included a flag for racial minorities as a control. We find the estimate of price remains entirely robust to it, with an estimated price for Flood Factor 10 at -4.79% (table not shown for brevity).

3.3.2 Estimates of FEMA Floodplain

While our main analysis focuses on using property-level flood factor as the measure of the flood risk, we also conduct our analysis using other measures of flood risk.

Specifically, we employ the same anti-IV estimation framework to estimate the implicit price of a property being included in the FEMA 100-year floodplain. The advantage of using FEMA designated floodplains is that homeowners likely know better whether their home resides inside a floodplain because of the mandatory insurance requirement. The disadvantage is that the floodplain designation is somewhat coarse in that a property is either in the floodplain or not. So, if a given property is at higher or lower risk than nearby properties, the floodplain designation is a noisy measure of the true underlying flood risk.

Table 3a shows that FEMA floodplain designation is highly correlated with Flood Factor estimates. For example, 70% of the properties with a Flood Factor of 10 reside in a floodplain, whereas only 9% of the properties with a Flood Factor of 1 reside in a floodplain. The fraction in floodplain at a given Flood Factor is monotonically increasing at higher Flood Factor categories (above 6), but the relationship is weak at lower Flood Factor categories (2-5). In general, the data suggest Flood Factor and floodplain designation are indeed highly correlated, but important differences still remain.

Table 3b summarizes the estimation results, using household income as the anti-IV. Column (1) shows the conventional OLS regression with census block fixed effect and a vector of house-level characteristics as controls. It produces a wrong-signed estimate of positive price of flood risk at 0.48%, suggesting that endogeneity still remains even after controlling for rich hedonics. Column (2) adds log household income as the additional control for unobserved quality, still producing a positive but somewhat smaller coefficient at 0.26%. Column (3) estimates the implicit price of being in a floodplain using log household income as the anti-IV, producing a correctly-signed estimate of -2.8%, suggesting that the implicit cost of being in a floodplain is about 2.8% of the home price. It is also a magnitude that is internally consistent with the estimate based on Flood Factor in Table 2 where the highest Flood Factor 10 costs about 4.9%, but we know a significant fraction of properties in a floodplain are at Flood Factor below level 10.

A priori, given the mandatory insurance requirement, it seems sensible that the implicit price of being included in a floodplain should be negative. Given that the insurance payout is unlikely to cover

the entire financial cost of experiencing flood, either because of insurance coverage limits or because of other costs associated with flood-related disruptions, it would lead to a further negative price. A simple measure of the cost of flood insurance is available from FEMA and finds that on average, homeowners spend $\$1,290$ annually on flood insurance premiums as part of the National Flood Insurance Program (FEMA, 2023). One could generate a reasonable present value of these stream of payments using the annuity discount rate given by Poterba and Solomon (2021) of 8% and an average length of homeownership at given residence of about 9 years during our years of study (Mateyka and Marlay, 2010). These premiums generate a present value of $\$8,058$. Given our estimates of 2.84% in Table 3b, and an average home price of $\$212,075$ generates a home price discount that is fairly close to the present value of these premiums at $\$6,023$. While we caution against doing a direct comparison of the insurance costs and reported property damages with the estimated amenity price due to the inherent conceptual difference, as the implicit price is driven by the marginal owner's willingness pay rather than the average costs incurred, it still provides us a sense over how the marginal owner values the amenity in comparison to the average cost.

3.4 Robustness

While we employ detailed census block fixed effects and census blocks tend to be small in area in general, there remains a concern that within-block geographies that the OLS specification should control for. Such amenities could also be correlated with both quality and flood risk, thus biasing the OLS estimates. For example, distance to the coast for coastal properties may capture the desirable amenity of having easy access to the beach, but it may also come with higher flood risks. Elevation may also capture the desirable amenity of having a nice view, but it may be associated with lower flood risks.

Empirically, if there remain substantial residual variations in these geographic features within census blocks, then omitting such controls can introduce bias to the conventional OLS estimates. On the other hand, whether omitting them would bias the anti-IV estimate depends on whether the candidate anti-instrument remains conditionally independent with the focal amenity without these controls. Concretely, in our example, it would have to be that the extent to which household income mismeasures underlying quality varies systematically by distance to beach or elevation, which seems unlikely.

Table 4 shows the estimates of flood risk with additional information on distance to the coast among coastal properties. To capture potential non-linearity in the relationship between distance to the coast and unobserved quality, we split the distance to coast into bins at $[0, 0.05)$, $[0.05, 0.1)$, $[0.1, 0.25)$, $[0.25, 0.5)$, $[0.5, 2)$ and more than 2 miles. Compared to our main results in Table 2,

controlling for the distance to coast (in miles) does not affect the price estimates in either OLS or anti-IV, suggesting once block fixed effects are controlled for, distance to coast does not appear to be a major confounder. Column (3) shows that the estimated price of Flood Factor 10 becomes -4.7%, compared to -4.9% without the additional distance controls. Table 4 also shows the estimates of flood risk with additional information on elevation. To capture potential non-linearity in the relationship between elevation and unobserved quality, we split elevation (measured in meters) into bins at $[0, 1)$, $[1, 2)$, ..., $[9, 10)$, $[10, 50)$, $[50, 100)$, $[100, 500)$ and higher than 500 meters. Column (6) shows that the estimated price of flood is also robust to controlling for elevation bins, with the price of Flood Factor 10 at -5.0%.

3.4.1 Estimates of Flood Risk in Non-Parametric Models

Because the anti-IV framework is not reliant on the linearity assumption, in this section, we illustrate how the method is applied when we do not impose function form assumptions on the equilibrium hedonic pricing function $P(Z, \Phi)$.

The top panel of Figure 7 shows the empirical analog of the "Rosen" surface in the non-parametric framework. Specifically, we divide all homes into 11 x 10 cells, each representing homes within a certain home price quantile (1-11) and Flood Factor category (1-10). We approximate the predicted anti-IV function $\hat{H}(P, Z) = \mathbb{E}[H|P, Z]$ by simply taking the average of the anti-IV H, namely household income, in each of the (P, Z) cells. Because we would want to condition on geographic fixed effects, we compute the average by running a regression of the anti-IV on the cell indicators with geographic fixed effects.[25] The resulting heat map plots cells with higher average incomes with lighter colors (yellow) and cells with lower average incomes with darker colors (purple). A visual inspection of the color pattern suggests that we can approximate the level sets of the anti-IV function $\hat{H}$ by grouping together cells of the same color, which generally consists of downward-sloping curves, indicating a negative price for flood risk.

The bottom panel of Figure 7 provides the local estimate of the amenity price using the estimation strategy described in Section 2.4.1. Specifically, at each cell (P, Z), we compute how $\hat{H}$ varies from one cell to the left to one cell to the right to estimate $\frac{\partial \hat{H}}{\partial Z}$. Analogously, how $\hat{H}$ varies from one cell from above to one cell from below to estimate $\frac{\partial \hat{H}}{\partial P}$. The negative ratio of the two $-\frac{\partial \hat{H}}{\partial Z}/\frac{\partial \hat{H}}{\partial P}$ provides a local estimate of the implicit price of incremental flood risk at that cell. We also visualize the magnitude of the estimated coefficients with colors: consistent with the linear estimate, we find significantly

[25] In Stata, it amounts to *reg H i.price#i.flood, absorb(FE)*. Note that these estimates are based on controlling for block-group fixed effects rather than block fixed effects to allow for a sufficient number of observations in each bin.

more negative prices of flood at high Flood Factor categories (e.g., the price of increasing the Flood Factor from 8.5 to 9.5.) and relatively close to zero for lower Flood Factor categories (e.g., the price of increasing the Flood Factor from 3.5 to 4.5). Moreover, we do not seem to find much variation across the quality segments, where the estimated prices in the rightmost column do not seem to vary significantly. This suggests, at least when measured as a percentage of prices, we do not find evidence that the capitalization rate of flood risk varies significantly across the quality segment.

3.5 Beyond Flood Risk

While the focus of this section is to illustrate how to apply the anti-IV method in estimating the price of flood, in theory, the proposed method could also help estimate the price of other residential amenities, especially other environmental amenities that are also known to be prone to the wrong-signed problem. However, as we discussed in Section 2.5.3, the anti-IV method, while well-suited to address issues related to the unobserved vertical quality, could still be biased if there exists missing amenities that are otherwise required by conditional independence. As a result, to estimate the price of flood risk, we have generally opted for a specification where we still employ granular geographic fixed effects to control for potentially missing neighborhood amenities. Indeed, Figure 8 shows that when we include coarser geographic fixed effects (e.g., census tract or county) than the most granular block fixed effects, the estimated price for flood becomes more negative.[26]

Because controlling for missing amenities at the neighborhood level would require extensive geographic fixed effects, it would, in turn, limit the set of relevant amenities of interest (e.g., heat risk) that do not have much residual variations within small geographic areas. That said, as another proof of concept, in Appendix D.5, we present a second application of the anti-IV method to estimating the price of air pollution where we use a novel PageRank-based, county-level measure of place quality as the candidate anti-IV. However, since there is limited variation in air quality within a county, we are inherently more limited in our ability to control for granular geographic fixed effects and thus lean more heavily on a more carefully constructed anti-IV where conditional independence could still hold.

4 Conclusion

In this paper, we develop a novel anti-instrumental variable approach that allows us to price a wide range of amenity trade-offs in the housing market in the midst of unobserved quality. Our estimator can be applied to commonly used observational datasets, and proves that the typical confounder of

[26] A potential explanation for why the inclusion of more granular fixed effects results in more muted prices is that, conditional on quality, other disamenities tend to come bundled with flood risk and "load onto" the coefficient for flood risk.

unobserved vertical quality that is to blame for classic "wrong-signed" results should no longer be viewed as endemic to observational approaches. These trade-offs are identified under a notably lax set of assumptions, central to which is that the researcher has available to her a rough "anti-instrument" for quality.

To investigate whether such a viable anti-instrument for quality might exist, we leverage property-level household income to help us account for unobserved house quality. The identification assumption is that, after including an extensive set of geographic fixed effects to control for variations in neighborhood-level amenities and another set of controls to account for house-level characteristics, household income is conditionally independent of flood risk.

Unlike most other estimates one could have obtained from observational data—which tend to be oppositely signed or imprecisely estimated, the anti-IV approach produces statistically significant estimates whereby the marginal American home-buyer pays 4.9% *less* to own a property with the highest flood risk at a Flood Factor of 10. The magnitude of our estimate also monotonically increases with the levels of projected flood risk. Together, these findings lend credibility to our anti-instrument approach and promise that the anti-instrument approach can be used more widely in economics, especially in hedonic models of prices. In a literature where finding credible estimates of flood risk has been challenging, our estimates will prove relevant for understanding the economic impact of flood risk, or more broadly, climate change, on residential real estate, a major segment of the economy.

The anti-instrumental variable approach is not without limitations. While the presence of a hedonic pricing function, as required by the Single Index Assumption, could be a sensible approximation to the housing market, it may not apply in other settings, especially in markets where a significant portion of prices are driven by idiosyncratic variation. Moreover, it also depends on the availability of a candidate anti-instrument that can plausibly satisfy conditional independence.

Still, we believe that the anti-IV method provides an important tool in a researcher's arsenal to address endogeneity issues that arise naturally when prices are jointly determined by observed and unobserved characteristics. If anything, given the prevalence of traditional instrumental variable approaches, researchers have always been looking for variables that can instrument for the endogenous, *observed* characteristic. Rather, our research shines a spotlight on a previously overlooked area, where we suggest searching for candidate variables that may instrument for the endogenous, *unobserved* error, opening up new sources of identification for these problems.

References

Altonji, Joseph G, Timothy Conley, Todd E Elder, and Christopher R Taber. 2024. "Methods for Using Selection on Observed Variables to Address Selection on Unobserved Variables1."

Altonji, Joseph, Todd Elder, and Christopher Taber. 2005. "Selection on Observed and Unobserved Variables: Assessing the Effectiveness of Catholic Schools." *Journal of Political Economy*.

Anderson, T. W., and Herman Rubin. 1950. "The Asymptotic Properties of Estimates of the Parameters of a Single Equation in a Complete System of Stochastic Equations." *The Annals of Mathematical Statistics*, 21(4): 570–582. Publisher: Institute of Mathematical Statistics.

Andrews, Isaiah, James H. Stock, and Liyang Sun. 2019. "Weak Instruments in Instrumental Variables Regression: Theory and Practice." *Annual Review of Economics*, 11(Volume 11, 2019): 727–753. Publisher: Annual Reviews.

Bajari, Patrick, and C Lanier Benkard. 2005. "Demand Estimation with Heterogeneous Consumers and Unobserved Product Characteristics: A Hedonic Approach." *Journal of Political Economy*.

Baldauf, Markus, Lorenzo Garlappi, and Constantine Yannelis. 2020. "Does climate change affect real estate prices? Only if you believe in it." *The Review of Financial Studies*, 33(3): 1256–1295.

Bayer, Patrick, Fernando Ferreira, and Robert McMillan. 2007. "A Unified Framework for Measuring Preferences for Schools and Neighborhoods." *Journal of Political Economy*, 115(4): 588–638. Publisher: The University of Chicago Press.

Bell, Alex. 2022. "Job Amenities and Earnings Inequality." Available at SSRN: http://dx.doi.org/10.2139/ssrn.4173522.

Beltrán, Allan, David Maddison, and Robert Elliott. 2019. "The impact of flooding on property prices: A repeat-sales approach." *Journal of Environmental Economics and Management*, 95: 62–86.

Bernstein, Asaf, Matthew T Gustafson, and Ryan Lewis. 2019. "Disaster on the horizon: The price effect of sea level rise." *Journal of financial economics*, 134(2): 253–272.

Bernstein, Asaf, Stephen B. Billings, Matthew T. Gustafson, and Ryan Lewis. 2022. "Partisan residential sorting on climate change risk." *Journal of Financial Economics*, 146(3): 989–1015.

Berry, Steven, and Ariel Pakes. 2007. "The pure characteristics demand model." *International Economic Review*, 48(4): 1193–1225.

Berry, Steven, James Levinsohn, and Ariel Pakes. 1995. "Automobile prices in market equilibrium." *Econometrica*, 63(4): 841–890.

Berry, Steven, James Levinsohn, and Ariel Pakes. 2004. "Differentiated Products Demand Systems from a Combination of Micro and Macro Data: The New Car Market." *Journal of Political Economy*, 112(1): 68–105.

Billings, Stephen B. 2019. "Technical summary - merging home mortgage disclosure data to property records from zillow (ztrax) 1995–2016." https://sites. google.com/a/colorado.edu/stephen-billings/code.

Bin, Okmyung, and Craig E Landry. 2013. "Changes in implicit flood risk premiums: Empirical evidence from the housing market." *Journal of Environmental Economics and management*, 65(3): 361–376.

Bishop, Kelly C., Nicolai V. Kuminoff, H. Spencer Banzhaf, Kevin J. Boyle, Kathrine Von Gravenitz, Jaren C. Pope, V. Kerry Smith, and Christopher D. Timmins. 2020. "Best Practices for Using Hedonic Property Value Models to Measure Willingness to Pay for Environmental Quality." *Review of Environmental Economics and Policy*, 14(2): 260–281.

Burbano, Vanessa, Olle Folke, Stephan Meier, and Johanna Rickne. 2022. "The Gender Gap in Meaningful Work: Explanations and Implications." *CEPR Discussion Papers*. Number: 17634 Publisher: C.E.P.R. Discussion Papers.

