A WARNING ABOUT USING PREDICTED VALUES FROM REGRESSION MODELS FOR EPIDEMIOLOGIC INQUIRY

A PREPRINT

Elizabeth L. Ogburn
Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health
Baltimre, MD 21205
eogburn@jhsph.edu

Kara E. Rudolph
Department of Epidemiology
Columbia Mailman School of Public Health
New York, NY

Rachel Morello-Frosch
Department of Environmental Science, Policy and Management
and School of Public Health
University of California, Berkeley
Berkeley, CA

Amber Khan
University of Washington School of Public Health
Department of Environmental and Occupational Health Sciences
Seattle, WA

Joan A. Casey
Department of Environmental Health Sciences
Columbia Mailman School of Public Health
New York, NY

October 6, 2020

Source of Funding

This research is based upon work supported by the Urban Institute through funds provided by the Robert Wood Johnson Foundation. We thank them for their support but acknowledge that the findings and conclusions presented in this report are those of the author(s) alone, and do not necessarily reflect the opinions of the Urban Institute or the Robert Wood Johnson Foundation.

A WARNING ABOUT USING PREDICTED VALUES FROM REGRESSION MODELS FOR EPIDEMIOLOGIC INQUIRY

ABSTRACT

In many settings researchers may not have direct access to data on one or more variables needed for an analysis, and instead may use regression-based estimates of those variables. Using such estimates in place of original data, however, introduces complications and can result in uninterpretable analyses. In simulations and observational data we illustrate the issues that arise when an average treatment effect is estimated from data where the outcome of interest is a prediction from an auxiliary model. We show that bias in any direction can result, both under the null and alternative hypotheses.

***K*eywords** proxy variables · measurement error · imputation

Introduction

In many settings researchers lacking direct access to data on variables needed for an analysis rely instead on regression-based estimates of those variables. Regression-based estimates are routinely used as exposures, outcomes, and effect modifiers or covariates in epidemiologic studies, including some appearing in top epidemiology journals [1, 2, 3, 4, 5, 6]. These include some types of small area estimates [7, 8], models to estimate the composition of biological samples [9], models that combine biomarker data with demographic covariates [10], and models of frailty in aging research [2, 11]. Our motivating example is a small area estimation model [7, 8] which purports to be useful for "informing local health policy-makers, improving community-based public health program planning and intervention strategy development, and facilitating public health resource allocation and delivery." But using regression-based estimates in place of original data can result in uninterpretable analyses and is rarely valid.

We use *primary analysis* to refer to the analysis that answers a question of interest and *primary data* to refer to the data used in the primary analysis. We use *auxiliary model* to refer to a regression model, fit to *auxiliary data*, that provides estimates of a variable, V, that is required for the primary analysis but is not available in the primary data. We show in simulations that the common use of predictions from auxiliary models in primary analyses can result in spurious association estimates, including estimates that are in the opposite direction from the truth.

Terminology and notation	
primary analysis	analysis that answers question of interest
primary data	data used in primary analysis
V	variable required for primary analysis but not available in primary data
auxiliary model	regression model used to estimate V
auxiliary data	data used for auxiliary model
Z	auxiliary data covariates used as independent variables in model for V
$g(Z)$	function that predicts V; result of the auxiliary model

Our warnings apply to any estimate of V that can be expressed as a function of covariates, i.e., the regressors in the auxiliary model. They do not apply to the use of predicted values from deterministic mathematical models, such as differential equation models of infectious disease epidemics or models that smooth between observations, e.g., kriging to estimate air pollution levels. They also do not apply to settings amenable to multiple imputation, where V is observed for some but not all subjects in the primary data.

What may go wrong when using estimates from auxiliary regression models

Suppose an auxiliary model uses covariates Z to estimate a variable, V, that researchers need for their primary analysis. As a toy example, suppose that V is blood pressure, and Z includes age, sex, body mass index (BMI), race/ethnicity, and socioeconomic status (SES). The result of the auxiliary model is a function $g(Z)$ (e.g., $Z'Z\beta$ if the auxiliary model is a linear regression) that outputs estimated values of V. Plugging the Z for a particular observation into $g(Z)$ estimates that observation's V.

