American Journal of Epidemiology

Vol. 178, No. 8
DOI: 10.1093/aje/kwt127
Advance Access publication:
September 5, 2013

Practice of Epidemiology

Measurement Error in Epidemiologic Studies of Air Pollution Based on Land-Use Regression Models

Xavier Basagaña*, Inmaculada Aguilera, Marcela Rivera, David Agis, Maria Foraster, Jaume Marrugat, Roberto Elosua, and Nino Künzli

* Correspondence to Dr. Xavier Basagaña, Centre for Research in Environmental Epidemiology (CREAL), Doctor Aiguader 88, 08003 Barcelona, Catalonia, Spain (e-mail: xbasagana@creal.cat).

Initially submitted August 20, 2012; accepted for publication May 22, 2013.

Land-use regression (LUR) models are increasingly used to estimate air pollution exposure in epidemiologic studies. These models use air pollution measurements taken at a small set of locations and modeling based on geographical covariates for which data are available at all study participant locations. The process of LUR model development commonly includes a variable selection procedure. When LUR model predictions are used as explanatory variables in a model for a health outcome, measurement error can lead to bias of the regression coefficients and to inflation of their variance. In previous studies dealing with spatial predictions of air pollution, bias was shown to be small while most of the effect of measurement error was on the variance. In this study, we show that in realistic cases where LUR models are applied to health data, bias in health-effect estimates can be substantial. This bias depends on the number of air pollution measurement sites, the number of available predictors for model selection, and the amount of explainable variability in the true exposure. These results should be taken into account when interpreting health effects from studies that used LUR models.

air pollution; bias (epidemiology); measurement error; regression analysis

Abbreviations: IMT, intima media thickness; LUR, land-use regression; MSE, mean squared error.

Epidemiologic studies on the health effects of long-term exposure to air pollution often rely on a reduced sample of air pollution measurements in the study area and use modeling techniques to assign air pollution exposure to all study participants, usually at their residential addresses. One of such techniques is land-use regression (LUR) modeling, which involves fitting a linear regression model to the pollutant using as potential predictors a set of variables (e.g., traffic-related and topographical variables) on which data are available for any location in the study area through geographical information systems. The process of LUR model development commonly includes a variable selection procedure. LUR models are considered a simple and cost-efficient technique for assessment of air pollution exposure and are increasingly being used in epidemiologic studies (1–3).

Predictions of outdoor residential exposure from statistical models are inevitably affected by measurement error. Thus, when exposure predictions are included as explanatory variables in a regression model for a health outcome, results will suffer from the consequences of exposure measurement error, namely an increase in the variance of the estimated regression coefficients and, in certain cases, bias. Szpiro et al. (4) developed a general framework for measurement error in spatial prediction that applies to LUR models. They showed that measurement error in such contexts has a Berkson-like component and a classical-type component. The first component has the important property that it increases the variance of the estimated coefficients in the health model but does not bias them. On the contrary, the second component, besides increasing the variance, can also introduce bias. In Szpiro et al.'s applications of the measurement error framework, bias in the health-effect estimates was very small, suggesting that the main effect of exposure measurement error in spatial prediction is an increase in variance estimates (4, 5).

Recently, it has been shown that bias in health-effect estimates is of concern when exposure estimates are derived from

LUR models that were built using a relatively small number of air pollution measurement sites and a large number of potential predictors (6). In this report, we explore the relative magnitude of bias and variance inflation in health-effect estimates in typical LUR model settings, and we investigate the contributions of parameter estimation, variable selection, sample size, and LUR model R^2 to the results.

MATERIALS AND METHODS

Consider the case where we are interested in the effect β_X of the true exposure X on a health outcome Y,

$$Y = \beta_0 + \beta_X X + \varepsilon, \quad (1)$$

where ε are the residuals of the model. Since X is not measured in all residential locations of the N study participants, but at $n < N$ (possibly nonoverlapping) locations, we fit a LUR model to the n measurement locations. This model is based on a selected subset of r potential predictors and is used to predict $\hat{Z}$, the predicted exposure at the N residential locations. Finally, a model for the regression of Y on $\hat{Z}$ is fitted, from which $\hat{\beta}_Z$ is obtained.