Calder-Wang, Sophie. 2021. "The Distributional Impact of the Sharing Economy on the Housing Market."

Chay, Kenneth Y., and Michael Greenstone. 2005. "Does Air Quality Matter? Evidence from the Housing Market." *Journal of Political Economy*, 113(2): 376–424. Publisher: The University of Chicago Press.

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.

Currie, Janet, and Matthew Neidell. 2005. "Air Pollution and Infant Health: What Can We Learn from California's Recent Experience?*." *The Quarterly Journal of Economics*, 120(3): 1003–1030.

Ekeland, Ivar, James J. Heckman, and Lars Nesheim. 2004. "Identification and Estimation of Hedonic Models." *Journal of Political Economy*, 112(S1): S60–S109. Publisher: The University of Chicago Press.

Epple, Dennis, and Holger Sieg. 1999. "Estimating Equilibrium Models of Local Jurisdictions." *Journal of Political Economy*, 107(4): 645–681. Publisher: The University of Chicago Press.

Epple, Dennis, Luis Quintero, and Holger Sieg. 2020. "A New Approach to Estimating Equilibrium Models for Metropolitan Housing Markets." *Journal of Political Economy*, 128(3): 948–983. Publisher: The University of Chicago Press.

FEMA. 2023. "Cost of Flood Insurance for Single-Family Homes under NFIP's Pricing Approach." https://www.fema.gov/flood-insurance/work-with-nfip/risk-rating/single-family-home.

Fogel, Jamie. 2021. "Valuing American Cities: A Revealed Preference Approach."

Folke, Olle, and Johanna Rickne. 2022. "Sexual Harassment and Gender Inequality in the Labor Market*." *The Quarterly Journal of Economics*, 137(4): 2163–2212.

Galiani, Sebastian, Alvin Murphy, and Juan Pantano. 2015. "Estimating Neighborhood Choice Models: Lessons from a Housing Assistance Experiment." *American Economic Review*, 105(11): 3385–3415.

Gibson, Matthew, and Jamie T Mullins. 2020. "Climate risk and beliefs in New York floodplains." *Journal of the Association of Environmental and Resource Economists*, 7(6): 1069–1111.

Giglio, Stefano, Bryan Kelly, and Johannes Stroebel. 2021. "Climate Finance." *Annual Review of Financial Economics*, 13(Volume 13, 2021): 15–36. Publisher: Annual Reviews.

Glaeser, Edward L., Jed Kolko, and Albert Saiz. 2001. "Consumer city." *Journal of Economic Geography*, 1(1): 27–50.

Goldberger, Arthur S. 1984. "Reverse Regression and Salary Discrimination." *The Journal of Human Resources*, 19(3): 293–318.

Heckman, James J., Rosa L. Matzkin, and Lars Nesheim. 2010. "Nonparametric Identification and Estimation of Nonadditive Hedonic Models." *Econometrica*, 78(5): 1569–1591. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3982/ECTA6388.

Kuminoff, Nicolai V., and Jaren C. Pope. 2014. "Do "Capitalization Effects" for Public Goods Reveal the Public's Willingness to Pay?" *International Economic Review*, 55(4): 1227–1250. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/iere.12088.

Kuminoff, Nicolai V., Christopher F. Parmeter, and Jaren C. Pope. 2010. "Which hedonic models can we trust to recover the marginal willingness to pay for environmental amenities?" *Journal of Environmental Economics and Management*, 60(3): 145–160.

Mateyka, Peter, and Matthew Marlay. 2010. "The duration and tenure of residence, 1996 to 2009."

Murfin, Justin, and Matthew Spiegel. 2020. "Is the risk of sea level rise capitalized in residential real estate?" *The Review of Financial Studies*, 33(3): 1217–1255.

Neidell, Matthew J. 2004. "Air pollution, health, and socio-economic status: the effect of outdoor air quality on childhood asthma." *Journal of Health Economics*, 23(6): 1209–1236.

Ortega, Francesc, and Süleyman Taşpınar. 2018. "Rising sea levels and sinking property values: Hurricane Sandy and New York's housing market." *Journal of Urban Economics*, 106: 81–100.

Oster, Emily. 2019. "Unobservable Selection and Coefficient Stability: Theory and Evidence." *Journal of Business & Economic Statistics*, 37(2): 187–204.

Page, Lawrence. 2001. "Method for node ranking in a linked database."

Poterba, James M, and Adam Solomon. 2021. "Discount rates, mortality projections, and money's worth calculations for US individual annuities." National Bureau of Economic Research.

Rosen, Sherwin. 1974. "Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition." *Journal of Political Economy*, 82(1): 34–55. Publisher: The University of Chicago Press.

Sorkin, Isaac. 2018. "Ranking Firms Using Revealed Preference*." *The Quarterly Journal of Economics*, 133(3): 1331–1393.

Stepner, Michael. 2013. "BINSCATTER: Stata module to generate binned scatterplots." *Statistical Software Components*. Publisher: Boston College Department of Economics.

Sullivan, Daniel M, and Alan Krupnick. 2018. "Using Satellite Data to Fill the Gaps in the US Air Pollution Monitoring Network."

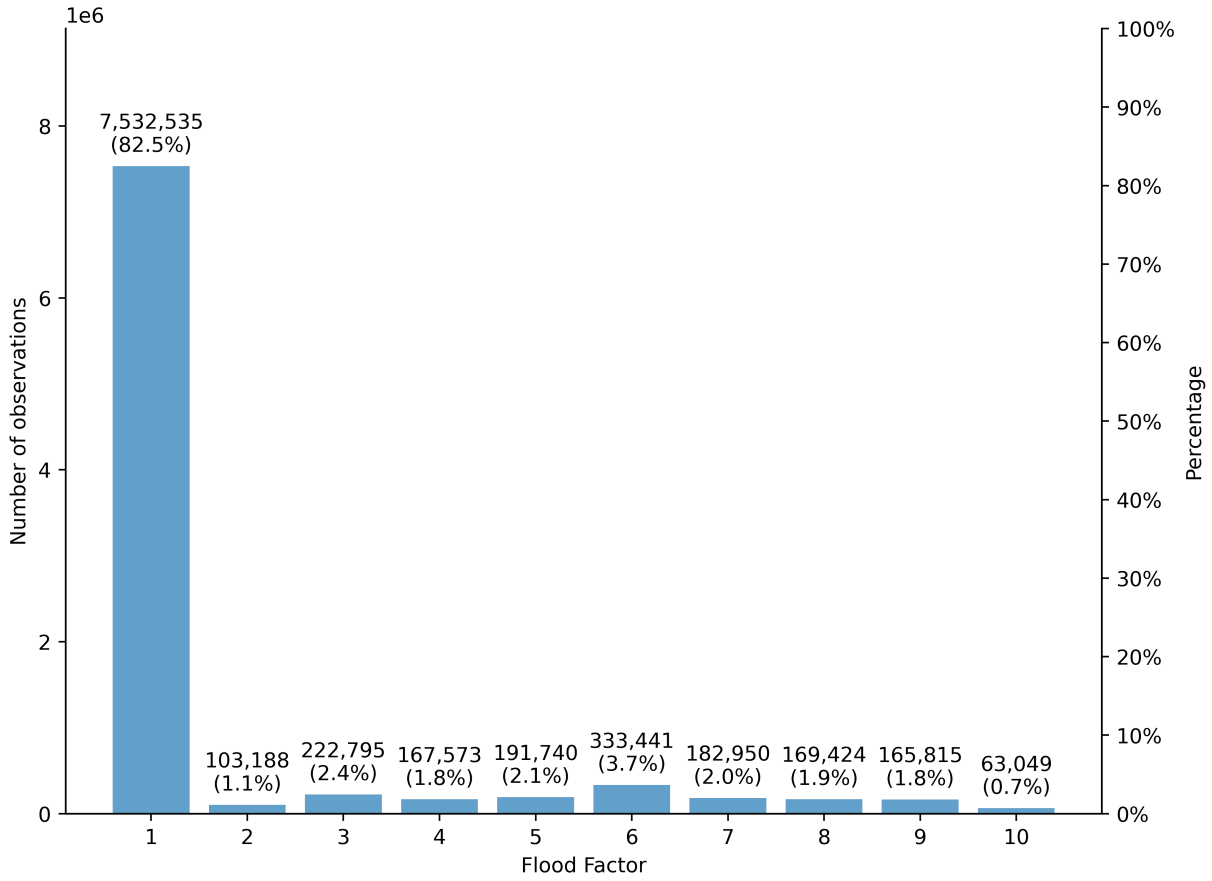
Vossler, Christian A, and J Scott Holladay. 2018. "Alternative value elicitation formats in contingent valuation: Mechanism design and convergent validity." *Journal of Public Economics*, 165: 133–145.

Wooldridge, Jeff. 2007. "What's New in Econometrics? Lecture 6 Control Functions and Related Methods." *NBER Methods Lecture*.

Zivin, Joshua Graff, Yanjun Liao, and Yann Panassie. 2023. "How hurricanes sweep up housing markets: Evidence from Florida." *Journal of Environmental Economics and Management*, 118: 102770.

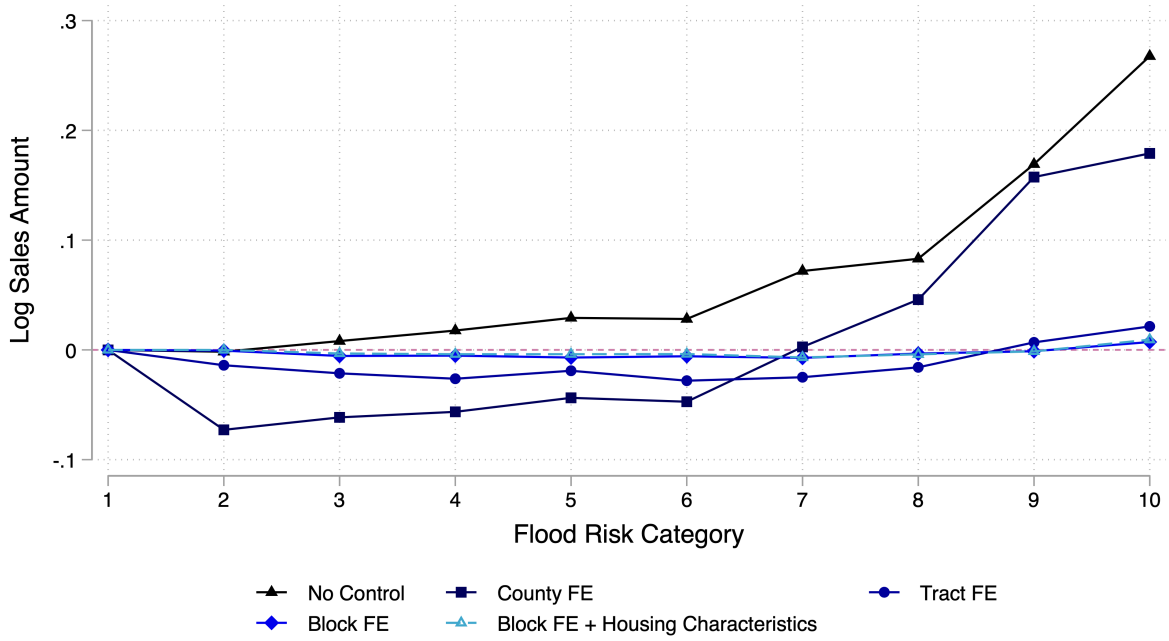
5 Figures

Figure 1: Distribution of Flood Risk Factor



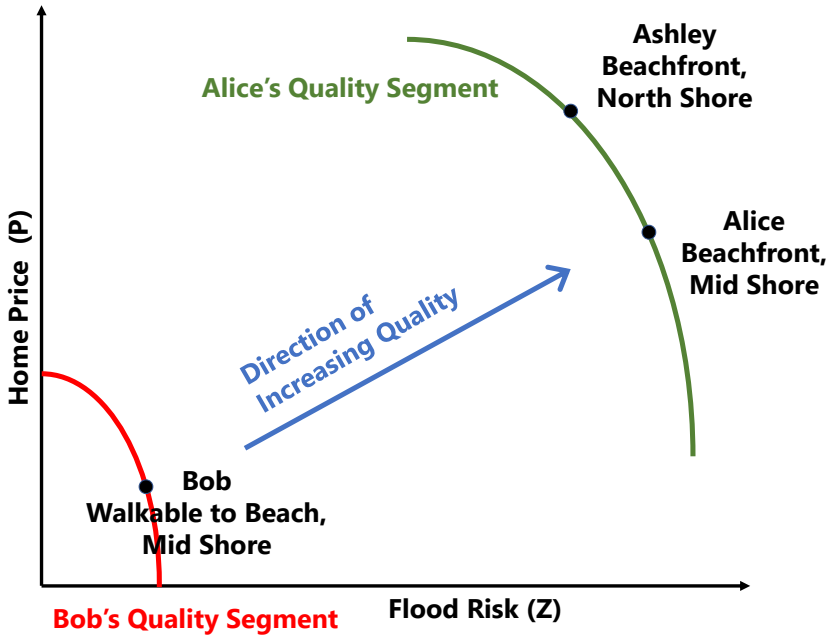
Notes: This figure shows the distribution of flood risk factors at the property level.

Figure 2: Illustration of the Perverse Sign Problem



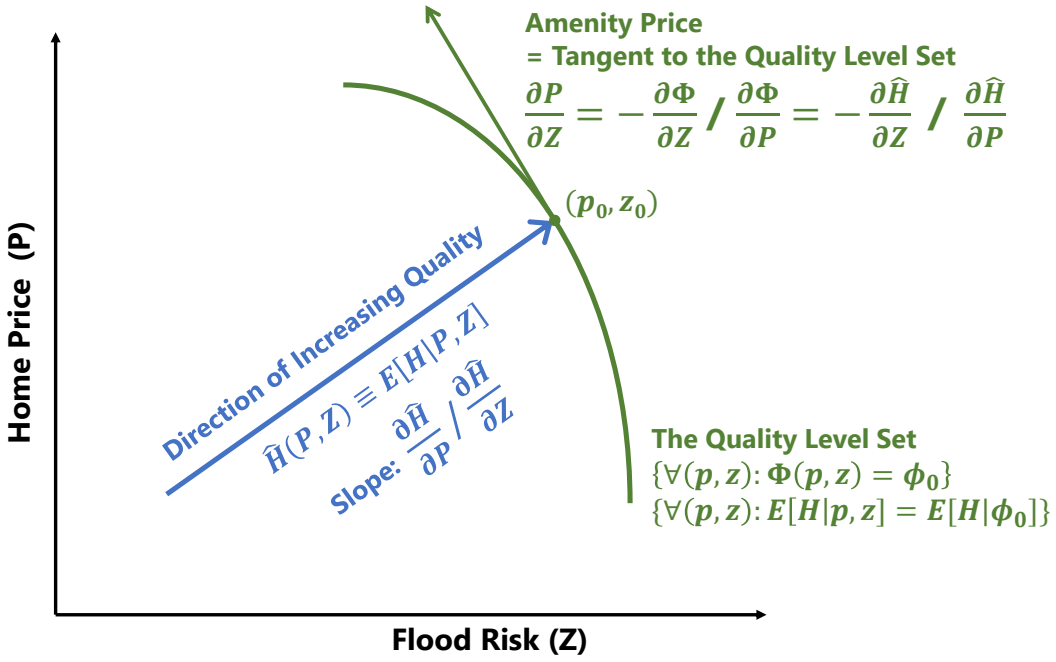
Notes: The black triangle dots plot the average coefficient when regressing log home prices on various Flood Risk categories (2-10), with 1 being the omitted category using home sales data from 2007 to 2017. We subsequently add controls for county fixed effects, census tract fixed effects, census block fixed effects, and an array of housing characteristics (e.g., number of bathrooms, building age deciles, total square footage, lot size). In all cases, the OLS regressions always a positive price for the highest flood risk category, which is wrong-signed.