If V is a confounder in a primary analysis, using the estimated V values $g(Z)$ to control for V cannot do better than optimally controlling directly for Z, because a function of Z can only contain less information than Z alone. In our toy example, controlling for estimated blood pressure can do no better than correctly specifying a model for the outcome given the exposure and age, sex, BMI, race/ethnicity, and SES. A paper published in *Epidemiology* recommended that researchers interested in estimating the effect of medical interventions on mortality among older adults using Medicare claims data control for confounding by frailty using a function of age, sex, race/ethnicity, and 20 claims for conditions, symptoms, and medical equipment (e.g., wheelchair, vertigo, and dementia) shown to be predictive of frailty [3]. If the association of interest is unconfounded after controlling for these 23 variables, in addition to other measured covariates, then all 23 variables should be included in any claims-based analysis. If confounding remains after conditioning on these 23 variables, then no function of them can control for confounding—no matter how predictive it is of frailty.

If V is an exposure or treatment of interest, including $g(Z)$ instead of V in a primary analysis only tells us about the relationship between a function of Z and the outcome. One problem is that $g(Z)$ may fail to capture features of V that drive the true effect of interest. Another is that it may introduce spurious associations. For example, suppose SES has a strong relationship with the outcome of interest. Then V=blood pressure may appear to be a strong predictor of the outcome simply because SES is an important component of $g(Z)$. In the *International Journal of Epidemiology*, researchers suggested estimating early- and mid-life cardiovascular disease (CVD) risk factors as a function of race/ethnicity, sex, and age cohort in order to use cohort studies of older adults to estimate the associations of early life risk factors on later life outcomes [1].[1] They suggest that estimated risk factors can be used to assess "the effects of early and midlife exposures on later life outcomes," but any apparent association between these risk factors and later life outcomes could also be driven by associations with age cohort, sex, or race/ethnicity unrelated to CVD risk factors.

If V is an outcome, using $g(Z)$ instead of V only tells us about how well the variables in our model of interest predict a function of Z. As above, $g(Z)$ may fail to capture features of V that are truly affected by the exposure of interest, or it may introduce spurious associations. In our toy example, if SES is strongly associated with an exposure of interest (as indeed it often is), this could drive an apparent but spurious relationship with predicted blood pressure. A paper published in this journal compared definitions of incident chronic kidney disease based on estimated glomerular filtration rate (eGRF), which is a function of age, sex, Black race/ethnicity, and creatinine, to a definition based only on creatinine, which is a biomarker for GRF [4]. The creatinine definition identified male sex as a risk factor for incident CKD whereas the eGRF definitions identified male sex as protective, suggesting "the nonlinear relation between serum creatinine and eGFR may partially explain this apparent discrepancy." However, the eGRF equation included sex as a factor thereby defining eGRF as 1.35 times higher for males than for females. Whenever eGRF is used as an outcome in an epidemiologic study, e.g., [5], this same discrepancy could bias estimates of associations with any exposure that is differential by sex either towards or away from the null. The use of race in eGRF and other clinical algorithms has come under increasing criticism [12, 13]; the phenomenon we have just described is one reason why using race as a predictor can be highly problematic.

When V is an exposure or an outcome, controlling for any elements of Z in the primary analysis can undermine the predictive power of $g(Z)$ to capture information about V. In the extreme case, controlling for Z could be tantamount to controlling for $g(Z)$ itself, forcing all estimated associations with $g(Z)$ to be null. In the CVD risk factor analysis [1], any estimate of the predictor $g(Z)$ that controls for sex, race/ethnicity, or age cohort would likely be attenuated and any estimate that stratifies by all three covariates would be null.

The problem of a missing variable is an extreme case of missing data, where V is missing for every observation. It is well-known from the literature on missing data imputation that the joint relationships among all the variables in a primary analysis must be correctly modeled in the auxiliary data in order for estimated V to be a valid replacement for the true V [14]. This may not be a problem in most missing data settings where all of the relevant variables are available in a single data set. But in our setting, this generally entails that all of the variables to be included in a primary analysis are also included in the auxiliary model for V, with all of the relationships correctly specified. This is almost never possible: researchers may not have access to the auxiliary data or oversight over how the auxiliary model is fit, and the auxiliary data may not include all of the primary analysis variables.