We used simulations to study the effects of exposure measurement error on the estimation of β_Z. Simulation parameters were based on data from the Girona Heart Registry (REGICOR) Study (7). Thus, the exposure X represented nitrogen dioxide, a marker of traffic-related air pollution, and the response Y represented log-transformed carotid intima-media thickness (IMT), a marker of subclinical atherosclerosis.

Data were simulated as follows. The original residential locations of the $N = 2{,}622$ study participants and their predictor variables, computed using geographical information systems, were kept fixed across simulations. Then, we generated true nitrogen dioxide values for all participants based on a model with 5 of the potential predictor variables plus random error. The 5 chosen variables were the ones obtained in the final LUR model in the original paper and are described in Web Table 1 (available at http://aje.oxfordjournals.org/) (7). The random error variance was set to 2 different values that resulted in proportions of explainable variability in the true exposure of 50% and 75%. These proportions reflect the usual range in real studies. In addition, we simulated IMT data for all participants using model 1 (equation 1), with X scaled so that $\beta_X = 0.05$ represents a 5% increase in IMT associated with an increase in nitrogen dioxide exposure from the fifth percentile to the 95th percentile. The residual variance of model 1 was set to the value of the residual variance in the final IMT model in the original publication (7). The values of all parameters are given in Web Appendix 1.

From these data, independent samples of $n = 20$, $n = 40$, or $n = 80$ measurement locations were randomly drawn. These numbers reflect common values in real studies using LUR models (1). Sampling was stratified with equal probability according to categories based on quartiles of true nitrogen dioxide concentrations, to ensure that all samples contained the full range of nitrogen dioxide concentrations. Using data from the n locations only, a new LUR model (the estimated LUR model) was derived using a supervised forward variable selection algorithm described elsewhere (6). The process was performed separately using a set of $r = 20$ or $r = 100$ potential predictors (see Web Appendix 1 for more details). Both sets included the 5 predictor variables that generated the true nitrogen dioxide data. Since many geographical information systems variables can be computed for different buffer sizes, it is common in practice to have a large number of predictors. The estimated LUR model was used to predict $\hat{Z}$ for all locations, which was then entered into a model for Y to obtain $\hat{\beta}_Z$. In a separate analysis, variable selection was not performed when deriving the estimated LUR model. Instead, a model with the 5 variables that generated the true nitrogen dioxide values was fitted to the n locations.

The entire simulation process was repeated 1,000 times. The in-sample (based on the n locations used to build the model) and out-of-sample (based on $N - n$ remaining measures) R^2 values for the estimated LUR model were computed as $R^2 = \max(0, 1 - \sum_i (X_i - \hat{Z}_i)^2 / \sum_i (X_i - \bar{X})^2)$, where $\bar{X}$ indicates the mean of X. Bias was computed as the mean of $\hat{\beta}_Z$ over all simulations minus the true value of β_X, 0.05. The attenuation factor, a measure of bias in the multiplicative scale, was computed as the mean of $\hat{\beta}_Z$ over all simulations divided by the true value of β_X. The standard errors for $\hat{\beta}_Z$ obtained when fitting a regression model for Y on $\hat{Z}$ are hereafter referred to as naive standard errors. The standard deviation of the 1,000 values of $\hat{\beta}_Z$ was divided by the average of the naive standard errors to compute the standard error inflation factor. This quantity indicates the amount by which the naive standard errors need to be inflated to properly account for the real variation in $\hat{\beta}_Z$. The coverage of 95% confidence intervals based on naive standard errors was computed as the proportion of confidence intervals that included the true value of β_X. The mean squared error (MSE) was computed as the squared bias of $\hat{\beta}_Z$ plus its variance across simulations. The percentage of the MSE accounted for by the bias term was calculated as bias^2/MSE.

A similar simulation for the case of logistic regression is described in Web Appendix 2.