Figure 3: Illustration of the Identification Issue from Unobserved Quality in the Hedonic Model



Notes: Alice's more expensive oceanfront home has high flood risk, while Bob's similar home, ten minutes from the shore, has lower flood risk. A simple regression would incorrectly estimate positive prices for flood risk. To account for such unobserved differences, like better beach access in Alice's home, one could compare homes of similar quality, such as Ashley's oceanfront home, with lower flood risk. However, fully controlling for all aspects of a house is challenging because houses are inherently unique, a limitation that conventional hedonic regressions face in estimating the price of such environmental amenities.

Figure 4: Illustration of Identification Using the Anti-Instrumental Variable Approach



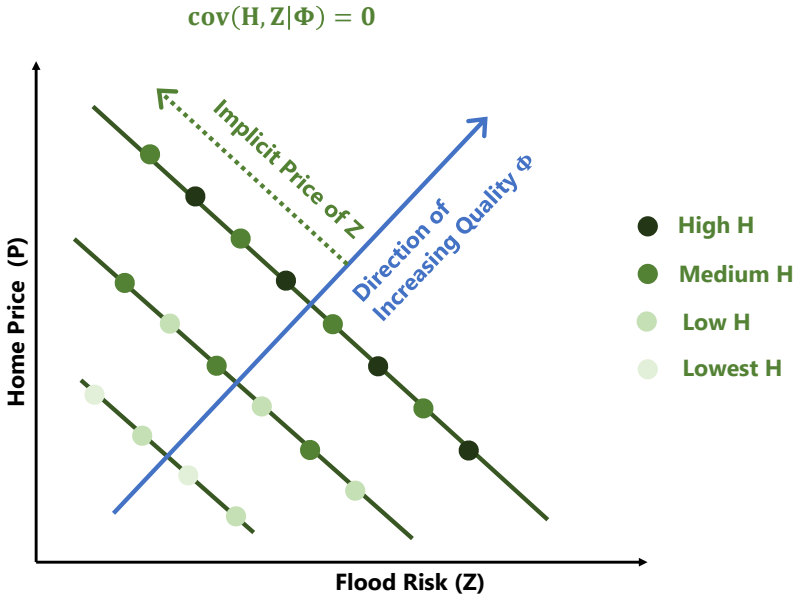
Notes: The figure above illustrates the intuition behind the identification strategy. By definition, the price of the amenity at any given observed (p_0, z_0) is determined by how prices change in response to a change in amenity, holding its unobservable quality $\phi_0 = \Phi(p_0, z_0)$ fixed. Thus, that the amenity price can be represented by the tangent direction of the level set of $\Phi(p, z)$ valued at ϕ_0, as shown above.
When there exists an anti-instrumental variable H that satisfies Conditional Independence, Lemma 2.4 says that the predicted anti-IV satisfies $\hat{H}(p, z) \equiv \mathbb{E}[H|p, z] = \mathbb{E}[H|\Phi = \Phi(p, z)]$. This means that the level set of the predicted anti-IV function $\hat{H}(\cdot, \cdot)$ valued at $\mathbb{E}[H|\phi_0]$ is *identical* to the level set of $\Phi(\cdot, \cdot)$ when valued at ϕ_0, thus producing the same tangent vector. Therefore, the amenity price can be estimated as the tangent vector of the level set of the predicted anti-IV function, namely,

$$\left.\frac{\partial P}{\partial Z}\right|_{(z_0, \phi_0)} = -\left.\frac{\partial \Phi}{\partial Z} \Big/ \frac{\partial \Phi}{\partial P}\right|_{(p_0, z_0)} = -\left.\frac{\partial \hat{H}}{\partial Z} \Big/ \frac{\partial \hat{H}}{\partial P}\right|_{(p_0, z_0)}. \tag{5.1}$$

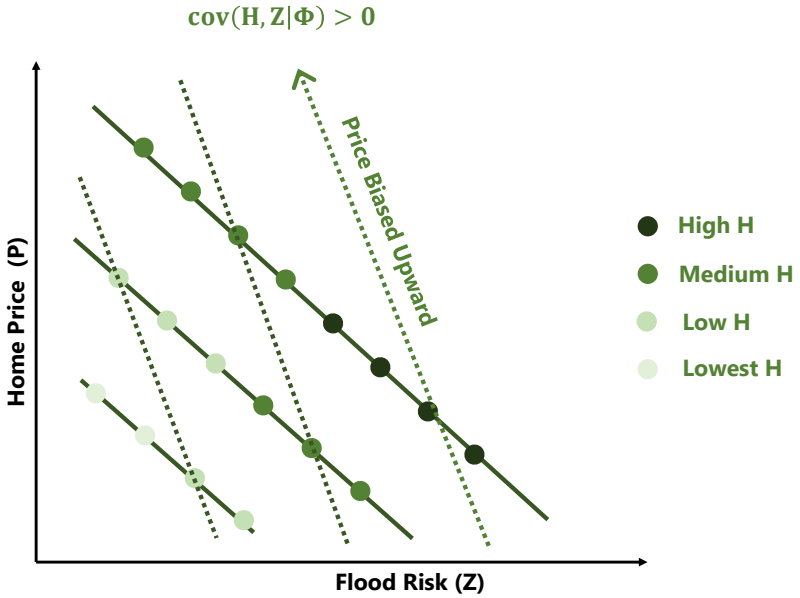
Moreover, the tangent direction is by definition perpendicular to the gradient of the predicted anti-IV function $\nabla \hat{H} = [\partial \hat{H}/\partial Z, \partial \hat{H}/\partial P]$, where the gradient represents the direction of increasing quality.

Figure 5: Conditional Independence

(a) Conditional Independence Satisfied $cov(H, Z|\Phi) = 0$



(b) Conditional Independence Violated $cov(H, Z|\Phi) = 0$



Notes: This figure illustrates the intuition behind how a violation of conditional independence leads to a biased estimate. The top panel shows that when the anti-IV and the focal amenity are conditionally independent, variations in the anti-IV H are not correlated with the focal amenity Z within each quality segment. However, in the bottom panel, when the anti-IV H and the focal amenity Z are positively correlated conditional on unobserved quality Φ, then, within each quality segment, higher H are more likely to appear at higher levels of Z. Consequently, the level set of H, namely, the set of P, Z that would generate the same H, would appear steeper than the true level set of Φ, leading to an upward biased estimate.

Figure 6: Pricing Flood Risk Using Anti-Instruments

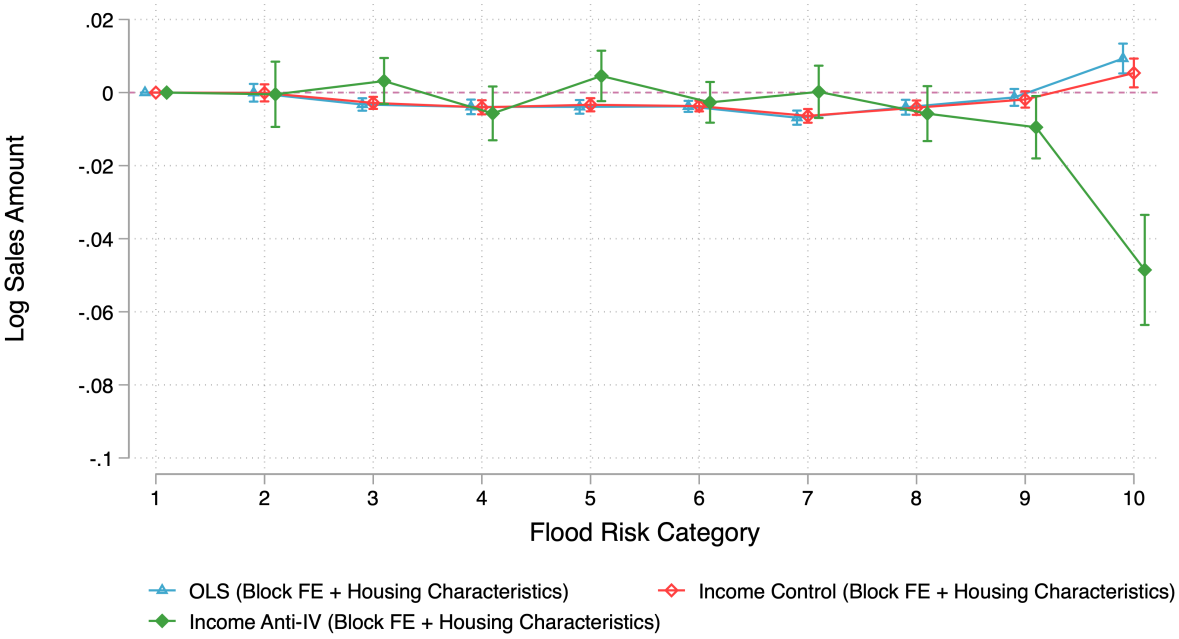
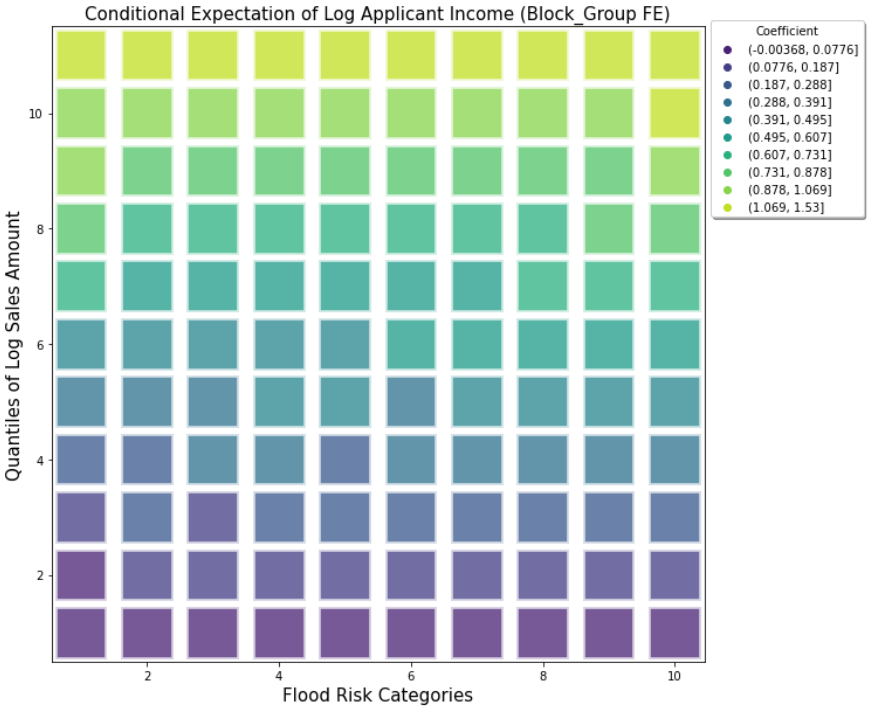
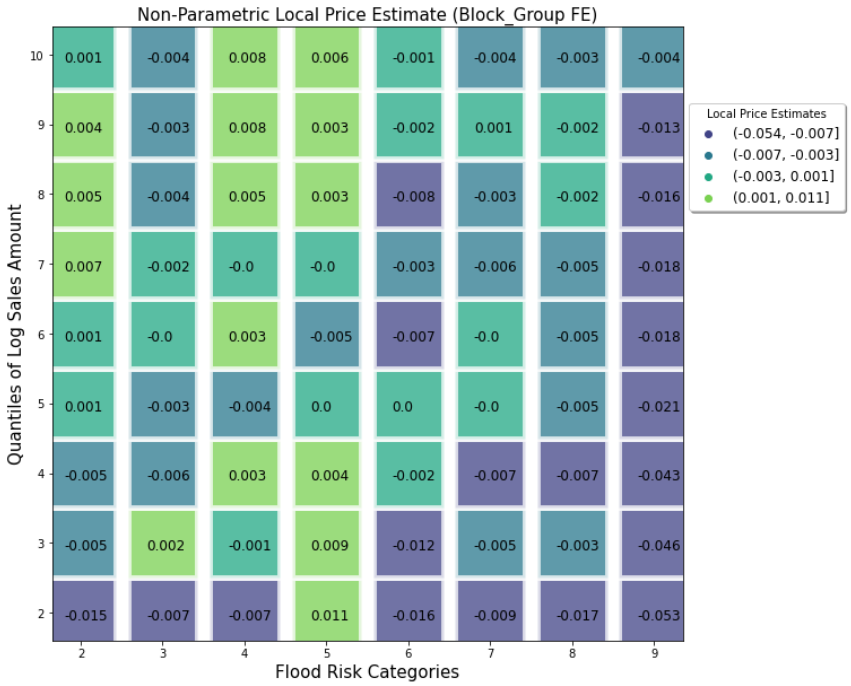


Figure 7: Non-parametric Estimation of Amenity Prices

(a) Average Household Income

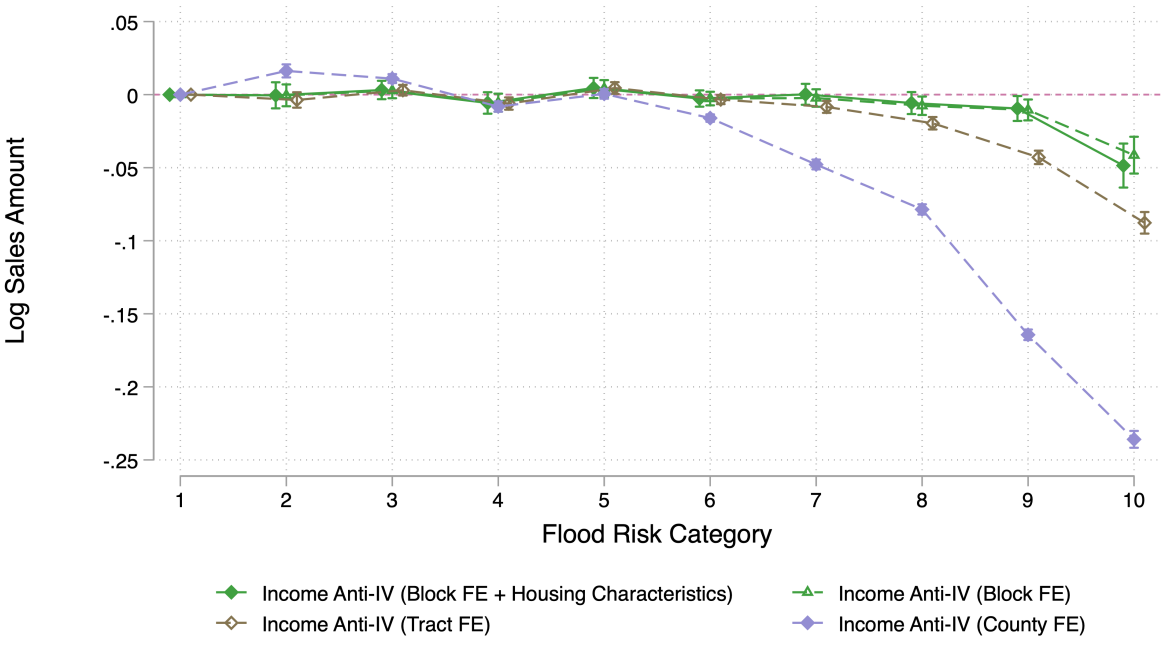


(b) Local Price Estimates



Notes: This figure categorized the homes into 110 (11x10) cells. Each cell represents homes that lie within certain house price quantile and flood factor category. The top panel plots the predicted anti-IV function, computed as the average household income in that cell, controlling for block-group fixed effects. The bottom panel plots the estimated price of a one-unit increase in the Flood Factor.