In the best-case scenario of a correctly-specified auxiliary model and informative predictors, the auxiliary model may be very predictive of V, with highly statistically significant coefficients, high R^2, etc. Nevertheless, these criteria are not sufficient to validate the use of predicted V values in subsequent analyses; no matter how "good" the auxiliary model is for V, if it does not correctly capture the joint relationships among all of the primary analysis variables, it can introduce

[1] They first estimated smoking as a function of race, sex, age cohort. Denote this function $s(race, sex, cohort)$. Next they estimated BMI as a function of race, sex, age cohort, and estimated smoking: $b(s(race, sex, cohort), race, sex, cohort)$ – so BMI is also a function only of race, sex, and age. They subsequently estimated other risk factors as functions of race/ethnicity, sex, cohort, and previously estimated risk factors.

meaningful bias into the primary analysis. The bias introduced by replacing V with its estimate will depend on many factors, and using the predicted V may be better or worse than ignoring V altogether or using a more naive proxy for V. In simulations we show that associations estimated using predicted V can go in the opposite direction from the true association, whether V is a confounder, an outcome, or an exposure.

When is it acceptable to use estimates from auxiliary regression models?

Let W be the collection of primary analysis variables that are not available in the auxiliary data. The strategy of predicting V using auxiliary data on V and Z is unproblematic when either of these two conditional independences holds:

$$Z \perp W|V \tag{1}$$

or

$$V \perp W|Z. \tag{2}$$

Equivalently, the joint distribution of V, Z, and W factorizes into one component that is a function of V and Z, and another component that is a function of W and either Z or V, but not both.

In this case, the joint relationships among Z,V, and W can be correctly modeled in the auxiliary data even though W does not appear in the auxiliary data, because the regression of V on Z only involves $f(Z, V)$ and not the other factor in full joint distribution. When the joint distribution factorizes in this way, using $g(Z)$ in place of V is akin to observing V with error, where the magnitude of the error is directly related to how well $g(Z)$ predicts V.

Motivating example

The Centers for Disease Control and Prevention (CDC) 500 Cities data uses small area estimation models to estimate the prevalence of health outcomes at the census tract-level based on county-level data. The auxiliary model uses county-level data to regress V onto age category, sex, race/ethnicity category, and county-level poverty rate. Census tract data on covariates, Z, coupled with the county-level fitted auxiliary model, is used to predict V at the census tract-level, $g(Z)$.[2] Although the CDC website recommends that these data be used for surveillance purposes, the original papers proposing the small area models suggest that they can be used to design and assess interventions [7, 8] and they have frequently been used as outcomes [15, 16, 17, 18], exposures [17], and covariates in epidemiologic research, sometimes simultaneously.

Simulations

We considered two different simulation settings to mimic analyses of the 500 Cities data that use auxiliary model predictions as outcome variables in the primary analysis.

We first simulated the county-level data: 4 covariates based on the distributions of the 4 predictors used in the 500 Cities small area estimation models, the exposure with approximately the same marginal distribution as day-night average noise, and the outcome with approximately the same conditional distribution as poor sleep conditional on the exposure of noise.We ran an auxiliary model regressing prevalence of poor sleep on the 4 auxiliary covariates.

Next, we simulated the primary data at the census tract-level, using the same data generating models as the county-level data. We simulated the true outcome, V, for each observation. To mimic settings in which researchers do not have access to the true V, we also calculated a predicted outcome $g(Z)$ for each observation by using the auxiliary model. We compared two analyses, both using the primary data: one using the true V to assess the association between exposure and V, and one using the predicted values $g(Z)$ that would be available to a researcher. The results are depicted in Figure 1. In panel A the association reverses direction when V is replaced with $g(Z)$. In panel B, we illustrate the perils of controlling for an auxiliary model predictor in the primary analysis. Although both the exposure-V and exposure-$g(Z)$ associations are positive, controlling for race/ethnicity results in a null association.

