Specification Errors, Measurement Errors, Confounding

Kerby Shedden

Department of Statistics, University of Michigan

October 20, 2021

An unobserved covariate

Suppose we have a data generating model of the form

$$y = \alpha + \beta x + \gamma z + \epsilon.$$

The usual conditions $E[\epsilon|\mathbf{x} = x, \mathbf{z} = z] = 0$ and $\mathrm{var}\,[\epsilon|\mathbf{x} = x, \mathbf{z} = z] = \sigma^2$ hold.

The covariate x is observed, but z is not observable.

If we regress y on x, the model we are fitting differs from the data generating model. What are the implications of this?

Does the fitted regression model $\hat{y} = \hat{\alpha} + \hat{\beta}x$ estimate $E[y|\mathbf{x} = x]$, and does the MSE $\hat{\sigma}^2$ estimate $\mathrm{var}[y|\mathbf{x} = x]$?

An unobserved independent covariate

The simplest case is where x and z are independent (and for simplicity $E[z] = 0$). The slope estimate $\hat{\beta}$ has the form

$$\begin{aligned}
\hat{\beta} &= \sum_i y_i(x_i - \bar{x}) / \sum_i (x_i - \bar{x})^2 \\
&= \sum_i (\alpha + \beta x_i + \gamma z_i + \epsilon_i)(x_i - \bar{x}) / \sum_i (x_i - \bar{x})^2 \\
&= \beta + \gamma \sum_i z_i(x_i - \bar{x}) / \sum_i (x_i - \bar{x})^2 + \sum_i \epsilon_i(x_i - \bar{x}) / \sum_i (x_i - \bar{x})^2
\end{aligned}$$

By the double expectation theorem,

$$E[\epsilon|\mathbf{x} = x] = E_{z|x} E[\epsilon|x, z] = 0$$

and since z and x are independent

$$E[\sum_i z_i(x_i - \bar{x})|x] = \sum_i (x_i - \bar{x}) E[z_i|x] = E[z] \times \sum_i (x_i - \bar{x}) = 0.$$

An unobserved independent covariate

Therefore $\hat{\beta}$ remains unbiased if there is an unmeasured covariate z that is independent of x. Specifically, $E[\hat{\beta}|X] = \beta$.

What about $\hat{\sigma}^2$? What does it estimate in this case?

An unobserved independent covariate

The residuals are

$$(I - P)y = (I - P)(\gamma z + \epsilon)$$

So the residual sum of squares is

$$y'(I - P)y = \gamma^2 z'(I - P)z + \epsilon'(I - P)\epsilon + 2\gamma z'(I - P)\epsilon.$$

The expected value is therefore

$$\begin{aligned} E[y'(I - P)y|x] &= \gamma^2 \mathrm{var}[z]\mathrm{rank}(I - P) + \sigma^2 \mathrm{rank}(I - P) \\ &= (\gamma^2 \mathrm{var}[z] + \sigma^2)(n - 2). \end{aligned}$$

Hence the $\hat{\sigma}^2$ has expected value $\gamma^2 \mathrm{var}(z) + \sigma^2$.

An unobserved independent covariate

Are our inferences correct?

We can set $\tilde{\epsilon} = \gamma z + \epsilon$ as being the error term of the model. Since

$$E[\tilde{\epsilon}|X = x] = 0 \qquad \text{cov}[\tilde{\epsilon}|X = x] = (\gamma^2 \text{var}[Z] + \sigma^2)I \propto I,$$

all the results about estimation of β in a correctly-specified model hold in this setting.

In general, we may wish to view any unobserved covariate as simply being another source of error, like ϵ. But we will see next that this cannot be done if z is dependent with x.

Confounding

As above, continue to take the data generating model to be

$$y = \alpha + \beta x + \gamma z + \epsilon,$$

but now suppose that x and z are correlated.

As before, z is not observed so our analysis will be based on y and x.

A variable such as z that is associated with both the dependent and independent variables in a regression model is called a confounder.

In this outcome, we often call y the outcome or response, and x is the exposure or treatment.

Confounding

Suppose x and z are standardized, and $\mathrm{cor}[x, z] = r$. Further suppose that $E[z|x] = rx$.