Figure 8: Role of Geographic Fixed Effects



6 Tables

Table 1: Summary Statistics for Main Variables

	Statistics							
	N	Mean	Stdev	10th-Percentile	25th-Percentile	Median	75th-Percentile	90th-Percentile
Sales Price (in $000)	9,132,510	212	175	70	106	164	258	396
Applicant Income (in $000)	9,132,510	78	106	26	37	56	87	138
Flood Factor	9,132,510	1.8	2	1	1	1	1	6
Flood Zone	9,113,364	.14	.35	0	0	0	0	1
YearBuilt	9,007,507	1,978	29	1,940	1,960	1,985	2,002	2,007
Full Bathrooms	8,562,877	1.7	.98	0	1	2	2	3
Total Square Feet	9,132,510	3,305	13,983	1,092	1,440	2,160	3,665	5,980
Lot Size Square Feet	7,799,017	164,773	8,141,983	3,485	6,000	8,712	16,117	45,302
Miles to Coast	715,457	.84	.56	.12	.34	.77	1.3	1.7
Elevation (m)	715,457	22	26	2.2	4.4	11	29	64

Notes: Sales price and applicant income (in $000) are harmonized to inflation adjusted values to 2020 dollars.

Table 2: Estimation of the Price of Flood Factor by Category

	Log Sales Amount							
	(1) OLS	(2) OLS	(3) OLS	(4) OLS	(5) OLS	(6) OLS	(7) Control	(8) Anti-IV
Flood Factor=1	0	0	0	0	0	0	0	0
	(.)	(.)	(.)	(.)	(.)	(.)	(.)	(.)
Flood Factor=2	-0.00166	-0.0728***	-0.0140***	-0.00945***	-0.000644	-0.0000525	-0.0000812	-0.000472
	(0.00218)	(0.00175)	(0.00142)	(0.00137)	(0.00140)	(0.00123)	(0.00119)	(0.00456)
Flood Factor=3	0.00806***	-0.0615***	-0.0213***	-0.0145***	-0.00551***	-0.00326***	-0.00282***	0.00317
	(0.00150)	(0.00120)	(0.000983)	(0.000951)	(0.000989)	(0.000870)	(0.000839)	(0.00321)
Flood Factor=4	0.0177***	-0.0564***	-0.0263***	-0.0178***	-0.00520***	-0.00391***	-0.00403***	-0.00569
	(0.00172)	(0.00138)	(0.00113)	(0.00110)	(0.00116)	(0.00102)	(0.000982)	(0.00376)
Flood Factor=5	0.0292***	-0.0437***	-0.0190***	-0.0138***	-0.00692***	-0.00391***	-0.00333***	0.00455
	(0.00161)	(0.00130)	(0.00107)	(0.00104)	(0.00109)	(0.000954)	(0.000921)	(0.00352)
Flood Factor=6	0.0282***	-0.0472***	-0.0280***	-0.0170***	-0.00580***	-0.00377***	-0.00370***	-0.00267
	(0.00123)	(0.00100)	(0.000834)	(0.000815)	(0.000876)	(0.000770)	(0.000743)	(0.00284)
Flood Factor=7	0.0719***	0.00271*	-0.0249***	-0.0175***	-0.00727***	-0.00686***	-0.00638***	0.000211
	(0.00165)	(0.00133)	(0.00110)	(0.00107)	(0.00112)	(0.000989)	(0.000955)	(0.00365)
Flood Factor=8	0.0831***	0.0458***	-0.0159***	-0.0111***	-0.00331**	-0.00401***	-0.00413***	-0.00576
	(0.00171)	(0.00138)	(0.00115)	(0.00112)	(0.00118)	(0.00104)	(0.00100)	(0.00383)
Flood Factor=9	0.169***	0.157***	0.00685***	-0.000242	-0.00119	-0.00132	-0.00188	-0.00949*
	(0.00173)	(0.00141)	(0.00124)	(0.00122)	(0.00134)	(0.00118)	(0.00114)	(0.00435)
Flood Factor=10	0.268***	0.179***	0.0214***	0.0149***	0.00721**	0.00932***	0.00536**	-0.0485***
	(0.00278)	(0.00226)	(0.00199)	(0.00200)	(0.00237)	(0.00208)	(0.00201)	(0.00769)
Log Applicant Income (in $000)							0.158***	2.312***
							(0.000215)	(0.00314)
Geographic FEs	None	County	Tract	Block_Group	Block	Block	Block	Block
Housing Characteristics	N	N	N	N	N	✓	✓	✓
N	9132510	9132435	9129615	9123033	8592563	8592563	8592563	8592563
Adjusted R^2	0.00247	0.371	0.621	0.666	0.755	0.810	0.823	-12.87

Notes: Standard errors in parentheses. Statistical significance at the 0.1%, 1%, and 5% levels are denoted by ***, **, and *, respectively.

Table 3: Alternative Measures of Flood Risk

(a) Comparison Between Floodplain and Flood Factor

Flood Factor	% In a Floodplain
1	9.50%
2	33.63%
3	27.51%
4	26.55%
5	27.78%
6	29.00%
7	37.22%
8	41.12%
9	57.17%
10	70.34%

(b) Estimation of the Price of Being in a Floodplain

	Log Sales Amount		
	(1) OLS	(2) Control	(3) Anti-IV
In a Flood Plain=1	0.00486***	0.00258***	-0.0284***
	(0.000806)	(0.000778)	(0.00298)
Log Applicant Income (in $000)		0.158***	2.313***
		(0.000215)	(0.00314)
Block Fixed Effects	✓	✓	✓
Housing Characteristics	✓	✓	✓
N	8575267	8575267	8575267
Adjusted R^2	0.810	0.823	-12.86

Notes: Standard errors in parentheses. Statistical significance at the 0.1%, 1%, and 5% levels are denoted by ***, **, and *, respectively.

Table 4: Robustness: Estimation of the Price of Flood Factor - Additional Controls

	Log Sales Amount					
	(1) OLS	(2) Control	(3) Anti-IV	(4) OLS	(5) Control	(6) Anti-IV
Flood Factor=1	0	0	0	0	0	0
	(.)	(.)	(.)	(.)	(.)	(.)
Flood Factor=2	-0.0000534	-0.0000883	-0.000564	-0.0000137	-0.0000735	-0.000888
	(0.00123)	(0.00119)	(0.00456)	(0.00123)	(0.00119)	(0.00456)
Flood Factor=3	-0.00328***	-0.00284***	0.00309	-0.00323***	-0.00283***	0.00264
	(0.000869)	(0.000839)	(0.00321)	(0.000870)	(0.000839)	(0.00321)
Flood Factor=4	-0.00397***	-0.00410***	-0.00584	-0.00389***	-0.00406***	-0.00636
	(0.00102)	(0.000982)	(0.00376)	(0.00102)	(0.000982)	(0.00376)
Flood Factor=5	-0.00399***	-0.00340***	0.00463	-0.00390***	-0.00336***	0.00404
	(0.000954)	(0.000921)	(0.00352)	(0.000954)	(0.000921)	(0.00352)
Flood Factor=6	-0.00385***	-0.00378***	-0.00279	-0.00380***	-0.00378***	-0.00349
	(0.000770)	(0.000743)	(0.00284)	(0.000770)	(0.000743)	(0.00284)
Flood Factor=7	-0.00705***	-0.00655***	0.000201	-0.00689***	-0.00645***	-0.000586
	(0.000989)	(0.000955)	(0.00365)	(0.000989)	(0.000955)	(0.00365)
Flood Factor=8	-0.00435***	-0.00444***	-0.00568	-0.00412***	-0.00430***	-0.00672
	(0.00104)	(0.00100)	(0.00383)	(0.00104)	(0.00100)	(0.00383)
Flood Factor=9	-0.00222	-0.00270*	-0.00928*	-0.00164	-0.00228*	-0.0110*
	(0.00118)	(0.00114)	(0.00435)	(0.00118)	(0.00114)	(0.00435)
Flood Factor=10	0.00682**	0.00316	-0.0468***	0.00809***	0.00409*	-0.0504***
	(0.00208)	(0.00201)	(0.00769)	(0.00209)	(0.00201)	(0.00769)
Log Applicant Income (in $000)		0.158***	2.312***		0.158***	2.312***
		(0.000215)	(0.00314)		(0.000215)	(0.00314)
Block Fixed Effects	✓	✓	✓	✓	✓	✓
Housing Characteristics	✓	✓	✓	✓	✓	✓
Additional Control	Miles to Coast	Miles to Coast	Miles to Coast	Elevation	Elevation	Elevation
N	8592563	8592563	8592563	8592563	8592563	8592563
Adjusted R^2	0.810	0.823	-12.87	0.810	0.823	-12.86

Notes: Standard errors in parentheses. Statistical significance at the 0.1%, 1%, and 5% levels are denoted by ***, **, and *, respectively.

Table 5: Robustness: Estimation of the Price of Flood Factor - Additional Controls

	Log Sales Amount (1) Anti-IV
Flood Factor=1	0 (.)
Flood Factor=2	-0.000479 (0.00454)
Flood Factor=3	0.00295 (0.00320)
Flood Factor=4	-0.00545 (0.00374)
Flood Factor=5	0.00472 (0.00351)
Flood Factor=6	-0.00255 (0.00283)
Flood Factor=7	0.000162 (0.00364)
Flood Factor=8	-0.00561 (0.00382)
Flood Factor=9	-0.00926* (0.00434)
Flood Factor=10	-0.0479*** (0.00766)
Log Applicant Income (in $000)	2.308*** (0.00312)
Fixed Effects	Y
Hedonic Controls	Y
Minority Control	Y
N	8592563
Adjusted R^2	-12.77

Notes: Standard errors in parentheses. Statistical significance at the 0.1%, 1%, and 5% levels are denoted by ***, **, and *, respectively.

A Proofs

This appendix provides the proofs for why an anti-instrumental variable that satisfies conditional independence can identify amenity prices. For the general (non-parametric) case, developed based on Section 3 of Bell (2022), we make a clear statement on how the price is formulated as the ratio of the two partial derivatives. The proof also clarifies the nature of local monotonicity required. For the special linear case, unlike Bell (2022), we provide an intuitive formulation that is based on the key properties of the linear projection residuals.

A.1 Proof of Identification: General Case

Theorem A.1. ***(The price of the amenity is identified using an anti-instrumental variable.)*** *Consider a market that satisfies the single index assumption as described in Assumption 2.1. If there exists an anti-IV H that satisfies conditional independence (Assumption 2.2) and strict monotonicity locally at (p, z) (Assumption 2.3), then the price of the amenity is identified and can be consistently estimated as*

$$\left.\frac{\partial P(Z,\Phi)}{\partial Z}\right|_{(z,\phi)} = -\frac{\partial \hat{H}(P,Z)}{\partial Z} \Big/ \left.\frac{\partial \hat{H}(P,Z)}{\partial P}\right|_{(p,z)} \tag{A.1}$$

where $\phi = \Phi(p, z)$, and $\hat{H}(\cdot, \cdot)$ refers to the predicted anti-IV

$$\hat{H}(p, z) \equiv \mathbb{E}[H \mid P = p, Z = z]. \tag{A.2}$$

Proof. Denote $\phi = \Phi(p, z)$. First, consider

$$\begin{aligned}
&\hat{H}(P = p, Z = z) && && \text{(A.3)}\\
&= \mathbb{E}[H|P = p, Z = z] && \text{by definition} && \text{(A.4)}\\
&= \mathbb{E}[H|P = p, Z = z, \Phi = \phi] && \text{by Single Index Assumption (invertability in } \Phi = \Phi(p, z)) && \text{(A.5)}\\
&= \mathbb{E}[H|Z = z, \Phi = \phi] && \text{by Single Index Assumption (existence of } P(z, \phi)) && \text{(A.6)}\\
&= \mathbb{E}[H|\Phi = \phi]. && \text{by Conditional Independence Assumption} && \text{(A.7)}
\end{aligned}$$

The statement above proves Lemma 2.4. Next, take the total derivative of the LHS of Equation (A.3):

$$\frac{d}{dZ}\Big(\hat{H}(P = p, Z = z)\Big) = \left.\frac{\partial \hat{H}(P,Z)}{\partial P}\right|_{(p,z)} \left.\frac{\partial P(Z,\Phi)}{\partial Z}\right|_{(z,\phi)} + \left.\frac{\partial \hat{H}(P,Z)}{\partial Z}\right|_{(p,z)} \tag{A.8}$$

Correspondingly, take the total derivative of the RHS of Equation (A.7):

$$\frac{d}{dZ}\Big(\mathbb{E}[H|\Phi = \phi]\Big) = 0 \tag{A.9}$$

Taken together, we have

$$\left.\frac{\partial P(Z,\Phi)}{\partial Z}\right|_{(z,\phi)} = -\frac{\partial \hat{H}(P,Z)}{\partial Z} \Big/ \left.\frac{\partial \hat{H}(P,Z)}{\partial P}\right|_{(p,z)}. \tag{A.10}$$

Note that the denominator

$$\left.\frac{\partial \hat{H}(P,Z)}{\partial P}\right|_{(p,z)} = \left.\frac{\partial \hat{H}(\Phi)}{\partial \Phi}\right|_{\phi} \times \left.\frac{\partial \Phi(P,Z)}{\partial P}\right|_{(p,z)} \neq 0 \tag{A.11}$$

because local monotonicity ensures $\partial \hat{H}(\phi)/\partial \Phi \neq 0$ and the Single Index Assumption ensures that $\partial \Phi(p,z)/\partial P \neq 0$. □

A.2 Proof of Identification: Linear Case

Theorem A.2. *Consider a market that satisfies (i) the linear single index assumption as follows*

$$P = \beta_0 + \beta_z Z + \Phi. \tag{A.12}$$

If there exists an anti-instrumental variable H that satisfies

$$\textit{(ii) Linear conditional independence:} \quad cov(H, Z^{\perp \Phi}) = 0 \tag{A.13}$$

$$\textit{(iii) Relevance:} \quad cov(H, \Phi) \neq 0 \tag{A.14}$$

where $Z^{\perp \Phi}$ denotes the residual of the linear projection from Z to Φ, then the price of the amenity is identified and can be consistently estimated as the ratio of two regression coefficients

$$\beta_z = -\frac{\delta_z}{\delta_p} \tag{A.15}$$

where δ_z and δ_p are OLS regression coefficients of H on Z and P respectively:

$$\mathbb{E}^*[H|Z,P] = \delta_0 + \delta_z Z + \delta_p P. \tag{A.16}$$

Proof. First, consider the following linear projection of H on Z and P

$$H = \delta_0 + \delta_z Z + \delta_p P + \mu \tag{A.17}$$

where we have, by construction,

$$cov(Z, \mu) = 0 \tag{A.18}$$

$$cov(P, \mu) = 0. \tag{A.19}$$

Next, replace the price with the underlying hedonic equation $P(Z, \Phi)$

$$H = \delta_0 + \delta_z Z + \delta_p P + \mu \tag{A.20}$$
$$= \delta_0 + \delta_z Z + \delta_p(\beta_0 + \beta_z Z + \Phi) + \mu \tag{A.21}$$
$$= (\delta_0 + \delta_p \beta_0) + (\delta_z + \delta_p \beta_z) Z + \delta_p \Phi + \mu. \tag{A.22}$$

Now, consider the following hypothetical regression, which is infeasible because Φ is unobserved,

$$H = \gamma_0 + \gamma_z Z + \gamma_\phi \Phi + \omega \tag{A.23}$$

where we have, again by construction,

$$cov(Z, \omega) = 0 \tag{A.24}$$
$$cov(\Phi, \omega) = 0. \tag{A.25}$$

Moreover, because H satisfies linear conditional independence, namely,

$$cov(H, Z^{\perp \Phi}) = 0, \tag{A.26}$$

it implies that the coefficient on the hypothetical regression must satisfy

$$\gamma_z = 0. \tag{A.27}$$

Moreover, since $cov(\Phi, \omega)$ the hypothetical linear regression of H on Φ is

$$H = \gamma_0 + \gamma_\phi \Phi + \omega \tag{A.28}$$

where $\gamma_\phi = \frac{cov(H,\Phi)}{var(\Phi)}$.