We also ran two simulations to illustrate the potential consequences of using $g(Z)$ as a covariate or exposure (Figure 2). For panel A, we simulated a continuous covariate V, conditional on an auxiliary predictor Z. We then simulated exposure and outcome to have a strong negative association conditional on V. We ran this same data-generating process in auxiliary and analysis data sets. We regressed V on Z in the auxiliary data, and then predicted $g(Z))$ in the analysis data. Conditional on $g(Z)$, exposure and outcome have a strong positive association (in the opposite direction of the

[2]Random effects were included only at the state and county level, meaning that within county the V is predicted from essentially an ordinary regression model.

truth). For panel B, in auxiliary and analysis data we simulated a continuous outcome outcome, and then simulated a continuous exposure V, conditional on an auxiliary predictor but independent of the outcome. We used the model fit from regressing V on the auxiliary predictor in the auxiliary data to calculate $g(Z)$ in the analysis data. Despite the true null relationship, $g(Z)$ and the outcome are strongly positively associated.

All code is available at [BLINDED].

Observational data analysis

Using logistic regression, we examined associations between day-night average noise and poor sleep using the 500 Cities data. Day-night average noise was estimated from a geospatial sound model over a 24-hour period with a 10-decibel penalty added between 10PM and 7AM [19]. Potential census tract-level confounding variables from the 2011-2015 U.S. American Community Survey included proportion: females, individuals >75 years, non-Hispanic White and non-Hispanic Black individuals, homeowners, and individuals living below the federal poverty threshold; and population density (individuals/km^2). We included modeled estimates of census tract level NO_2 and $PM_{2.5}$ from the Center for Air, Climate and Energy Solutions [20]. [3]

Although our results appear to show a strong association between noise and poor sleep, these results could be driven by an association between noise and one of the auxiliary predictors used in the CDC small-area estimation models. For example, socioeconomic status has previously been shown to be strongly associated with day-night average noise in the United States [21].

Conclusion

Predictions from auxiliary regression models should not generally be used to learn about associations, mechanisms, or causal effects. While the predictions themselves can be useful for descriptive purposes, such as health surveillance or predicting burden of disease, including them as outcomes, exposures, or covariates in larger models for the purpose of estimating associations or causal effects can result in bias.

The exception to this general principle is when the joint distribution of the variables in the union of the auxiliary and primary data (X, V, and W) factorizes into two terms, one of which involves only X and V and the other of which may involve W and either X or V but not both. This is the case whenever V is a deterministic function of X. In future work we plan to formalize this principle, quantify departures from the factorization, and characterize the relationship between such departures and potential bias.

References

[1] Adina Zeki Al Hazzouri, Eric Vittinghoff, Yiyi Zhang, Mark J Pletcher, Andrew E Moran, Kirsten Bibbins-Domingo, Sherita H Golden, and Kristine Yaffe. Use of a pooled cohort to impute cardiovascular disease risk factors across the adult life course. *International journal of epidemiology*, 48(3):1004–1013, 2019.

[2] Jodi B Segal, Jin Huang, David L Roth, and Ravi Varadhan. External validation of the claims-based frailty index in the national health and aging trends study cohort. *American journal of epidemiology*, 186(6):745–747, 2017.

[3] Carmen C Cuthbertson, Anna Kucharska-Newton, Keturah R Faurot, Til Stürmer, Michele Jonsson Funk, Priya Palta, B Gwen Windham, Sydney Thai, and Jennifer L Lund. Controlling for frailty in pharmacoepidemiologic studies of older adults: validation of an existing medicare claims-based algorithm. *Epidemiology (Cambridge, Mass.)*, 29(4):556, 2018.

[4] Lori D Bash, Josef Coresh, Anna Köttgen, Rulan S Parekh, Tibor Fulop, Yaping Wang, and Brad C Astor. Defining incident chronic kidney disease in the research setting: The aric study. *American journal of epidemiology*, 170(4):414–424, 2009.

[5] Ana Navas-Acien, Maria Tellez-Plaza, Eliseo Guallar, Paul Muntner, Ellen Silbergeld, Bernard Jaar, and Virginia Weaver. Blood cadmium and lead and chronic kidney disease in us adults: a joint analysis. *American journal of epidemiology*, 170(9):1156–1164, 2009.

