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

January 14, 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.

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 regression 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, under both the null and alternative hypotheses.

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

Introduction

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. Examples include small area estimates of variables only directly measured at a geographic scale larger than the scale of the analysis [1, 2, 3, 4]; models to estimate the composition of biological samples that are mixtures of different cell types [5]; and models of frailty in aging research [6]. Using regression-based estimates in place of original data, however, can result in uninterpretable analyses.

We use *primary analysis* to refer to the analysis that a researcher uses to answer 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. Use of predictions from auxiliary models in primary analyses, though common in practice [7, 8, 9, 10, 11, 12], is usually invalid. We show in simulations that it can result in spurious association estimates, including estimates that are in the opposite direction from the truth.

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

If V is a confounder in a primary analysis, using the estimated V values $g(X)$ to control for V can do no better than optimally controlling directly for X. In our toy example, controlling for estimated blood pressure can do no better than simply controlling for age, sex, BMI, race, and SES directly (and optimally). If V is an exposure or treatment of interest, including $g(X)$ instead of V in a primary analysis only tells us about the relationship between a function of X and the outcome. For example, suppose age has a strong relationship with the outcome of interest. Then blood pressure may appear to be a strong predictor of the outcome simply because age is an important component of $g(X)$. If V is an outcome, using $g(X)$ instead of V only tells us about how well the variables in our model of interest predict a function of X. For example, if age is strongly predicted by an exposure of interest, this could drive an apparent but spurious relationship with predicted blood pressure. When V is an exposure or an outcome, controlling for any elements of X in the primary analysis can undermine the predictive power of $g(X)$ to capture information about V. In the extreme case, controlling for X could be tantamount to controlling for $g(X)$ itself, in an analysis designed to assess associations with $g(X)$, forcing all estimated associations to 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 [13]. This may not be a problem in most missing data settings, because 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 (if it did, researchers could simply run their primary analysis on the auxiliary data).

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 for V, if it does not correctly capture the joint relationships among all of the primary analysis variables, it can introduce meaningful bias into the primary analysis. Of course, the bias introduced by replacing V with its estimate will depend on many factors, and sometimes using the predicted V may be better than ignoring V altogether or using a more naive proxy for V.

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 X is unproblematic when either of these two conditional independences holds:

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

or

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

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

In this case, the joint relationships among X,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 X only involves $f(X, V)$ and not the other factor in full joint distribution. This will be the case, for example, if X perfectly predicts V, that is V is exactly equal to $g(X)$. Then we would say that $g(X)$ is a deterministic, rather than a regression model for V. In our toy example, this criterion would be met if the primary analysis were about the relationship between blood pressure and W=*sphygmometer measurement error*, since age, sex, BMI, SES, and race are all independent of W.

When the joint distribution factorizes in this way, using $g(X)$ in place of V is akin to observing V with error, where the magnitude of the error is directly related to how well $g(X)$ predicts V.

Motivating example

As a case study, we consider the Centers for Disease Control and Prevention (CDC) 500 Cities data, which uses small area estimation models to estimate the prevalence of various health outcomes, health behaviors, and health prevention measures at the census tract level, based on county-level data. In this case, the auxiliary model uses county-level data to regress Y as a function of age category, sex, race/ethnicity category, and county-level poverty rate. Census tract data on covariates, coupled with the county-level fitted auxiliary model, is used to predict Y at the census tract level. Although the CDC website recommends that these data be used for surveillance purposes, they have been used for epidemiologic research [1, 2, 3, 4]. We use the 500 Cities data and simulations to illustrate how use of auxiliary model-predicted outcomes in a primary analysis can result in severely biased effect estimates.

Simulations

We considered four different simulation settings to mimic analyses of the 500 Cities data that use auxiliary model predictions as outcome variables in the primary analysis. The auxiliary model is fit using county-level data and outcome variable predictions are then made at the census tract-level. The primary analysis estimates the census tract-level relationship between day-night average noise and the predicted outcome of interest (poor sleep; prevalence of adults aged 18 years that usually slept less than 7 hours in a 24-hour period), controlling for covariates.