Finally, we examine the following correlation

$$cov(\Phi, \mu) = cov(P - \beta_0 - \beta_z Z, \mu) \tag{A.29}$$
$$= cov(P, \mu) - \beta_z cov(Z, \mu) \tag{A.30}$$
$$= 0 \tag{A.31}$$

where the first equation requires the invertability of the pricing function in Φ, which is guaranteed by the linear hedonic pricing function in this case. The last step is true by construction due to Eq (A.18) and (A.19). Thus, because the error term μ is uncorrelated with both Z and Φ, Eq (A.22) also represents a linear projection of H on Z and Φ. Because there exists only one linear projection of H on Z and Φ, it must be that Eq (A.22) and Eq (A.23) represent the same regression. Therefore, we

have

$$\delta_z + \delta_p \beta_z = \gamma_z = 0 \tag{A.32}$$

Or, equivalently,

$$\beta_z = -\frac{\delta_z}{\delta_p} \tag{A.33}$$

where the last step uses the relevance assumption because

$$\delta_p = \gamma_\Phi = \frac{cov(H, \Phi)}{var(\Phi)} \neq 0. \tag{A.34}$$

□

B An Illustrative DGP Simulation with Cobb-Douglass Demand

In this section, we present a simple model of the housing market under Cobb-Douglas utility and competitive supply of amenities. Moreover, we show that there could be a natural one-to-one correspondence ("duality") between the household budget constraint and the housing quality index, which motivates us to use measures of the household budget constraints (e.g., household income) as the anti-IV variable for amenity pricing.

B.1 Model Set-Up

- **Demand:** Households have Cobb-Douglas utility over housing amenity Z, housing quality ϕ, and the consumption of a numeraire good, C

$$U_i(Z_i, \Phi_i, C_i) = Z_i^{\theta_i^Z} \Phi_i^{\theta^\phi} C_i^{\theta_i^C} \tag{B.1}$$

  where $\theta_i^Z + \theta^\phi + \theta_i^C = 1$. The preference for amenity is heterogeneous, measured by θ_i^Z. The preferences for quality is θ^ϕ. Moreover, we assume that the preference for quality is vertical, namely θ^ϕ is the same across all households. The preference for the outside consumption good θ_i^C.

- **Supply:** There is one price-taking representative firm on the supply side with a cost function for producing amenity and quality: $TC(Z, \Phi)$. For simplicity, we assume linear cost function, then we have

$$TC(Z, \Phi) = c_z Z + c_\Phi \Phi \tag{B.2}$$

- **Market Clearing:** Let $P(Z, \Phi)$ denote the price of a home with amenity Z and quality Φ. Then, in a competitive equilibrium, marginal cost equals to marginal revenue. So, the cost function determines the price of amenity and quality respectively:

$$\frac{\partial P}{\partial Z} = \frac{\partial TC}{\partial Z} = c_z \tag{B.3}$$

$$\frac{\partial P}{\partial \Phi} = \frac{\partial TC}{\partial \Phi} = c_\phi \tag{B.4}$$

  Therefore, the price of a home is

$$P(Z, \Phi) = \beta_z Z + \beta_\phi \Phi \tag{B.5}$$

  where $\beta_z = c_z$ and $\beta_\phi = c_\phi$. Without loss of generality, we assume the intercept to be zero (i.e., zero profit for the producers).

B.2 Correspondence Between Housing Quality and Household Budget

Given the model set-up, each household solves its utility maximization problem subject to market prices of the homes and its own budget constraint. We denote their budget constraint by η_i:

$$\max_{Z,\phi,C} U_i(Z, \Phi, C) \qquad \text{s.t.} \quad \beta_z Z + \beta_\phi \Phi + C \leq \eta_i. \tag{B.6}$$

Solving the utility maximization yields that the expenditure on amenity is a θ_i^Z fraction of the household's budget and the expenditure on quality is a θ^ϕ fraction of the household's budget:

$$Z_i^* = \frac{\theta_i^Z}{\beta_z} \eta_i \tag{B.7}$$

$$\Phi_i^* = \frac{\theta^\phi}{\beta_\phi} \eta_i \tag{B.8}$$

Since the preference for quality Φ is vertical (i.e., not heterogeneous by household), there is a one-to-one correspondence between the amount of housing quality chosen Φ_i^* and the household's budget constraint η_i. As a result, instead of trying to control for unobserved housing quality, it would be equivalent to control for the household budget set η_i, for which we can find a anti-IV.

In other words, while there may be various contributors to one's budget constraints that are unobservable to the econometrician such as one's wealth, inheritance, debt obligations etc., the anti-IV approach allows us to still estimate the amenity price as long as we can observe a anti-IV variable for the budget set. For instance, one may consider household income to be relevant for a household's budget; yet, conditional on the total budget, the variations in household income may be thought of plausibly orthogonal to the preference for amenity θ_i^Z, and thus orthogonal to the amount of amenity chosen Z_i^*.

C "Missing" Amenities

In Section 2.5.3, we noted that when there are missing amenities, the amenity price estimated by our anti-IV method may differ from the price one would estimate from a quasi-experiment.

Our anti-IV method estimates the price a home buyer would pay to obtain the additional amenity, factoring in the existing correlation between the observed amenity and unobserved amenities as well as unobserved quality in the market. This may differ from the impact on home prices due to an exogenous and isolated change in the same amenity, as found in quasi-experiments. Although the latter is useful when comparing the costs of producing such amenities, we believe the former is more pertinent for assessing housing inequality, specifically, the actual amount home-buyers would need to pay to attain a certain amenity level.

To compare our estimate with the amenity price derived from quasi-random variations, we introduce an analogous "bias" formula for our anti-IV methods. Unlike the OLS methods, we demonstrate that the "bias" of the anti-IV estimator remains stable even in the presence of unobserved quality. This suggests that our anti-IV method can offer a more reliable price estimate compared to traditional OLS methods, even when there are "missing" amenities.

C.1 Set-Up

Consider a market where prices are determined by two amenities Z_1, Z_2 and quality Φ. We focus on the linear case where the pricing equation is

$$P = \beta_0 + \beta_1 Z_1 + \beta_2 Z_2 + \beta_\phi \Phi \tag{C.1}$$

where $\beta_1, \beta_2, \beta_\phi \neq 0$, so the Single Index Assumption is satisfied. Further, consider that a anti-IV variable H exists and satisfies linear conditional independence:

$$cov(H, Z_1^{\perp \Phi}) = 0, \qquad cov(H, Z_2^{\perp \Phi}) = 0. \tag{C.2}$$

Then, the amenity price could be consistently estimated by the anti-IV method using the ratio of coefficients formula derived in Theorem 2.7 if all amenities are observed:

$$-\frac{\delta_1}{\delta_p} = \beta_1 \tag{C.3}$$

where the coefficients are based on the linear regression

$$\mathbb{E}^*[H|Z_1, Z_2, P] = \delta_0 + \delta_1 Z_1 + \delta_2 Z_2 + \delta_p P. \tag{C.4}$$

C.2 Formula for Missing Amenity Bias - OLS

To start, we first derive the bias using OLS while controlling for H

$$\mathbb{E}^*[P|Z_1, H] = \beta_0^{\text{ols}} + \beta_1^{\text{ols}} Z_1 + \beta_H^{\text{ols}} H.$$

We derive the OLS bias formula when quality is imperfectly controlled for

$$\hat{\beta_1}^{\text{ols}} = \frac{cov(P, Z_1^{\perp H})}{var(Z_1^{\perp H})} \tag{C.5}$$

$$= \frac{cov(\beta_0 + \beta_1 Z_1 + \beta_2 Z_2 + \beta_\phi \Phi, Z_1^{\perp H})}{var(Z_1^{\perp H})} \tag{C.6}$$

$$= \beta_1 + \beta_2 \frac{cov(Z_2, Z_1^{\perp H})}{var(Z_1^{\perp H})} + \beta_\phi \frac{cov(\Phi, Z_1^{\perp H})}{var(Z_1^{\perp H})} \tag{C.7}$$

where the first bias component $cov(Z_2, Z_1^{\perp H})$ is driven by the correlation between observed and unobserved amenity and the second bias component $cov(\Phi, Z_1^{\perp H})$ is caused by the fact that H is an imperfect control of Φ.

Similarly, we can sign the direction of the bias based on the signs of the correlation terms. To provide some intuition, suppose that we have positive priced amenities $\beta_1 > 0$ and $\beta_2 > 0$ that are positively correlated $cov(Z_2, Z_1^{\perp \Phi}) > 0$. When H perfectly measures Φ, the OLS estimator is biased upward $\hat{\beta_1^{ols}} > \beta_1$. Moreover, if H does not measure Φ well, for example, in the extreme case when H is not correlated with Φ at all, then the additional bias term $cov(\Phi, Z_1^{\perp H})$ will likely to be positive.

As an illustration, suppose Z_1 represents "clean air" and Z_2 represents "clear roads", and to the extent that they are positively correlated when controlling for housing quality Φ, then the OLS estimator provides a positively biased estimator. To interpret this "bias", it essentially means that a home-buyer cannot live in a neighborhood with clean air without also having to pay for the fact that these are less congested neighborhoods, and that is because the presence of these amenities in the market is positively correlated even when conditioning on quality.

C.3 Formula for Missing Amenity Bias - Anti-IV

Next, we derive the "bias" formula for the anti-IV estimator when there are missing amenities. Suppose Z_2 is missing, then, the anti-IV estimator becomes

$$\hat{\beta}_1^{\text{anti-IV}} = -\frac{\gamma_1}{\gamma_p} \tag{C.8}$$

where the coefficients are based on the linear regression of H on P and Z_1

$$\mathbb{E}^*[H|Z_1, P] = \gamma_0 + \gamma_1 Z_1 + \gamma_p P. \tag{C.9}$$

Recall the residual term of a linear regression is uncorrelated with the regressor

$$\mathbb{E}^*[H|Z_1, Z_2, P] = \delta_0 + \delta_1 Z_1 + \delta_2 Z_2 + \delta_p P \tag{C.10}$$

$$cov(\epsilon_{Z_1, Z_2, P}, Z_1) = 0 \tag{C.11}$$

$$cov(\epsilon_{Z_1, Z_2, P}, P) = 0. \tag{C.12}$$

We can rewrite the numerator of the anti-IV estimator as follows:

$$\gamma_1 = \frac{cov(H, Z_1^{\perp P})}{var(Z_1^{\perp P})} \tag{C.13}$$

$$= \frac{cov(\delta_0 + \delta_1 Z_1 + \delta_2 Z_2 + \delta_p P + \epsilon_{Z_1,Z_2,P},\, Z_1^{\perp P})}{var(Z_1^{\perp P})} \tag{C.14}$$

$$= \delta_1 + \delta_2 \frac{cov(Z_2, Z_1^{\perp P})}{var(Z_1^{\perp P})} \tag{C.15}$$

Similarly, we can rewrite the denominator of the anti-IV estimator as follows:

$$\gamma_p = \frac{cov(H, P^{\perp Z_1})}{var(P^{\perp Z_1})} \tag{C.16}$$

$$= \frac{cov(\delta_0 + \delta_1 Z_1 + \delta_2 Z_2 + \delta_p P + \epsilon_{Z_1,Z_2,P},\, P^{\perp Z_1})}{var(P^{\perp Z_1})} \tag{C.17}$$

$$= \delta_p + \delta_2 \frac{cov(Z_2, P^{\perp Z_1})}{var(P^{\perp Z_1})}. \tag{C.18}$$

Note that the covariance term $cov(\epsilon_{Z_1,Z_2,P}, Z_1^{\perp P}) = 0$ and $cov(\epsilon_{Z_1,Z_2,P}, P^{\perp Z_1}) = 0$, because the residual term $\epsilon_{Z_1,Z_2,P}$ is by construction uncorrelated with P and Z_1. Putting the two together, the anti-IV estimator can be written as

$$\hat{\beta}_1^{\text{anti-IV}} = -\frac{\gamma_1}{\gamma_p} = -\frac{\delta_1 + \delta_2 \times \psi_1}{\delta_p + \delta_2 \times \psi_p} \tag{C.19}$$

where ψ_1 and ψ_p is defined by the linear regression of Z_2 on P and Z_1:

$$\mathbb{E}^*[Z_2|P, Z_1] = \psi_0 + \psi_1 Z_1 + \psi_p P. \tag{C.20}$$

Given that $-\frac{\delta_1}{\delta_p} = \beta_1$, we have that

$$\text{if} -\frac{\psi_1}{\psi_p} = \beta_1, \quad \text{then} \quad \hat{\beta}_1^{\text{anti-IV}} = \beta_1. \tag{C.21}$$

The derivation of the anti-IV bias in Equation (C.19) also clarifies that the bias due to missing amenities in the anti-IV estimator is *not* sensitive to how well the quality Φ itself is observed. In particular, the anti-IV "bias" is determined by the regression coefficients of Z_2 on P and Z_1,[27] which does not contain H. The "bias" is also affected by δ_2, which comes from the regression of H on P, Z_1, and Z_2 and is not sensitive to the measurement errors in the outcome variable.[28] Therefore, even

[27] One may also note that this ratio $-\psi_1/\psi_p$ can be viewed as the anti-IV estimator for the price of amenity when we use Z_2 as the anti-IV variable. Of course, this model is likely misspecified unless the corresponding conditional independence assumption happens to hold true. Namely, $cov(Z_2, Z_1^{\perp \tilde{\Phi}}) = 0$ where $\tilde{\Phi} = \beta_2/\beta_\phi Z_2 + \Phi$ measures an adjusted quality. Equivalently, the anti-IV estimator coincides the quasi-random estimate when $cov(Z_1, Z_2) = cov(Z_1, \tilde{\Phi})cov(Z_2, \tilde{\Phi})/var(\tilde{\Phi})$.