[6] Brendan Darsie, Michael G Shlipak, Mark J Sarnak, Ronit Katz, Annette L Fitzpatrick, and Michelle C Odden. Kidney function and cognitive health in older adults: the cardiovascular health study. *American journal of epidemiology*, 180(1):68–75, 2014.

[3] For this analysis, we ignore issues related to modeling noise and air pollution levels, but note that these models include deterministic components that differentiate them from the auxiliary regression models of primary concern.

[7] Xingyou Zhang, James B Holt, Hua Lu, Anne G Wheaton, Earl S Ford, Kurt J Greenlund, and Janet B Croft. Multilevel regression and poststratification for small-area estimation of population health outcomes: a case study of chronic obstructive pulmonary disease prevalence using the behavioral risk factor surveillance system. *American journal of epidemiology*, 179(8):1025–1033, 2014.

[8] Xingyou Zhang, James B Holt, Shumei Yun, Hua Lu, Kurt J Greenlund, and Janet B Croft. Validation of multilevel regression and poststratification methodology for small area estimation of health indicators from the behavioral risk factor surveillance system. *American journal of epidemiology*, 182(2):127–137, 2015.

[9] Eugene Andres Houseman, William P Accomando, Devin C Koestler, Brock C Christensen, Carmen J Marsit, Heather H Nelson, John K Wiencke, and Karl T Kelsey. Dna methylation arrays as surrogate measures of cell mixture distribution. *BMC bioinformatics*, 13(1):86, 2012.

[10] Andrew S Levey, Juan P Bosch, Julia Breyer Lewis, Tom Greene, Nancy Rogers, and David Roth. A more accurate method to estimate glomerular filtration rate from serum creatinine: a new prediction equation. *Annals of internal medicine*, 130(6):461–470, 1999.

[11] Dane R Van Domelen and Karen Bandeen-Roche. A note on proposed estimation procedures for claims-based frailty indexes. *American journal of epidemiology*, 2019.

[12] Nwamaka Denise Eneanya, Wei Yang, and Peter Philip Reese. Reconsidering the consequences of using race to estimate kidney function. *Jama*, 322(2):113–114, 2019.

[13] Darshali A Vyas, Leo G Eisenstein, and David S Jones. Hidden in plain sight—reconsidering the use of race correction in clinical algorithms, 2020.

[14] Xiao-Li Meng. Multiple-imputation inferences with uncongenial sources of input. *Statistical Science*, pages 538–558, 1994.

[15] Yan Li, Shelley H Liu, Li Niu, and Bian Liu. Unhealthy behaviors, prevention measures, and neighborhood cardiovascular health: a machine learning approach. *Journal of Public Health Management and Practice*, 25(1):E25–E28, 2019.

[16] Kevin M Fitzpatrick, Xuan Shi, Don Willis, and Jill Niemeier. Obesity and place: Chronic disease in the 500 largest us cities. *Obesity research & clinical practice*, 12(5):421–425, 2018.

[17] Shelley H Liu, Bian Liu, and Yan Li. Risk factors associated with multiple correlated health outcomes in the 500 cities project. *Preventive medicine*, 112:126–129, 2018.

[18] Yan Wang, James B Holt, Fang Xu, Xingyou Zhang, Daniel P Dooley, Hua Lu, and Janet B Croft. Using 3 health surveys to compare multilevel models for small area estimation for chronic diseases and health behaviors. *Preventing chronic disease*, 15, 2018.

[19] Daniel Mennitt, Kirk Sherrill, and Kurt Fristrup. A geospatial model of ambient sound pressure levels in the contiguous united states. *The Journal of the Acoustical Society of America*, 135(5):2746–2764, 2014.

[20] Hankey S Sheppard L Szpiro AA Marshall JD. 2018 S.-Y. K, Bechle M. Concentrations of criteria pollutants in the contiguous u.s., 1979 – 2015: Role of prediction model parsimony in integrated empirical geographic regression. *Under review*.

[21] Joan A Casey, Rachel Morello-Frosch, Daniel J Mennitt, Kurt Fristrup, Elizabeth L Ogburn, and Peter James. Race/ethnicity, socioeconomic status, residential segregation, and spatial variation in noise exposure in the contiguous united states. *Environ Health Perspect*, 125(7):077017, 2017.