Due to the linearity of $E[y|x, z]$:

- If x increases by one unit and z remains fixed, the expected response increases by β units.
- If z increases by one unit and x remains fixed, the expected response increases by γ units.

However, if we select a pair of cases with x values differing by one unit at random (without controlling z), their z values will differ on average by r units. Therefore the expected responses $E[y]$ for these two cases differ by $\beta + r\gamma$ units.

Known confounders

There is a popular informal "typology" of confounders:

- Known and measured confounders ("known knowns")
- Known and unmeasured confounders ("known unknowns")
- Unknown and unmeasured confounders ("unknown unknowns")

Known and measured confounders

Suppose we are mainly interested in the relationship between a particular variable x and an outcome y. A measured confounder is a variable z that can be measured and included in a regression model along with x. A measured confounder generally does not pose a problem for estimating the "effect" of x, unless it is highly collinear with x.

Example: Suppose we are studying the health effects of second-hand smoke exposure (x). We measure the health outcome (y) directly. Subjects who smoke (z) are at risk for many of the same bad outcomes that may be associated with second-hand smoke exposure. Thus, it would be very important to determine which subjects smoke, and include that information as a covariate (a measured confounder) in a regression model used to assess the effects of second-hand smoke exposure.

Known and measured confounders

Caution: Just because a confounder is known and measured does not mean that simply including it as a main effect in the regression is sufficient to account for its role. That is, the working model $E[y|x,z] = \alpha + \beta x + \gamma z$ is a simple additive model that only serves as a starting point for "controlling" for the role of z. Perhaps the actual mean structure is $E[y|x,z] = x + z^2$ or $E[y|x,z] = xz + z^2/(1+x^2)$.

Known but unmeasured confounders

A known but unmeasured confounder is a variable that we know about, and for which we may have some knowledge of its distribution, but it is not measured in our particular data set.

For example, we may know that certain occupations (like working in certain types of factories) may produce risks similar to the risks of exposure to second-hand smoke. If occupation data is not collected in a particular study, this is an unmeasured confounder.

Since we do not have data for unmeasured confounders, their omission may produce bias in the estimated effects for variables of interest. If we have some understanding of how a certain unmeasured confounder operates, we may be able to use a sensitivity analysis to get a rough idea of how much bias is present.

Unknown confounders

An unknown confounder is a variable that affects the outcome of interest, but is unknown to us. An unknown confounder is necessarily unmeasured.

For example, there may be unknown genetic or environmental factors that are associated with both second hand smoke exposure (x) and the outcome (y).

Randomizaton and confounding

Unknown confounders and unmeasured confounders place major limits on our ability to interpret regression models causally or mechanistically.

Randomization: One way to substantially reduce the risk of confounding is to randomly assign the values of x. In this case, there can be no systematic association between x and z (for any z), and in large enough samples the actual (sample-level) association between x and z will be very low, so very little confounding is possible.

Randomization is a form of intervention or manipulation, and can only be done in specific situations where it is possible to assign the values of x, rather than observe them.

Randomizaton and confounding

In small samples, randomization can only gaurantee approximate orthogonality against unmeasured confounders.

The average result of many randomized studies is unbiased, but individual randomized studies may be biased if by chance there are imbalances, or chance associations between x and a confounder z.

For example if we have studies with $n = 8$ people, always consisting of four females and four males, and we randomly select four people to have $x = 1$ and four people to have $x = 0$, then a particular study could easily be severely unbalanced, e.g. all of the subjecs with $x = 1$ might be female.

If we have a known confounder such as sex, then we can do stratified randomization, i.e. randomly assign two females to treatment and two females to control, and similarly for males.

Confounding in linear models

For simplicity, suppose that z has mean 0 and variance 1, and we use least squares to fit a working model

$$\hat{y} = \hat{\alpha} + \hat{\beta} x$$

We can work out the limiting value of the slope estimate as follows.

$$\begin{aligned} \hat{\beta} &= \frac{\sum_i y_i (x_i - \bar{x})}{\sum_i (x_i - \bar{x})^2} \\ &= \frac{\sum_i (\alpha + \beta x_i + \gamma z_i + \epsilon_i)(x_i - \bar{x})/n}{\sum_i (x_i - \bar{x})^2/n} \\ &\to \beta + \gamma r. \end{aligned}$$

Note that if either $\gamma = 0$ (z is independent of y given x) or if $r = 0$ (z is uncorrelated with x), then β is estimated correctly.