[28] Note that more measurement error in H does not change the estimate of δ_2 because H is on the LHS as the outcome variable in the regression, which is a key difference from the OLS case where H is included in the RHS as a regression

though neither the anti-IV method nor the OLS method is inherently designed to address the issue of missing amenities, the anti-IV estimate remains robust to unobserved qualities even in the presence of missing amenities.

Moreover, the bias term $-\frac{\phi_1}{\psi_p}$ for the anti-IV estimator encapsulates both the correlation between the unobserved amenity Z_2 and the observed amenity Z_1, but also the correlation between the unobserved amenity Z_2 with the unobserved quality, indirectly through its impact on price P.

As an illustration, suppose again Z_1 represents "clean air" and Z_2 represents "clear roads". They are positively correlated when controlling for housing quality Φ, but they are also positively correlated with housing quality, as richer households tend to consume more of both amenities and housing quality. Then, to interpret the anti-IV "bias", it means that to access cleaner air, a home buyer can achieve this through a combination of purchasing more "clear roads" and shifting to a segment of higher housing quality.

C.4 Simulation

In this section, we illustrate the property of both OLS and the proxy method when there is an unobserved amenity by running simulations using a DGP based on Cobb-Douglas preferences and competitive supply as outlined previously in the paper.

C.4.1 DGP for Simulation

The DGP is based on Cobb-Douglas demand over two amenities Z_1 and Z_2.

$$U_i(Z_1, Z_2, \Phi, C) = Z_1^{\theta_i^{Z_1}} Z_2^{\theta_i^{Z_2}} \Phi^{\theta^\phi} C^{\theta_i^C} \tag{C.22}$$

where $\theta_i^{Z_1} + \theta_i^{Z_2} + \theta^\phi + \theta_i^C = 1$. The true price for the amenities is $(\beta_1, \beta_2) = (0.3, 0.3)$. The first amenity Z_1 is observed by the researcher, but Z_2 is not.

Households budget sets η_i are simulated as follows: The set of all possible values of household budgets η_i is $\mathcal{E} = \{10n | n = 1, 2, ..., 100\}$. For each value in $\mathcal{E}$, there are exactly 50,000 households with this amount of budget, so there are a total of 5,000,000 households in the simulation. We simulate under both the case when η_i is perfectly observed, and when η_i is observed with a random noise as $h_i = 0.7\eta_i + \epsilon_i$, where $\epsilon_i \sim \mathcal{N}(0, 100)$ is the random noise.

The preference parameter for quality, θ^ϕ, is 0.5. The preference parameter for the amenities, $\theta_i^{Z_1}$ and $\theta_i^{Z_2}$, are drawn from a joint distribution, independently of η_i. Each has a uniform marginal distribution on $[0, 0.25)$, and the correlation between them is $corr(\theta_i^{Z_1}, \theta_i^{Z_2}) = \rho$. The correlation ρ is a key determinant of the direction of missing amenity bias, since it captures the co-movement of the demand for the two amenities as the expenditure on each amenity is driven by one's preferences. Therefore, we examine the results under various correlation ρ in our simulations. Finally, $\theta_i^C = 1 - \theta_i^{Z_1} - \theta_i^{Z_2} - \theta^\phi$.

C.4.2 Simulation Results

In the first set of simulations, we simulated the data by varying the correlation between the preference for the observed and unobserved amenity to examine how the anti-IV and OLS results respond to the change in ρ.

The results from the simulations are presented in Figure C.1. The x-axis represents the correlation between preference for amenities in each simulation, while the colored series depicts the estimated price for amenity one using different methods.

First, let's consider a scenario where all household budget sets, and therefore the amount of housing quality chosen, is perfectly observed. In this case, the solid blue and green lines indicate the estimated price for amenity using OLS and anti-IV, respectively. Notice that in the absence of correlation between preferences, the OLS method (solid blue) restores the original price β_1. The anti-IV method (solid green) also restores the correct price but at a different point when there is some positive correlation in preference such that $-\psi_1/\psi_p = \beta_1$. Moreover, these two lines intersect at both ends, where the preferences for the two amenities are perfectly correlated. This is to be expected in the case of $\rho = \pm 1$, because the price of Z_2 cannot be separately identified from Z_1, and both methods are reduced to estimating the price of the single combined amenity.

Next, when the household budget sets are observed with noise, the amount of housing quality chosen is measured with noise. In this case, the OLS estimator yields a highly biased result (dotted blue line), while the anti-IV estimator remains unaffected and robust to components of unobserved quality (dotted green line).

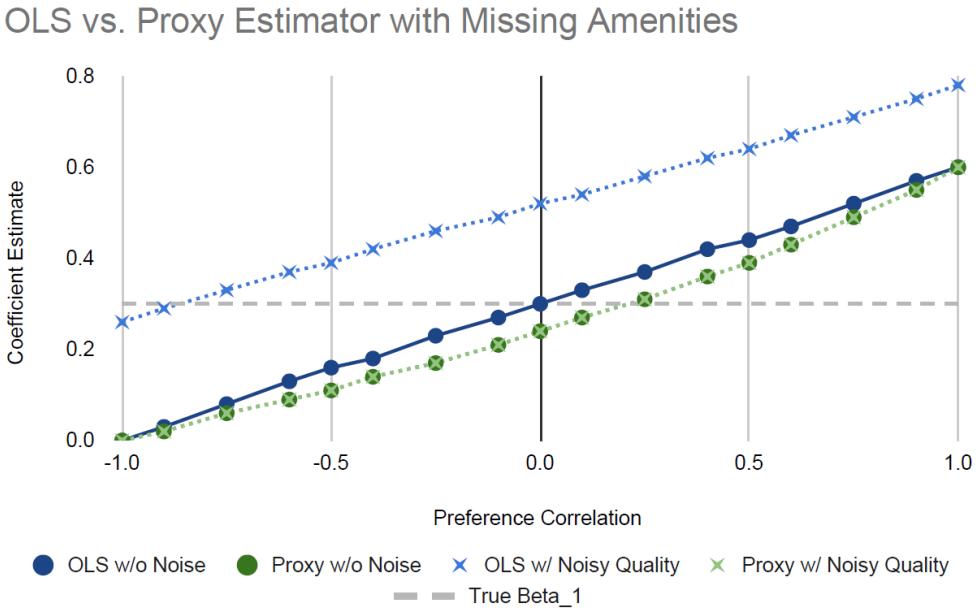
Furthermore, when we explore how sensitive our estimate is to the relationship between the anti-IV h and the budget set η, we simulate the data by holding the total variance of η fixed while varying the signal to total ratio of h.[29] If the variance of h is significantly higher than the variance of ϵ, then most of the variation in η is captured in h, and we will have a high signal-to-total ratio, and vice versa.

Figure C.2 shows that as the signal-to-total ratio decreases (going from the right of the graph to the left), the bias from OLS with anti-IV control increases steadily, while the bias for the anti-IV method is unchanged across different levels of signal to total ratio.[30] This simulation result is consistent with our derivation where the missing amenity "bias" of the anti-IV estimator is robust to varying levels of signal-to-total ratio of the anti-IV variable, which is a property that OLS does not have.

[29] To simulate this, households have a random draw of both anti-IV and noise, with $h_i \sim \mathcal{N}(0, 100)$ and $\epsilon_i \sim \mathcal{N}(0, 100)$, and the budget set is given as $\eta_i = \lambda h_i + (1 - \lambda)\epsilon_i$, where λ denotes varying levels of how informative h is.

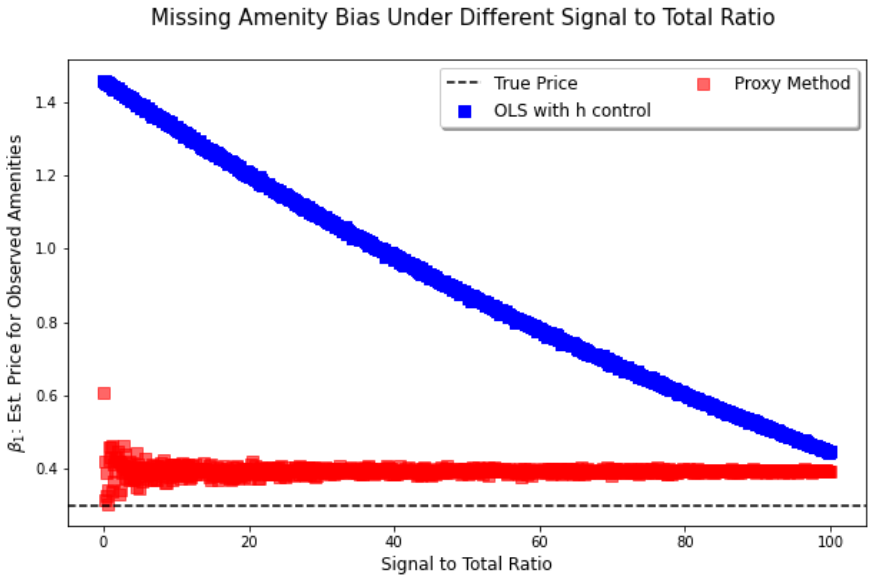
[30] When the signal-to-total ratio is very close to zero, the anti-IV method suffers from a "weak" anti-IV problem, analogous to the weak instrument problem of an IV estimator, where the denominator in the ratio-of-coefficients becomes close to zero and leading to larger standard errors.

Figure C.1: Missing Amenities Bias v.s. Preference Correlation



Notes: This figure compares the OLS and anti-IV estimator with a missing amenity under varying levels of preference correlations for the two underlying amenities. OLS estimates are shown in blue lines, and the anti-IV estimates are shown in green lines. Both methods can suffer from the missing amenity bias. However, in the presence of imperfectly measured quality, the bias in the OLS estimate (dotted blue line) is substantial, whereas the anti-IV estimator is not affected (dotted green line).

Figure C.2: Missing Amenity Bias v.s Signal to Total Ratio



Notes: This figure shows that as the signal-to-total ratio decreases (going from the right of the graph to the left), the bias from OLS with anti-IV control increases steadily, while the bias for the anti-IV method is unchanged across different levels of signal-to-total ratio.

D An Anti-IV Approach to Valuing Air Quality

D.1 Data

D.1.1 Geographic Mobility

Our measure of population mobility across geographic regions comes from two sources. The first data source is the 2018-2019 county-to-county migration data from the Internal Revenue Service (IRS) Statistics of Income (SOI) Division, which is "based on year-to-year address changes reported on individual income tax returns filed with the IRS."[31] The data is publicly available and covers the entire universe of population with Forms 1040 filings. The main caveats of the data are that it does not cover migration at finer geographic level or by demographics, and it does not include non-filers. Furthermore, flows between counties with fewer than 20 returns are suppressed for confidentiality protection. To measure geographic mobility at a finer resolution, we also use the Infutor data. The Infutor data contain individual-level mobility data across the United States that records the past ten residences at the street address level for each individual. Coverage is thought to be broad, though variable over time. For this version of the paper, we focus on a subsample of the data that contains all the individuals who have ever lived in the state of California to measure migration flows between California ZIP Code Tabulation Areas (ZCTA) in the year 2019.

D.1.2 Home Prices

Our house price data comes from the Zillow Home Value Index (ZHVI), which reflects the typical value for homes in the 35th to 65th percentile range. We collect monthly ZHVI for single-family residences with 1,2,3,4, and 5+ bedroom at the county-level in 2019 and take the monthly average over the year.

D.1.3 Geographic Amenities

The main amenity data we use for our analysis is the Air Quality Index (AQI) from EPA's Air Quality System, which is an aggregate measure of five pollutant levels. We collect daily county-level AQI data in 2019 from EPA Air Data and use the median across all days in the year. Since the AQI data is only available for 1063 counties out of the total 3033 counties, we interpolate the AQI using the inverse distance weighting method commonly used in the economics literature on air pollution (Neidell 2004; Currie and Neidell 2005).[32] For these counties, the interpolated AQI is calculated as the average of AQI readings from other counties weighted by the inverse of the squared distance between the counties.

D.1.4 Other Data Sources

Finally, we complement our data with county-level socio-demographics from the 2019 American Community Survery (ACS) 5-year data. This includes household median income, percent high school educated or above, and percent college educated or above.

[31] https://www.irs.gov/pub/irs-soi/1819inpublicmigdoc.pdf

[32] We are exploring other measures of air pollution, including those derived from satellite data used by Sullivan and Krupnick (2018). Results thus far are qualitatively robust to restricting the sample to non-interpolated counties.

D.2 Geographic PageRank (GPR)

In this section, we propose and construct a novel measure of locational desirability called "Geographic PageRank (GPR)," which we later use as a anti-IV variable for the purpose of valuing geographic amenities.

D.2.1 The Construction of Geographic PageRank

The key idea behind "Geographic PageRank" is to capture a ranking of geographic locations using a recursive logic based on migration: by revealed preference, when households move, they must on average move to a "better" place than their previous place, and places that draw people from "better" places are considered "best" places. As such, we use a matrix of migration to arrive at a ranking of places based on such recursive definition of common desirability. To implement the idea of a recursive-based ranking using flows, we use the famous PageRank algorithm introduced into labor economics by Sorkin (2018), which was invented by Larry Page (2001) and had been the foundational piece underlying much of Google's search rankings. Our approach to adapting PageRank from the labor to housing literature also compliments that of Fogel (2021), who was to our knowledge the first to implement PageRank at the geographic level, but his focus is to use it to show a null effect of local labor demand shocks on city value as measured by PageRank.

At a conceptual level, our proposed Geographic PageRank algorithm ranks a location by its share in the stationary distribution of household locations in accordance to the migration matrix. Specifically, let the migration matrix $M_{i,j}$ represent the fraction of population migrating from location j to i. In other words, the matrix specifies the destination locations for everyone from location j, and $M_{i,j}$ represents a transition matrix (in a column format) such that every column sums to one

$$\forall j : \sum_i M_{i,j} = 1. \tag{D.1}$$

In its simplest form, the PageRank algorithm computes the column vector v such that

$$Mv = v. \tag{D.2}$$

Equivalently, v represents the eigenvector associated with the transition matrix M.[33] In this sense, the resulting eigenvector v represents the stationary distribution of the population if they migrated according to the transition matrix M indefinitely: namely, the transition matrix M on the distribution v recovers v. As a last step, we obtain the ordinal "rank" of location i based on the rankings of the value of each element v_i.

To provide some intuition behind the PageRank algorithm, consider the following stylized example. In Panel (A) of Figure D.3, there are three locations A, B, and C. Assume that they are equal-sized cities and we observe that 10% of C's residents migrated to A and another 10% migrated to B.

[33] In practice, the PageRank algorithm also allows for a damping factor d such that every period a small fraction $1 - d$ from each node will be randomly redistributed to all other nodes to ensure numerical stability. We find that the ordinal rank of the nodes are not sensitive to the choice of the damping factor.

Performing the PageRank algorithm on the resulting migration matrix M results an equal ranking between A and B, which are both ranked above location C.

In Panel (B) of Figure D.3, we also consider three locations A, B, and C. However, in this case, we find 5% of C's residents migrated to A, and 15% migrated to B. Moreover, we also find 5% of B's residents also migrated to A. In this case, the PageRank algorithm ranks A above B, and then above C. Notably, the PageRank algorithm contains more information than simple counts of net-migration: even though both location A and B see a net-in migration of 10% in both examples, the PageRank algorithm ranks A above B in the second example, as A is drawing residents from B, which is in turn more desirable than C.