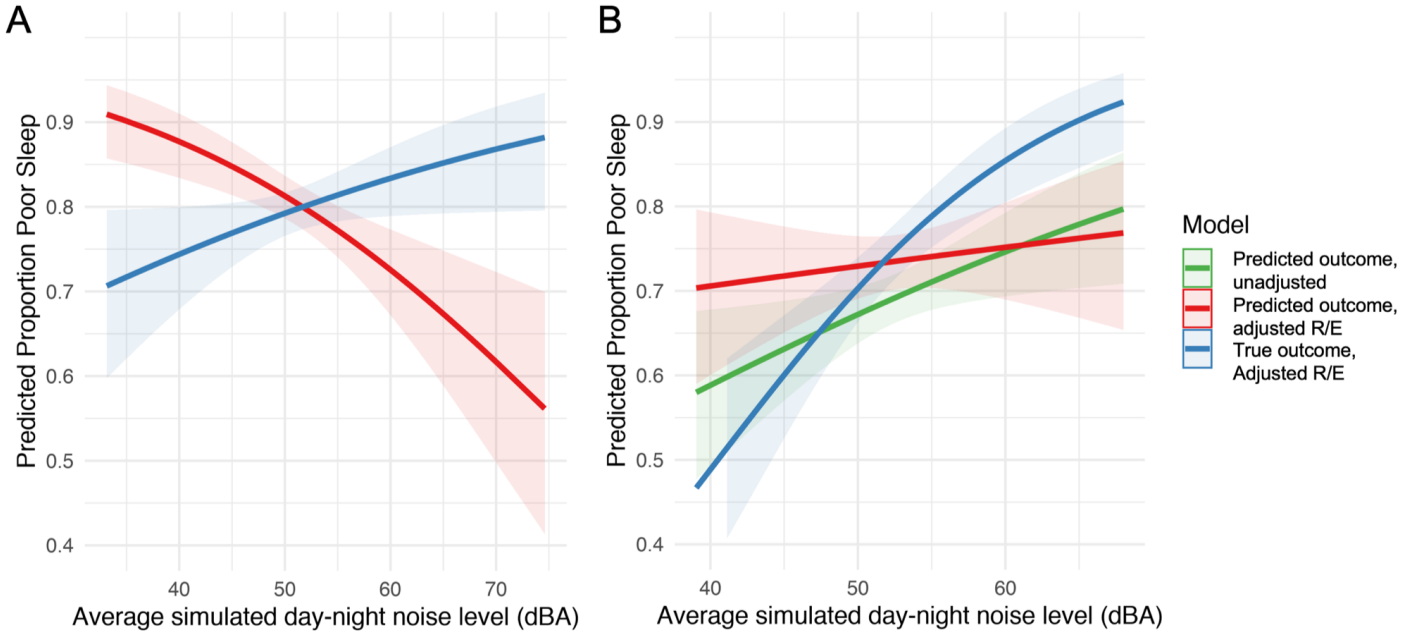

Figure 1: Simulated associations between day-night average noise level and proportion with poor sleep under two different scenarios. The figure displays predicted values the corresponding 95% confidence intervals from logistic regression models. The models in panel A are unadjusted, comparing the associations of the true (blue) and estimated (red) outcomes with exposure. In panel B the unadjusted associations between exposure and $g(Z)$ (green) and the association between exposure and V adjusted for race/ethnicity (blue) are positive, but the association between exposure and $g(Z)$ adjusted for race/ethnicity (red) is null.