Marginalization

For any population model defined by the first and second moments $E[y|x,z]$ and $\mathrm{var}[y|x,z]$, we can marginalize the model as follows:

$$E[y|x] = E_{z|x}E[y|x,z]$$

$$\mathrm{var}[y|x] = E_{z|x}\mathrm{var}[y|x,z] + \mathrm{var}_{z|x}E[y|x,z].$$

This marginalization can be done to any model, but what do we get if we marginalize the additive model for which $E[y|x,z] = \alpha + \beta x + \gamma z$ and $\mathrm{var}[y|x,z] = \sigma^2$?

Further, if we use least squares to model data $\{(y_i, x_i)\}$, are the results consistent for the marginalizations of the population model?

Marginalization

For the basic additive model,

$$\begin{aligned} E[y|x] &= E[E[y|x,z]|x] \\ &= E[\alpha + \beta x + \gamma z|x] \\ &= \alpha + \beta x + \gamma E[z|x]. \end{aligned}$$

$$\begin{aligned} \mathrm{var}[y|x] &= E_{z|x}\mathrm{var}[y|x,z] + \mathrm{var}_{z|x}E[y|x,z] \\ &= \sigma^2 + \mathrm{var}_{z|x}[\alpha + \beta x + \gamma z] \\ &= \sigma^2 + \gamma^2 \mathrm{var}[z|x]. \end{aligned}$$

Note that the marginalized linear model may be nonlinear in x. Also, while y is homoscedastic given x and z, it may be heteroscedastic when we only condition on x.

Confounding and mean structures

Suppose we regress y on x, ignoring z. Since

$$\hat{\beta} \to \beta + \gamma r,$$

and it is easy to show that $\hat{\alpha} \to \alpha$, the fitted model is approximately

$$\hat{y} \approx \alpha + \beta x + \gamma r x = \alpha + (\beta + \gamma r)x.$$

How does the fitted model relate to the marginal model $E[y|x]$? Since

$$E[y|x] = \alpha + \beta x + \gamma E[z|x],$$

the fitted regression model agrees with $E[y|x]$ as long as

$$E[z|x] = rx.$$

Confounding and variance structures

Turning now to the variance structure of the fitted model, the limiting value of $\hat{\sigma}^2$ is

$$\begin{aligned}
\hat{\sigma}^2 &= \sum_i (y_i - \hat{\alpha} - \hat{\beta} x_i)^2/(n-2) \\
&\approx \sum_i (\gamma z_i + \epsilon_i - \gamma r x_i)^2/n \\
&\to \sigma^2 + \gamma^2(1-r^2).
\end{aligned}$$

Ideally this should estimate the marginal variance $\mathrm{var}[y|x]$.

Confounding and variance structures

By the law of total variation,

$$\mathrm{var}[y|x] = \sigma^2 + \gamma^2 \mathrm{var}[z|x].$$

Thus for $\hat{\sigma}^2$ (obtained from regressing y on x while ignoring z) to estimate $\mathrm{var}[y|x]$ we need

$$\mathrm{var}[z|x] = 1 - r^2.$$

The Gaussian case

Suppose

$$y = \begin{pmatrix} a \\ b \end{pmatrix}$$

is a Gaussian random vector, where $y \in \mathcal{R}^n$, $a \in \mathcal{R}^q$, and $b \in \mathcal{R}^{n-q}$.

Let $\mu = Ey$ and $\Sigma = \mathrm{cov}[y]$. We can partition μ and Σ as

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \qquad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix},$$

where $\mu_1 \in \mathcal{R}^q$, $\mu_2 \in \mathcal{R}^{n-q}$, $\Sigma_{11} \in \mathcal{R}^{q \times q}$, $\Sigma_{12} \in \mathcal{R}^{q \times n-q}$, $\Sigma_{22} \in \mathcal{R}^{n-q \times n-q}$, and $\Sigma_{21} = \Sigma_{12}'$.