Beyond its widespread application in computer science, Sorkin (2018) is the first to use the PageRank algorithm in economics and it is applied to ranking a set of U.S. firms in the labor market context. Indeed, Sorkin (2018) proceeds to use the ranking of firms as a way to measure how much wage dispersion is due to job amenities. We highlight that there is a natural parallel between the challenges of the empirical valuation of job amenities and the empirical valuation of neighborhood amenities, even though there has not been much overlap between the two literature in recent works. As such, our work fits into this important theoretical and methodological gap.

D.2.2 A Ranking of U.S. Counties

To implement the Geographic PageRank algorithm at the national level, we use the outflow version of the IRS 2018-2019 county-to-county Migration Data, which records "the number of residents leaving a State or county and where they went." Table D.1 shows the summary statistics for the IRS migration data. 2,824 out of the total 3,033 counties has more than 20 migration flows to another county. Within these counties, the mean total outflow is 2090.6 with a standard deviation of 6802.9, and the median total outflow is 348. There are 47,974 origin-destination county pairs out of all the 9,196,056 possible pairs with more than 20 migration flows. Within these county pairs, the mean migration flow is 123.1 with a standard deviation of 468.9, and the median migration flow is 41.

Next, we obtain the corresponding rankings of all U.S. counties using the Geographic PageRank algorithm described above.[34] Figure D.4 illustrate the rankings on the map: green shades indicate higher-ranked (i.e., more desirable) counties and red shades indicate lower-ranked (i.e., less desirable) counties. We generally find higher-ranked counties along the coastal regions. Table D.2 shows the top 20 counties by its Geographic PageRank. The top 10 cities are Phoenix, Houston, Los Angeles, Dallas, Chicago, Fort Worth, Seattle, Austin, Minneapolis, and San Antonio. We observe that many of them are in the Sun Belt, reflecting recent migratory trends.

In general, places that see population growth are on average ranked higher. The top panel of Figure D.6 shows the binscatter between Geographic PageRank and net migration, clearly indicating a positive relationship. However, there remains significant dispersion for the same level of net population change, and the average relationship is also non-linear, as shown in the scatter plot in the bottom panel of Figure D.6. The remaining dispersion suggests that the recursive nature of the PageRank algorithm

[34] Our specific implementation assumes a damping factor of $d = 0.85$, which corresponds roughly to the average stay rate.

synthesizes more information from the full migration matrix M, as opposed to the simple summation operation that computes the net migration alone.

D.2.3 More Granular Rankings from Infutor Data

Although considerably noisier at present, our dataset of residential mobility from Infutor also allows us to rank finer-grained geographies. Figure D.5 shows a map of Geographic PageRank for California's ZCTA's, based on moves recorded by Infutor in 2019. In this version of the paper, we do not show additional results leveraging ZCTA-level moves due to the increased noise associated with small sample sizes at finer-grained geographies. However, we are in the process of cleaning and incorporating additional years and geographic coverage of the data, which will allow for more reliable estimates to study more granular mobility patterns.

D.2.4 Limitations

Two key limitations of the application of PageRank to geographies are worth noting. First, there are reasons to suspect that migratory tendencies might not always be equatable with place quality. If people's preferences for amenities systematically change as they age, then migratory flows may be less indicative of general place quality, and more indicative of the presence of particular amenities.[35] In preliminary results not shown here, we have found evidence that PageRank tends to over-value warm climates more than other candidate anti-IVs we have examined (such as income), which gives credence to the concern that migratory flows might not necessarily point in the direction of quality, but rather in the direction of amenities that become increasingly desirable as one ages. This limitation must be grappled with if one wishes to apply Geographic PageRank to pricing weather as an amenity, which Glaeser, Kolko and Saiz (2001) find to be among the most important amenities in driving home price differentials. For this reason, in future work, we will use the Infutor data to investigate the extent to which rankings of places are stable or change over individuals' life cycles by computing ranks based on specific age cohorts over time.

The second limitation of the PageRank algorithm when applied to geographies is that it is not invariant to aggregation of geographic units. To be more concrete, for example, New York City is divided into five counties (i.e., the five boroughs) in the data with Manhattan ranked at the 21st place. However, if we created a synthetic "New York City County" that includes all the boroughs, it would have ranked at the 5th place. One may take the problem of aggregation to its logical extreme by aggregating more and more counties to generate an artificially top-ranked "super" county. Therefore, in order to obtain a meaningful interpretation of the rankings, we need to start with locations that we believe are somewhat comparable in their geographic scope.

D.3 Results

In this section, we illustrate that the anti-IV method can be used to obtain sensible estimates of the price of air pollution. We demonstrate how it works in both the general non-linear case and the more

[35] In a related vein, as described by Sorkin (2018), correlated negative shocks (e.g., an economic recession) could lead migratory flows to tend toward worse rather than better places.

specific linear case. We show that the anti-IV-based estimates are able to address the "wrong-signed" problems that seem to be pervasive in the estimation of similar environmental amenities. We also find that the estimates are in-line with other existing reduced-form estimates using natural experiments.

First, we obtain a correctly-signed negative price for air pollution in the general setting without assuming linearity of the index mapping function, as shown in Figure D.7. Specifically, the estimation procedure goes as follows: First, we divide both price and amenities into quantiles, and in this case, we have divided them each into $N = 11$ equally sized quantiles. Then, the Cartesian product of each price quantile and each air pollution quantile generates a total of N^2 cells. For each cell, we collect all households who belong to that cell and compute the average Geographic PageRank of members of the cell, which captures $\mathbb{E}[H|P, Z]$, namely, the predicted anti-IV variable conditional on price and amenity. Hence, Figure D.7 is the direct empirical analogue of our illustration for identification as shown previously in Figure 4.

The pattern in Figure D.7 indicates a trade-off between home price and air quality. It shows a clear increase in the predicted Geographic PageRank in the diagonal direction: the best-ranked places (i.e. dark green) are in the upper-right hand corner and the worst-ranked places (i.e. dark red) are in the bottom-left corner. As such, the upward-sloping blue line in Figure D.7 indicates the direction of expanding choice sets, whereas the perpendicular direction (i.e., the dark green line) shows the level sets: the set of cells all with the same average PageRank but with varying combinations of price and amenity. In other words, the direction perpendicular to the direction of choice-set expansion shows the implied trade-off between home price and clean air: places with cleaner air are priced higher than places with more polluted air, conditioning on being in the same housing quality segment, which are proxied by the predicted Geographic PageRank.

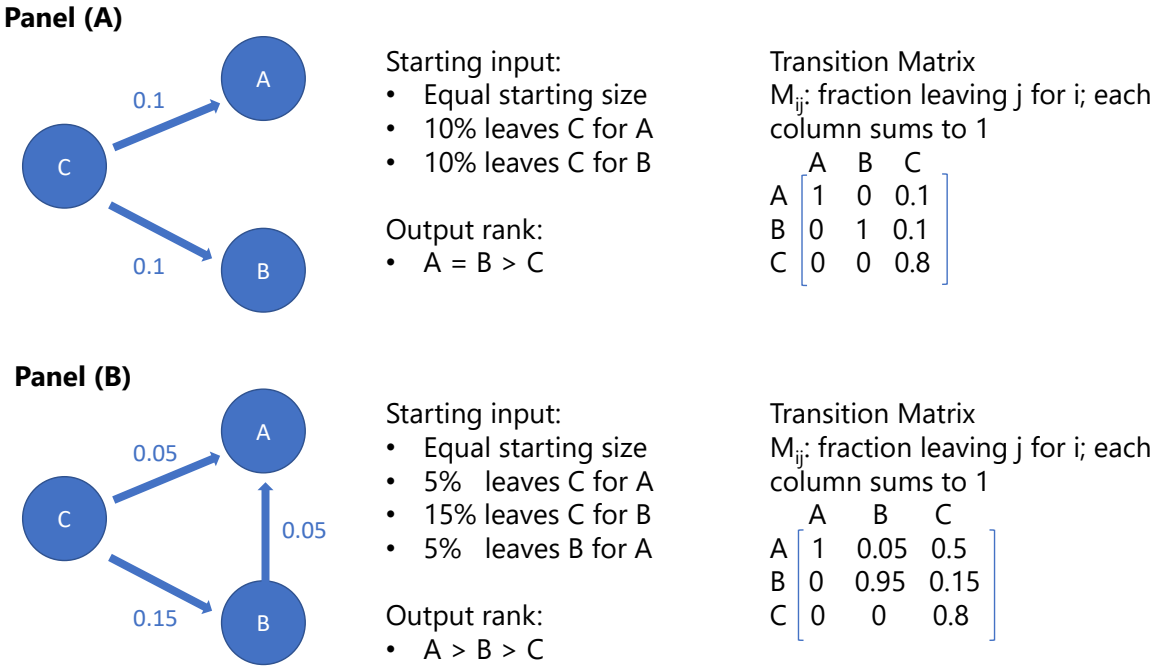
Next, if we take on the lens of linearity, then we can use the "ratio-of-coefficients" method to obtain a single coefficient estimate of the amenity price. Recall that linearity imposes the additional restriction that the index mapping function $\phi(p, z)$ must be linear in both price p and amenity z. To visually benchmark the magnitude of our estimate in a linear model, Figure D.8 compares the estimated price elasticity with known estimates in the literature, specifically, with Chay and Greenstone (2005). The shaded region indicates plausible point estimates of the housing price elasticity with respect to total suspended particulates (TSPs), which ranged from -0.20 to -0.35. The first three dots in Figure D.8 shows that the conventional OLS or OLS with controls appear to generate wrong-signed and downward-biased estimates. The last dot in Figure D.8 shows a significantly negative elasticity of -.40 estimated by our anti-IV-based approach, and the confidence interval overlaps with the range of point estimates in Chay and Greenstone (2005). Though it should not necessarily be expected ex ante that they are the same given the nuances in how price is conceptualized (discussed in Section 2.5.3), the fact that our estimate of the price of air pollution is indistinguishable from the range of estimates in the quasi-experimental literature lends some support to the credibility of our approach.

Building on the visual evidence of Figure D.8 , Table D.3 shows the estimated home price elasticity with respect to air pollution for different candidate anti-IV variables and estimation strategies. We

find that using either income or PageRank as an anti-IV—but not as a control—we get an intuitively signed price for air pollution, indicating that buyers face higher home prices to live in cleaner areas. Column (1) shows the naive OLS estimates, which reproduces the wrong-signed problem, indicating a positive price for air pollution. Column (2) shows that the conventional method of adding average household income as an additional control reduces the amount of bias, but the coefficient still remains positive. In column (3), instead of using income as a control, we estimate the price elasticity using income as an anti-IV. We find that, using household income as an anti-IV, it produces a correctly-signed negative price for air pollution. Next, instead of income, we can use the county-level Geographic PageRank either as a control or anti-IV; column (4) shows that controlling for GPR also has the effect of reducing the OLS coefficient, but still exhibits a positive sign, whereas column (5) shows our main result. The estimated price elasticity using GPR anti-IV produces a correctly-signed negative price elasticity.

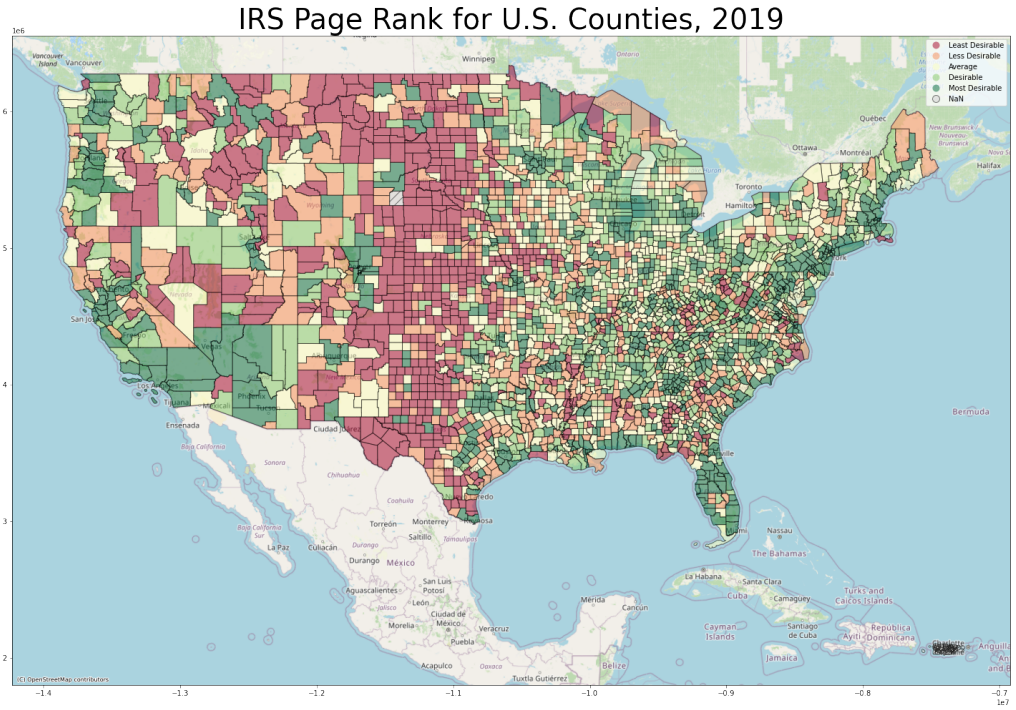
D.4 Figures: An Anti-IV Approach to Valuing Air Quality

Figure D.3: Example of PageRank Algorithm



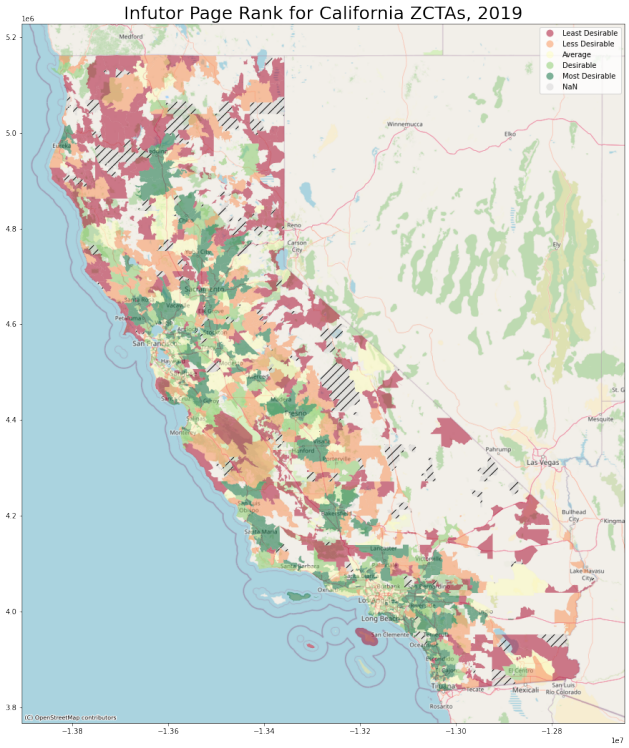
Notes: The figure above illustrates a simple example of the PageRank algorithm. In the top panel (A), there are three locations A, B, and C, where there is a net migration from location C to both A and B in equal proportions. The adjacency matrix $M_{i,j}$ indicates the transition probabilities in the column format, namely, $M_{i,j}$ represents the fraction departing from node j for node i. As such, the normalization requires $\forall i : \sum_j M_{i,j} = 1$. The resulting rank is $A = B > C$. By contrast, the bottom panel (B) illustrates another example where there is not only a migration from C, but also some migration from B to A. Note that the net in-migration for both examples for location A and B are the same, namely, a net migration of 10%. However, because location A is drawing people from the more desirable location B, as opposed to C, the resulting rank has A being ranked above B, namely $A > B > C$.