RESULTS

The first 2 columns of Table 1 illustrate how the estimated R^2 in the subsample of n measurement sites overestimates the prediction ability of the model in a new data set, which is better estimated by the out-of-sample R^2. This was especially the case for models developed with small n's. The out-of-sample R^2 increased with n but reached the true value used in the simulation (75%) only when no variable selection was performed—that is, the right variables were always used to fit the estimated LUR model. In that scenario, the estimated health model coefficient had approximately 10% attenuation when the estimated LUR model was based on 20 measurement sites. When 80 measurement sites were used, attenuation was only 1%. Variable selection in the development of the estimated LUR model introduced more attenuation of the health model coefficients. The strongest attenuation was found when the estimated LUR model could select among 100 predictor variables. In that case, the estimated health

Table 1. Results for the Properties of $\hat{\beta}_Z$, Obtained From 1,000 Simulations,[a] When the Proportion of Explained Variability in Nitrogen Dioxide Levels Was 75%

Scenario[b]	Exposure Model		Properties of $\hat{\beta}_Z$					MSE of $\hat{\beta}_Z$	
	In-Sample R^{2c}	Out-of-Sample R^{2c}	Coefficient[c]	Naive SE × 10^{4c}	Attenuation[c,d]	SE Inflation	Coverage[e], %	MSE × 10^5	% of MSE Due to Bias
Without variable selection[f]									
$n = 20$	0.83	0.67	0.045	8.0	0.91	7.4	17	5.5	37
$n = 40$	0.80	0.74	0.048	7.9	0.96	5.1	28	2.0	17
$n = 80$	0.79	0.76	0.049	7.9	0.99	3.4	40	0.8	7
With variable selection									
$r = 20$ variables									
$n = 20$	0.79	0.39	0.033	7.5	0.66	25.3	9	63.9	44
$n = 40$	0.75	0.55	0.040	7.8	0.81	20.9	18	35.5	26
$n = 80$	0.73	0.63	0.045	8.2	0.90	15.8	34	19.1	12
$r = 100$ variables									
$n = 20$	0.90	0.16	0.021	6.5	0.43	24.2	2	106.9	77
$n = 40$	0.80	0.42	0.034	7.5	0.68	22.5	9	54.4	48
$n = 80$	0.75	0.59	0.043	8.2	0.86	16.1	28	22.5	22

Abbreviations: LUR, land-use regression; MSE, mean squared error; SE, standard error.

[a] The simulation process is described in detail in Web Appendix 1.

[b] "*n*" refers to the number of measurement sites used to build the LUR model and *r* to the number of available predictors that could potentially be selected in the LUR model.

[c] Average over the 1,000 simulations.

[d] "Attenuation" refers to the mean of $\hat{\beta}_Z$ divided by the true value of β_X, 0.05.

[e] Coverage of the 95% confidence intervals based on naive standard errors.

[f] A model with the 5 variables that generated the "true nitrogen dioxide" values was fitted to the *n* locations.

model coefficient was less than half of the true value when the model was developed with 20 measurement sites. Attenuation was still strong-to-moderate in the remaining cases.

The naive standard errors of the health model coefficient severely underestimated the true variation in all scenarios, as reflected by the standard error inflation factor. Without variable selection, standard errors were too low by 3- to 7-fold, and values above 20-fold were reached when variable selection was performed. This led to very small coverage in all scenarios. For example, the 95% confidence interval based on naive standard errors of a study with $r = 20$ and $n = 80$ would include the true value only 34% of the time. The MSE, which combines bias and variance, was approximately halved when *n* was doubled. The percentage of MSE that was accounted for by bias ranged from 77% with 20 measurement sites and 100 predictor variables to 7% when no variable selection was performed and 80 measurement sites were available.

The same trends, but with worse overall performance, were observed when the model used to simulate true nitrogen dioxide values had a proportion of explainable variability of 50% (Table 2). In that case, the attenuation was still strong with 80 measurement sites and the percentage of MSE due to bias was larger in all scenarios. The results for logistic regression were very similar to those for linear regression in terms of attenuation, but naive standard errors showed less underestimation, leading to higher coverage values (Web Tables 2 and 3).