The Gaussian case

It is a fact that $a|b$ is Gaussian with mean

$$E[a|b] = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(b - \mu_2)$$

and covariance matrix

$$\mathrm{cov}[a|b] = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}.$$

The Gaussian Case

Now we apply these results to our model, taking x and z to be jointly Gaussian.

The mean vector and covariance matrix are

$$E\begin{bmatrix} z \\ x \end{bmatrix} = 0 \qquad \operatorname{cov}\begin{bmatrix} z \\ x \end{bmatrix} = \begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix}.$$

so we get

$$E[z|x] = rx \qquad \operatorname{cov}[z|x] = 1 - r^2.$$

These are exactly the conditions stated earlier that guarantee the fitted mean model converges to the marginal regression function $E[y|x]$, and the fitted variance model converges to to the marginal regression variance $\operatorname{var}[y|x]$.

Consequences of confounding

How does the presence of unmeasured confounders affect our ability to interpret regression models?

Population average covariate effect

Suppose we specify a value x_* in the covariate space and randomly select two subjects i and j having x values $x_i = x_* + 1$ and $x_j = x_*$. The inter-individual difference is

$$y_i - y_j = \beta + \gamma(z_i - z_j) + \epsilon_i - \epsilon_j,$$

which has a mean value (marginal effect) of

$$E[y_i - y_j | x_i = x_* + 1, x_j = x_*] = \beta + \gamma(E[z|x = x_* + 1] - E[z|x = x_*]),$$

which agrees with what would be obtained by least squares analysis as long as $E[z|x] = rx$.

Population average covariate effect

The variance of $y_i - y_j$ is

$$2\sigma^2 + 2\gamma^2 \mathrm{var}[z|x],$$

which also agrees with the results of least squares analysis as long as $\mathrm{var}[z|x] = 1 - r^2$.

Individual treatment effect

Now suppose we match two subjects i and j having x values differing by one unit, and who also having the same values of z.

This is what one expect to see as the pre-treatment and post-treatment measurements following a treatment that changes an individual's x value by one unit, if the treatment does not affect z (the within-subject treatment effect).

Individual treatment effect

The mean difference (individual treatment effect) is

$$E[y_i - y_j | x_i = x_* + 1, x_j = x_*, z_i = z_j] = \beta$$

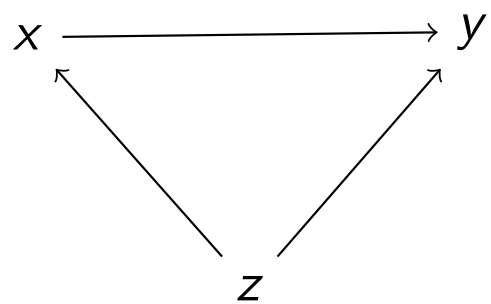
and the variance is

$$\mathrm{var}[y_i - y_j | x_i = x_* + 1, x_j = x_*, z_i = z_j] = 2\sigma^2.$$

These do not in general agree with the estimates obtained by using least squares to analyze the observable data for x and y. Depending on the sign of $\gamma\theta$, we may either overstate or understate the individual treatment effect β, and the population variance of the treatment effect will always be overstated.

Types of covariates

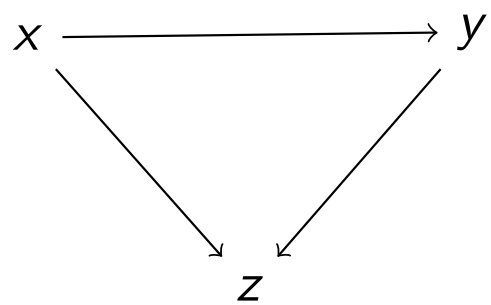
Expressed as a causal diagram, a confounder z relates to an exposure x and an outcome y as follows:



In addition to confounders, there are many other ways that a variable z can impact our ability to understand the relationship between variables of primary interest x and y.

Types of covariates

If we reverse the directionality between z and x, y, then z is no longer a confounder and instead becomes a collider.