Figure D.4: County-level PageRank using IRS County Migration Data



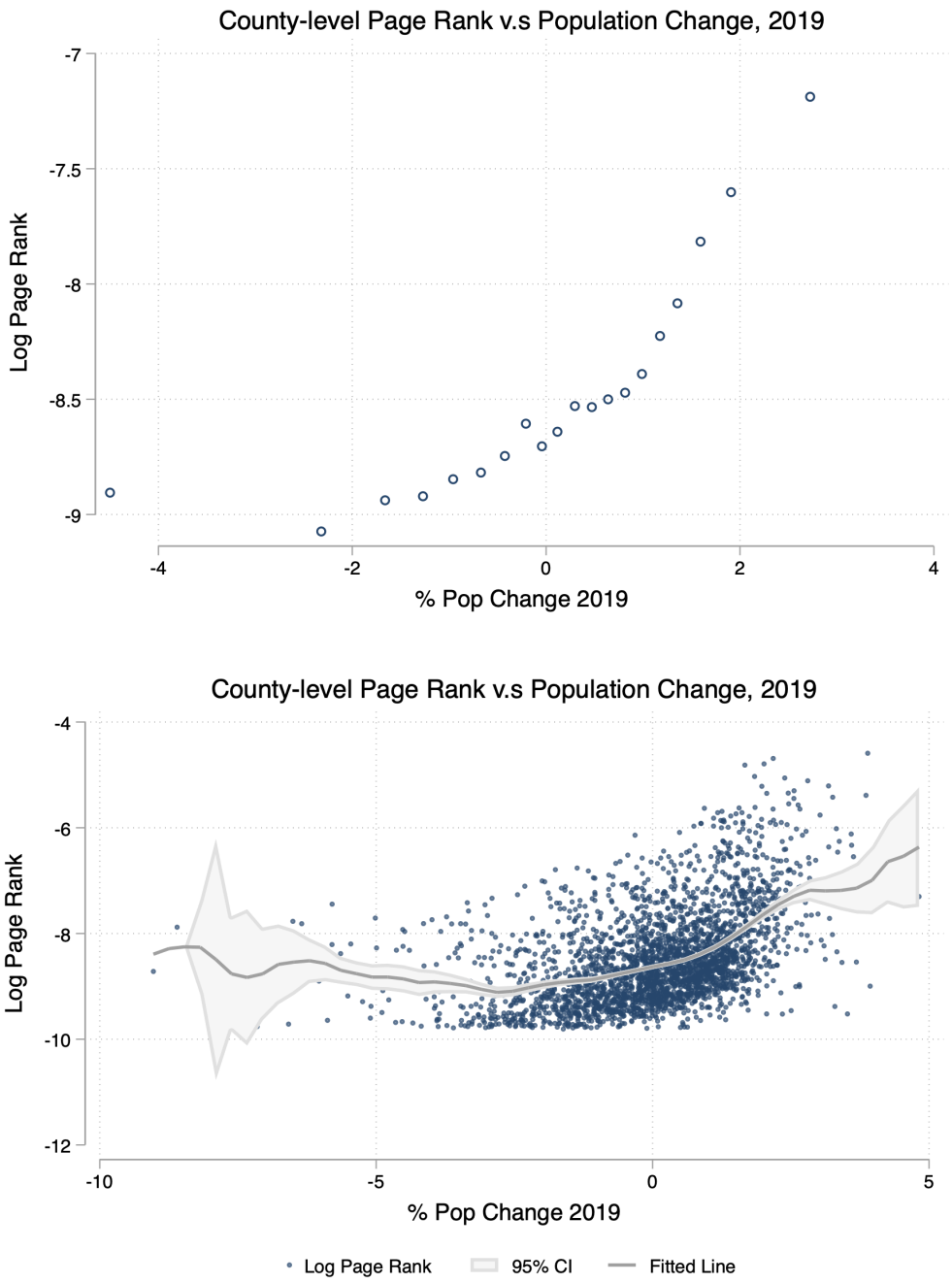
Notes: This map shows geographic PageRank at the county-level constructed using nationwide IRS county migration flow data in 2019. Green shades indicate higher-ranked (i.e., more desirable) counties and red shades indicate lower-ranked (i.e., less desirable) counties.

Figure D.5: ZCTA-level PageRank using Infutor Migration Data in California



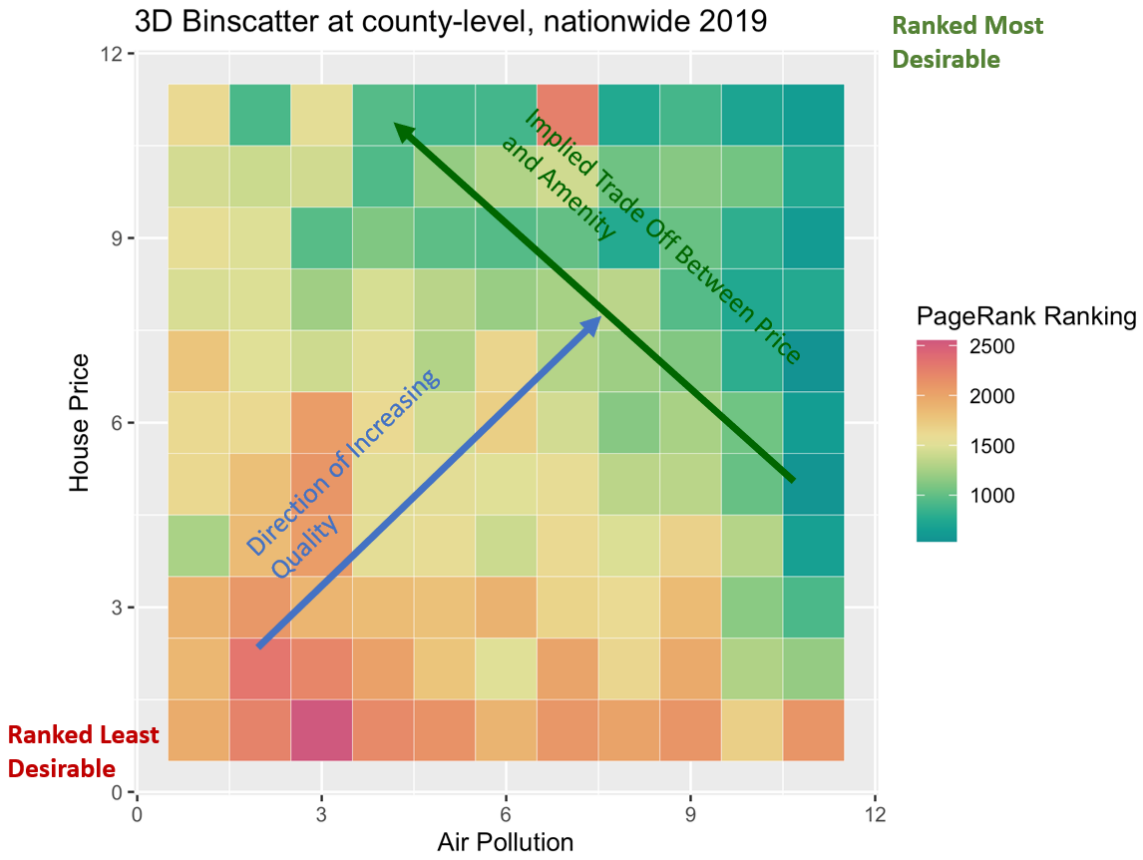
Notes: This map shows geographic PageRank at the ZCTA-level constructed using Infutor data in 2019. Green shades indicate higher-ranked (i.e., more desirable) ZCTAs and red shades indicate lower-ranked (i.e., less desirable) ZCTAs.

Figure D.6: Geographic PageRank v.s. Net Migration



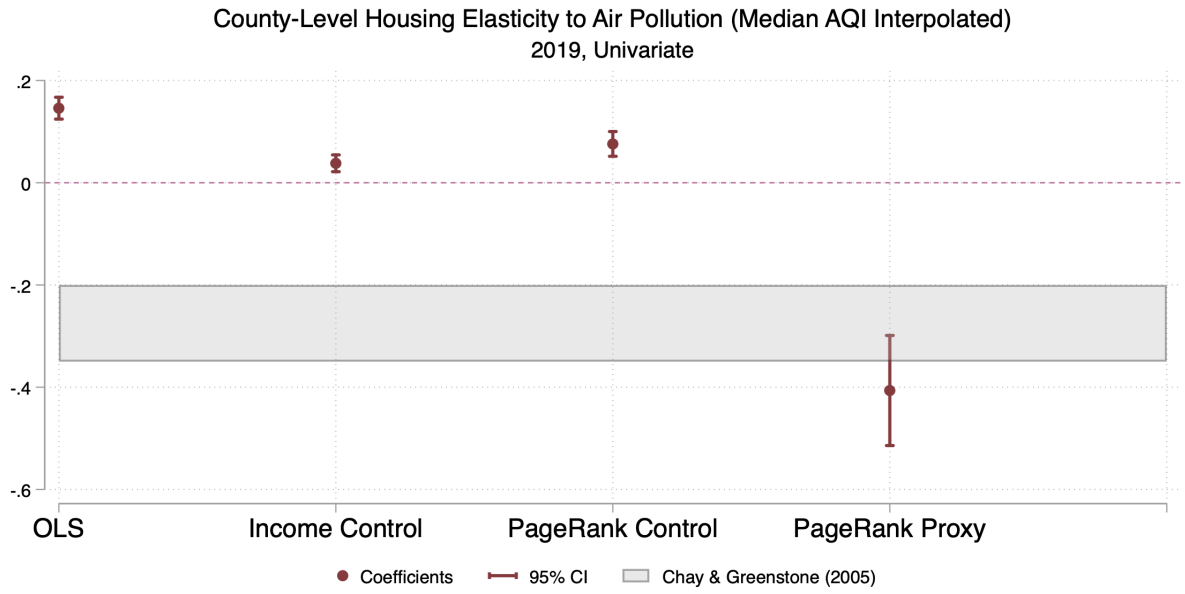
Notes: The binscatter (top panel) and the scatter plot (bottom panel) shows the relationship between Geographic PageRank and net migration. The y-axis represents the logarithm of the elements of the PageRank eigenvector. In general, places that see population growth are ranked higher. However, there remains significant dispersion for the same level of net migration and the average relationship is also non-linear.

Figure D.7: 3D Binscatter of Predicted PageRank by Price and Air Pollution



Notes: This figure categorized the counties into 121 (11x11) cells. Each cell represents counties that lie within certain house price decile x air pollution decile. For example, the upper-right-most cell represents counties in top income quantile and top air pollution quantile. The color of each cell represents the median Geographic PageRank Ranking of all the counties in this cell. The PageRank algorithm ranks the most desired place as number 1. Green shades represent higher-ranked (i.e., more desirable) places and red shades represent lower-ranked (i.e., less desirable) places.

Figure D.8: Pricing Air Pollution Using Anti-IV Approach



Notes: The estimated coefficients on this figures correspond to the estimated price elasticity to median AQI from regression results in Table D.3.

D.5 Tables: An Anti-IV Approach to Valuing Air Quality

Table D.1: Summary Statistics for 2018-2019 IRS County-to-County Migration Data

	County		Origin-Destination Pairs
	Outflows	Inflows	Migration Flows
	(1)	(2)	(3)
Mean	2090.62	2142.20	123.06
SD	6802.89	6415.80	468.86
Min	20	20	20
25th Percentile	126.5	128.5	27
Median	348	362.5	41
75th Percentile	1130.5	1211	81
Max	145204	111327	20908
N	2824	2756	47974

Notes: Migration flows between counties with less than 20 counts are suppressed in the data. Column (1) reports the county-level summary statistics for the total migration flows out of each county into other counties, and Column (2) reports the county-level summary statistics for the total migration flow into each county from other counties. Column (3) reports the summary statistics for migration flows at the level of origin county-destination pairs.

Table D.2: Top Ranked Counties in the U.S. (2019)

FIPS	Ranking	County Name	MSA Name
4013	1	Maricopa County, Arizona	Phoenix-Mesa-Scottsdale
48201	2	Harris County, Texas	Houston-The Woodlands-Sugar Land
6037	3	Los Angeles County, California	Los Angeles-Long Beach-Anaheim
48113	4	Dallas County, Texas	Dallas-Fort Worth-Arlington
17031	5	Cook County, Illinois	Chicago-Naperville-Elgin
48439	6	Tarrant County, Texas	Dallas-Fort Worth-Arlington
53033	7	King County, Washington	Seattle-Tacoma-Bellevue
48453	8	Travis County, Texas	Austin-Round Rock
27053	9	Hennepin County, Minnesota	Minneapolis-St. Paul-Bloomington
48029	10	Bexar County, Texas	San Antonio-New Braunfels
13121	11	Fulton County, Georgia	Atlanta-Sandy Springs-Roswell
6073	12	San Diego County, California	San Diego-Carlsbad
32003	13	Clark County, Nevada	Las Vegas-Henderson-Paradise
48085	14	Collin County, Texas	Dallas-Fort Worth-Arlington
48121	15	Denton County, Texas	Dallas-Fort Worth-Arlington
8031	16	Denver County, Colorado	Denver-Aurora-Lakewood
37119	17	Mecklenburg County, North Carolina	Charlotte-Concord-Gastonia
12057	18	Hillsborough County, Florida	Tampa-St. Petersburg-Clearwater
39049	19	Franklin County, Ohio	Columbus
12095	20	Orange County, Florida	Orlando-Kissimmee-Sanford

Notes: A list of top ranked counties in the U.S. based on 2019 migration statistics using Geographic PageRank.

Table D.3: Estimation of Home Price Elasticity to Median AQI

	OLS	Income		PageRank	
	(1)	(2) Control	(3) Proxy	(4) Control	(5) Proxy
Air Pollution (2019)	0.146***	0.0380***	-0.0794***	0.0759***	-0.406***
	(0.0109)	(0.00841)	(0.0121)	(0.0124)	(0.0550)
HH Median Income		1.380***	2.882***		
		(0.0344)	(0.0710)		
PageRank				0.210***	1.661***
				(0.0162)	(0.143)
Constant	11.59***	-3.809***	-20.57***	13.26***	24.75***
	(0.0213)	(0.384)	(0.790)	(0.124)	(1.164)
Room Fixed Effects	✓	✓	✓	✓	✓
Observations	8322	8322	8322	8322	8322
Adjusted R^2	0.410	0.692		0.485	

Notes: This table shows the estimated price elasticity to median AQI at the county-level in 2019 using home price index data from Zillow and interpolated median AQI data from EPA. Column 1 shows the OLS result from simply regressing house price index on median AQI. Column 2 and Column 4 regress house price index controlling for household median income and geographic PageRank, respectively. Column 3 and Column 5 apply the anti-IV method, using household median income and geographic PageRank as the anti-IV, respectively. The geographic PageRank is the 2019 nationwide version constructed using IRS migration inflow data, and HH median incomes comes from ACS. Standard errors are clustered at the county level.