DISCUSSION

We used simulations based on real data and current practices to study the consequences of using LUR model predictions as an exposure variable in a model for a health outcome. Measurement error associated with LUR modeling had important impacts not only on the variance of the health-effect estimate but also on bias. The bias was in the form of attenuation towards the null hypothesis, and it was stronger in cases where variable selection was performed with a large number of predictor variables and a small number of measurement sites, which is the most common case in practice (1). In such settings, bias accounts for a large part of the MSE of the health-effect estimate.

The results presented in this paper fit into the measurement error framework proposed by Szpiro et al. (4). The measurement error induced by LUR models can be divided into a classical-type part and a Berkson-like part. The classical-type component is the one that can introduce bias in the health-

Table 2. Results for the Properties of $\hat{\beta}_Z$, Obtained From 1,000 Simulations[a], When the Proportion of Explained Variability in Nitrogen Dioxide Levels Was 50%

Scenario[b]	Exposure Model		Properties of $\hat{\beta}_Z$					MSE of $\hat{\beta}_Z$	
	In-Sample R^{2c}	Out-of-Sample R^{2c}	Coefficient[c]	Naive SE × 10^{4c}	Attenuation[c,d]	SE Inflation	Coverage[e], %	MSE × 10^5	% of MSE Due to Bias
Without variable selection[f]									
$n = 20$	0.61	0.28	0.037	10.9	0.74	8.6	7	25.1	65
$n = 40$	0.55	0.41	0.044	11.3	0.89	5.8	17	7.4	41
$n = 80$	0.52	0.46	0.048	11.5	0.95	4.1	30	2.9	22
With variable selection									
$r = 20$ variables									
$n = 20$	0.59	0.12	0.025	9.8	0.49	18.5	5	97.3	66
$n = 40$	0.52	0.26	0.035	10.5	0.70	16.2	11	52	44
$n = 80$	0.49	0.34	0.040	10.9	0.80	14.8	22	36.3	28
$r = 100$ variables									
$n = 20$	0.82	0.02	0.011	6.5	0.23	17.7	1	162.8	92
$n = 40$	0.63	0.11	0.023	8.9	0.46	16.5	2	93.3	77
$n = 80$	0.54	0.26	0.034	10.4	0.68	14.7	8	48.3	52

Abbreviations: LUR, land-use regression; MSE, mean squared error; SE, standard error.

[a] The simulation process is described in detail in Web Appendix 1.

[b] "*n*" refers to the number of measurement sites used to build the LUR model and *r* to the number of available predictors that could potentially be selected in the LUR model.

[c] Average over the 1,000 simulations.

[d] "Attenuation" refers to the mean of $\hat{\beta}_Z$ divided by the true value of β_X, 0.05.

[e] Coverage of the 95% confidence intervals based on naive standard errors.

[f] A model with the 5 variables that generated the "true nitrogen dioxide" values was fitted to the *n* locations.

effect estimate. This bias has been small in several reported examples, but all of them were based on exposure models derived with 100 measurements or more (4, 5). Our examples show that this bias can be substantial when the number of measurement sites is small, as is usually the case in studies using LUR modeling (1, 3). The classical-type measurement error arises because of the uncertainty in estimating the exposure model parameters. This is why bias was present even if the variables that generated the true nitrogen dioxide values were used in the estimated LUR model, but it was aggravated when model selection was performed. When the LUR model is fitted with a small number of measurement sites, it will typically overfit that set of points (8, 9). One can estimate the attenuation factor by comparing the true and predicted nitrogen dioxide values in a validation data set (6).

Apart from the effect on bias, using LUR model predictions as the exposure variable has important consequences for standard errors. Naive standard errors do not account for measurement error, and their use can lead to large underestimations of the true variation of the estimates, producing confidence intervals with low coverage. Several methods have been suggested for correcting for measurement error in similar contexts, but they are not adapted to settings with variable selection, one of the key factors leading to bias and variance inflation (4, 10–12). Such adaptation is not straightforward (13, 14). Besides, the choice of method for model selection will have an impact on the results. In fact, stepwise methods are known to lead to overfitting and other problems (6, 8, 15). Other strategies, such as those based on shrinkage (9, 15), should be evaluated as potential alternatives.