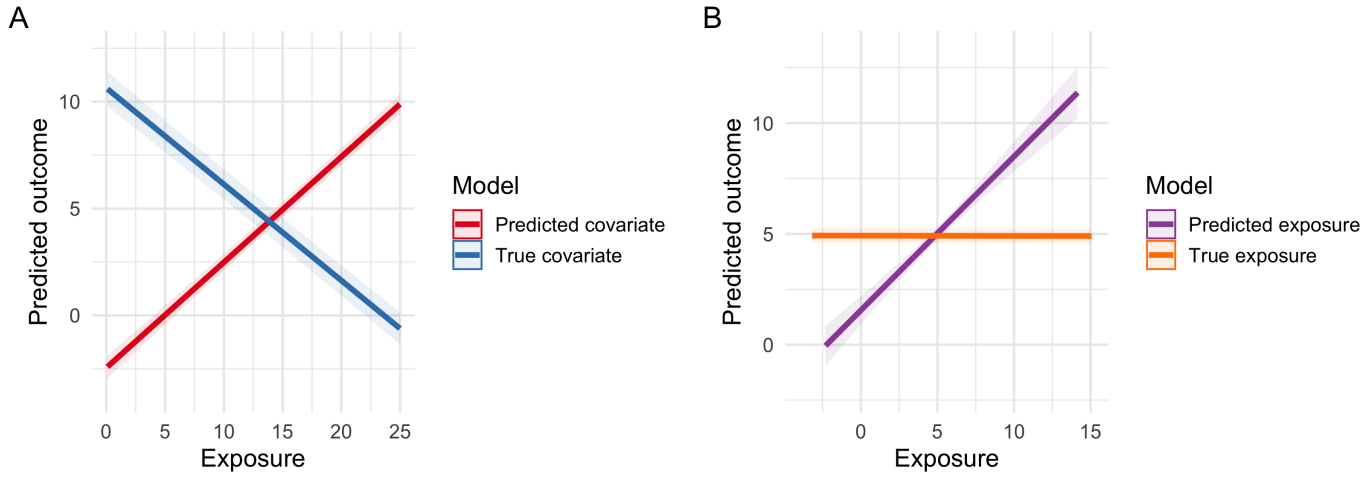

Figure 2: Simulated associations between exposure and outcome. The figure displays predicted values and the corresponding 95% confidence intervals from linear regression models. Panel A illustrates the association between exposure and outcome adjusted for a true (blue) versus predicted (red) covariate and the model in panel B illustrates the unadjusted association between an outcome and a predicted (purple) and true (orange) exposure.

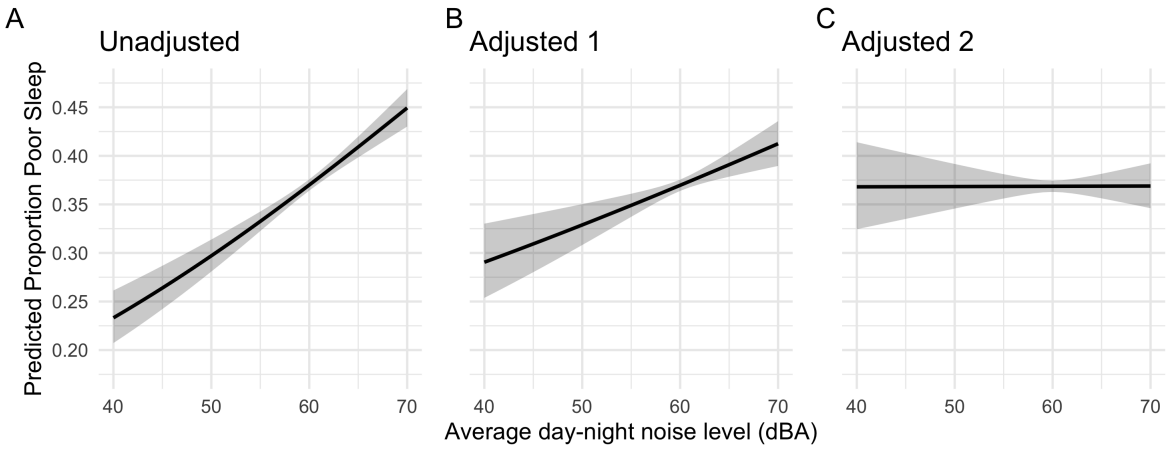

Figure 3: Association between day-night average noise and predicted proportion poor sleep in the nationwide 500 Cities dataset. The figure displays predicted values from logistic regression models. Panel A displays the unadjusted model. Adjusted model 1 (panel B) includes only covariates that are not in the auxiliary model: population density, $N0^2$, $PM_{2.5}$, and homeownership. Adjusted model 2 (panel C) includes the same covariates plus those from the auxiliary model: age category, sex, race/ethnicity category, and county-level poverty rate.