We first simulated the county-level data. We simulated 4 covariates based on the distributions of the 4 predictors used in the 500 Cities small area estimation models: age, sex, race/ethnicity category, and proportion below the federal poverty threshold. We simulated additional covariates based on the distributions of population density, proportion of home ownership, and ambient levels of nitrogen dioxide (NO_2) and particulate matter ($PM_{2.5}$) from the census tracts included in the 500 Cities data. We then simulated the exposure, X, in some settings conditional on covariates, with approximately the same marginal distribution as day-night average noise. Finally, we simulated poor sleep (Y conditional on covariates and/or exposure, depending on the simulation setting. After simulating the auxiliary data, we

ran an auxiliary model regressing prevalence of poor sleep on age, sex, race/ethnicity, and proportion below the federal poverty threshold.

Next, we simulated data at the census tract-level, using the same data generating models as the county-level data. These are the primary data. This ensures that the auxiliary model fit to the county-level data is also valid for the tract-level data, which may not always be the case in practice but allows us to isolate issues due to using estimates from auxiliary models. We simulated the true outcome, Y, for each observation. To mimic settings in which researchers do not have access to the true Y, we also calculated a predicted outcome, $\hat{Y}$, for each observation by using the auxiliary model. In each scenario, we compared two analyses, both using the primary data: one using the true Y to assess the association between X and Y, and one using the predicted values $\hat{Y}$ that would be available to a researcher. The results are depicted in Figure 1. Settings (A)-(C) show that associations using predicted outcomes can be in conflict with the true underlying associations. For example, in setting (A), there is no association between X and Y, but there is an apparent association between X and $\hat{Y}$ and in (C) we observe the inverse. In (B) the association reverses direction when Y is replaced with $\hat{Y}$. In setting (D), we illustrate the perils of controlling for a auxiliary model predictor, race/ethnicity, in the primary analysis. Code is available at `https://github.com/joanacasey/auxiliary_mod_perils`.

Observational data analysis

Using logistic regression, we examined unadjusted and adjusted associations between day-night average noise and poor sleep using the 500 Cities data. Day-night average noise, noise over a 24-hour period with a 10-decibel penalty added between 10PM and 7AM, came from a geospatial sound model [14]. We also assembled potential census tract-level confounding variables from the 2011-2015 U.S. American Community Survey: proportion females, individuals > 75 years of age, non-Hispanic White and non-Hispanic Black individuals, homeowners, and individuals living below the federal poverty threshold; and population density (individuals/km^2). Finally, we included modeled estimates of census tract level NO_2 and $PM_{2.5}$ from the Center for Air, Climate and Energy Solutions [15]. For this analysis, we ignore issues related to modeling air pollution levels, which can suffer from some of the same problems we describe from other auxiliary models.

Although our results appear to show a strong association between noise and poor sleep, these results are consistent with the true association being null or protective, as in the simulations in Figures 1(A) and (B). The 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 [16].

Conclusion

We argue that predictions from auxillary regression models–like those from the 500 Cities small area estimation 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, that is if X, V, and W, factorizes into two terms, one of which involves only X and V and the other of which may involve 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] PI Eke, X Zhang, H Lu, L Wei, G Thornton-Evans, KJ Greenlund, JB Holt, and JB Croft. Predicting periodontitis at state and local levels in the united states. *Journal of dental research*, 95(5):515–522, 2016.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] Tatyana Shamliyan, Kristine MC Talley, Rema Ramakrishnan, and Robert L Kane. Association of frailty with survival: a systematic literature review. *Ageing research reviews*, 12(2):719–736, 2013.

[7] Adyasha Maharana and Elaine Okanyene Nsoesie. Use of deep learning to examine the association of the built environment with prevalence of neighborhood adult obesity. *JAMA Network Open*, 1(4):e181535–e181535, 2018.

[8] Matthew Browning and Alessandro Rigolon. Do income, race and ethnicity, and sprawl influence the greenspace-human health link in city-level analyses? findings from 496 cities in the united states. *International journal of environmental research and public health*, 15(7):1541, 2018.

[9] Yuru Huang, Dina Huang, and Quynh C Nguyen. Census tract food tweets and chronic disease outcomes in the us, 2015–2018. *International journal of environmental research and public health*, 16(6):975, 2019.

[10] Joseph L Servadio, Abiola S Lawal, Tate Davis, Josephine Bates, Armistead G Russell, Anu Ramaswami, Matteo Convertino, and Nisha Botchwey. Demographic inequities in health outcomes and air pollution exposure in the atlanta area and its relationship to urban infrastructure. *Journal of Urban Health*, 96(2):219–234, 2019.

[11] 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.