ACKNOWLEDGMENTS

Author affiliations: Centre for Research in Environmental Epidemiology (CREAL), Barcelona, Spain (Xavier Basagaña, Inmaculada Aguilera, David Agis, Maria Foraster); Hospital del Mar Medical Research Institute (IMIM), Barcelona, Spain (Xavier Basagaña, Inmaculada Aguilera, David Agis, Maria Foraster); CIBER Epidemiología y Salud Pública (CIBERESP), Barcelona, Spain (Xavier Basagaña, Inmaculada Aguilera, David Agis, Maria Foraster, Roberto Elosua); Department of Population Health, Centre de recherche du Centre Hospitalier de l'Université de Montréal (CHUM), Montreal, Quebec, Canada (Marcela Rivera); Department of Experimental Health Sciences, Universitat Pompeu Fabra, Barcelona, Spain (Maria Foraster); Epidemiology and Cardiovascular Genetics Group, IMIM, Barcelona, Spain (Jaume Marrugat, Roberto Elosua); Swiss Tropical and Public Health

Institute, Basel, Switzerland (Nino Künzli); and Department of Public Health, Faculty of Medicine, University of Basel, Basel, Switzerland (Nino Künzli).

This work was supported by the Fondo de Investigación Sanitaria (grant PI060258); the Fundacio La Marató de TV3 (grant 081632); and the Instituto de Salud Carlos III (grant FI-9/00989 to M.F.).

We thank Dr. Mark Nieuwenhuijsen for helpful comments.

Conflict of interest: none declared.

REFERENCES

1. Hoek G, Beelen R, de Hoogh K, et al. A review of land-use regression models to assess spatial variation of outdoor air pollution. *Atmos Environ*. 2008;42(33):7561–7578.
2. Jerrett M, Arain A, Kanaroglou P, et al. A review and evaluation of intraurban air pollution exposure models. *J Expo Anal Environ Epidemiol*. 2005;15(2):185–204.
3. Ryan PH, LeMasters GK. A review of land-use regression models for characterizing intraurban air pollution exposure. *Inhal Toxicol*. 2007;19(suppl 1):127–133.
4. Szpiro AA, Sheppard L, Lumley T. Efficient measurement error correction with spatially misaligned data. *Biostatistics*. 2011;12(4):610–623.
5. Szpiro AA, Paciorek CJ, Sheppard L. Does more accurate exposure prediction necessarily improve health effect estimates? *Epidemiology*. 2011;22(5):680–685.
6. Basagaña X, Rivera M, Aguilera I, et al. Effect of the number of measurement sites on land use regression models in estimating local air pollution. *Atmos Environ*. 2012;54:634–642.
7. Rivera M, Basagaña X, Aguilera I, et al. Long-term exposure to traffic related air pollution and subclinical atherosclerosis: the REGICOR study. *Environ Health Perspect*. 2013; 121(2):223–230.
8. Harrell FE. *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis*. New York, NY: Springer Publishing Company; 2001.
9. Harrell FE Jr, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med*. 1996;15(4):361–387.
10. Gryparis A, Paciorek CJ, Zeka A, et al. Measurement error caused by spatial misalignment in environmental epidemiology. *Biostatistics*. 2009;10(2):258–274.
11. Lee D, Shaddick G. Spatial modeling of air pollution in studies of its short-term health effects. *Biometrics*. 2010; 66(4):1238–1246.
12. Madsen L, Ruppert D, Altman NS. Regression with spatially misaligned data. *Environmetrics*. 2008;19(5): 453–467.
13. Breiman L. Little bootstrap and other methods for dimensionality selection in regression: X-fixed prediction error. *J Am Stat Assoc*. 1992;87(419):738–754.
14. Greenland S. Invited commentary: variable selection versus shrinkage in the control of multiple confounders. *Am J Epidemiol*. 2008;167(9):523–529.
15. Hastie T, Tibshirani R, Friedman JH. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York, NY: Springer Publishing Company; 2001.