[12] Kyungsoon Wang and Dan Immergluck. The geography of vacant housing and neighborhood health disparities after the us foreclosure crisis. *Cityscape*, 20(2):145–170, 2018.

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

[14] 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.

[15] 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*.

[16] 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.

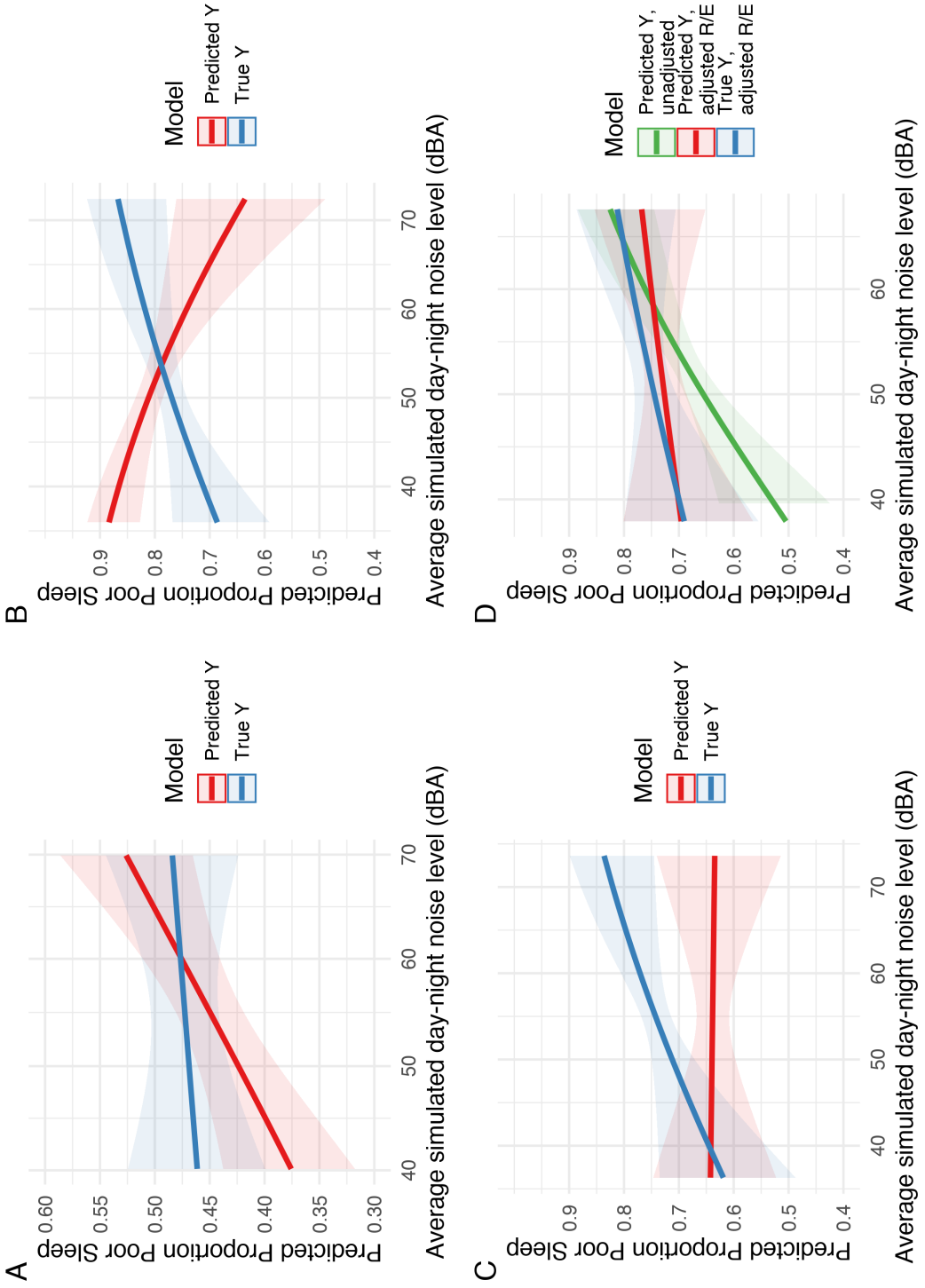

Figure 1: Simulated associations between day-night average noise level and proportion with poor sleep under four different scenarios. The figure displays predicted values from logistic regression models. Models (A)-(C) were unadjusted and model (D) controlled race/ethnicity.

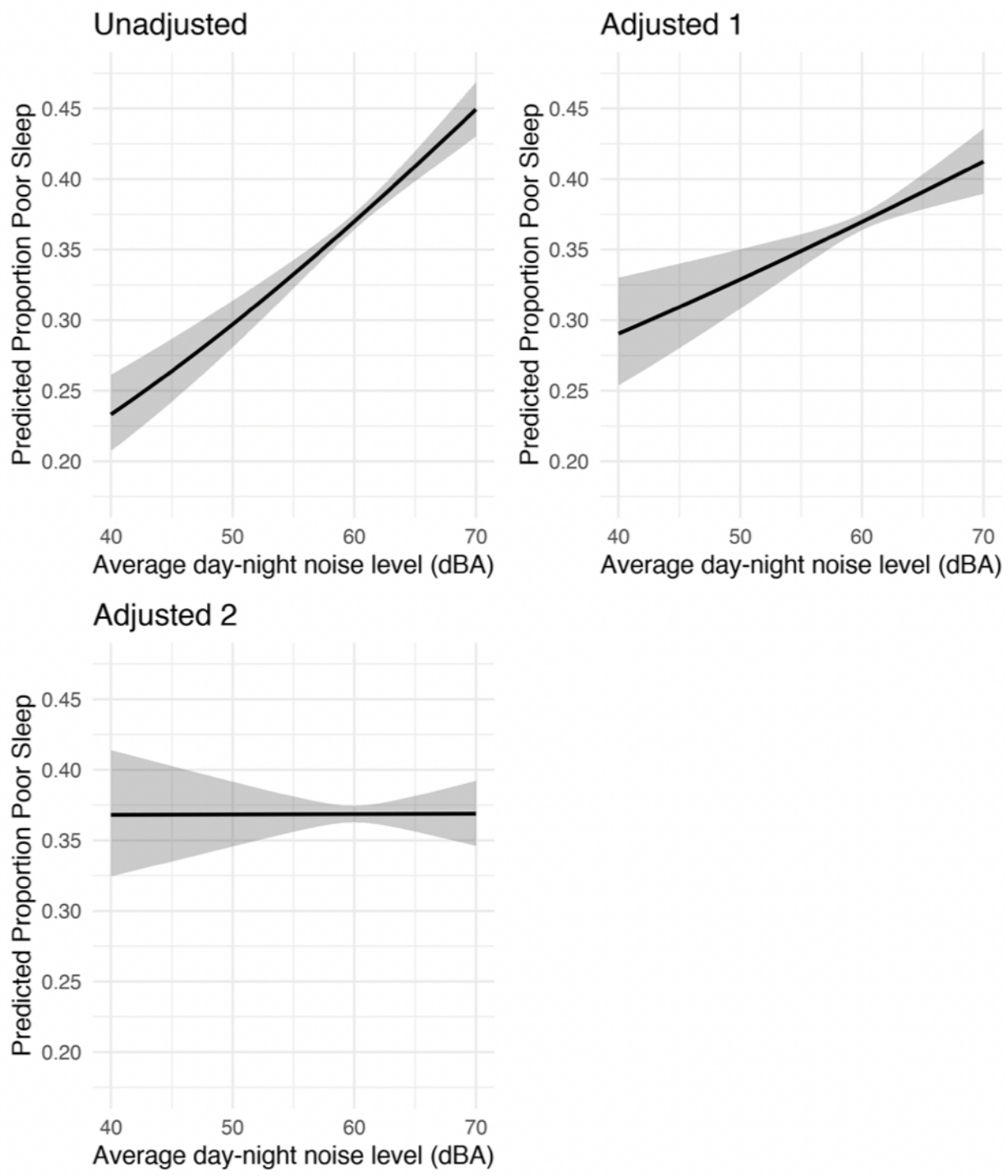

Figure 2: Association between day-night average noise and predicted proportion poor sleep in the nationwide 500 Cities dataset. The figure displays predicted values from logistic regression models. Adjusted model 1 includes only covariates that are not in the auxiliary model: population density, $N0^2$, $PM_{2.5}$, and homeownership. Adjusted model 2 includes the same covariates plus those from the auxiliary model: proportion females, individuals > 75 years of age, non-Hispanic White and non-Hispanic Black individuals, and individuals living below the federal poverty threshold.