If z is a confounder, then you must somehow control for z to obtain an undistorted understanding of the relationship between x and y. If z is a collider the opposite is true – controlling for z induces distortion in the relationship between x and y.

Types of covariates

A precision variable is a variable that explains some of the variation in y that is unrelated to x.

Including or excluding a precision variable in an analysis does not impact the bias, but can impact precision. In most cases, including a precision variable increases the precision with which the relationship between x and y is estimated

Types of covariates

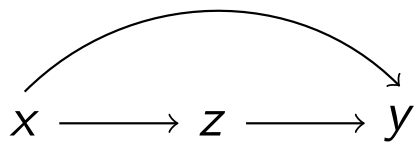
A mediator is a variable that lies on the causal pathway between an exposure x and an outcome y. In the diagram below, z is a mediator.

$$x \longrightarrow z \longrightarrow y$$

Controlling for a mediator will usually reduce or eliminate the apparent relationship between x and y, and doing so gives insight into the underlying mechanism behind the relationship between x and y.

Types of covariates

More generally, an exposure can have direct effects on an outcome, and indirect or mediated effects on an outcome, carried through a mediator z.



Types of covariates

A moderator or effect modifier is a variable that explains heteropgeneity in the relationship between x and y.

An interaction can be seen as an effect modifier. If $E[y|x, z] = x + zx = (1 + z)x$, we can interpret the slope of y on x as being different based on the value of z.

Finally note that in many cases we cannot be sure if a variable is a confounder, collider, mediator, or moderator, and many variables can occupy multiple of these roles at the same time.

Measurement error for linear models

Suppose the data generating model is

$$y = z\beta + \epsilon,$$

with the usual linear model assumptions, but we do not observe z.

Rather, we observe

$$x = z + \tau,$$

where τ is a random vector of covariate measurement errors with $E[\tau] = 0$. Assuming $x_1 = 1$ is the intercept, it is natural to set the first column of τ equal to zero.

This is called an errors in variables model, or a measurement error model.

Measurement error for linear models

When covariates are measured with error, least squares point estimates may be biased and inferences may be incorrect.

Intuitively it seems that slope estimates should be "attenuated" (biased toward zero). The reasoning is that as the measurement error grows very large, the observed covariate x becomes equivalent to noise, so the slope estimate should go to zero.

Measurement error for linear models

Let X and Z now represent the $n \times p+1$ observed and ideal design matrices and let T denote the $n \times p+1$ matrix of measurement errors. The least squares estimate of the model coefficients is

$$\begin{aligned}\hat{\beta} &= (X'X)^{-1}X'y \\ &= (Z'Z + Z'T + T'Z + T'\tau)^{-1}(Z'y + T'y) \\ &= (Z'Z/n + Z'T/n + T'Z/n + T'T/n)^{-1}(Z'y/n + T'Z\beta/n + T'\epsilon/n).\end{aligned}$$

We will make the simplifying assumption that the covariate measurement error is uncorrelated with the covariate levels, so

$$Z'T/n \to 0,$$

and that the covariate measurement error τ and observation error ϵ are uncorrelated, so

$$\tau'\epsilon/n \to 0.$$

Measurement error for linear models

Under these circumstances,

$$\hat{\beta} \approx (Z'Z/n + T'T/n)^{-1} Z'y/n.$$

Let M_z be the limiting value of $Z'Z/n$, and let $M_\tau = E[\tau\tau']$ be the limiting value of $T'T/n$. Thus the limit of $\hat{\beta}$ is

$$\begin{aligned}(M_z + M_\tau)^{-1} Z'y/n &= (I + M_z^{-1} M_\tau)^{-1} M_z^{-1} \mathbf{Z}'y/n \\ &\to (I + M_z^{-1} M_\tau)^{-1} \beta \\ &\equiv \beta_0.\end{aligned}$$

and hence the limiting bias is

$$\beta_0 - \beta = ((I + M_z^{-1} M_\tau)^{-1} - I)\beta.$$

Measurement error for linear models

What can we say about the bias?

Note that the matrix $M_z^{-1} M_\tau$ has non-negative eigenvalues, since it shares its eigenvalues with the positive semi-definite matrix

$$M_z^{-1/2} M_\tau M_z^{-T/2}.$$

It follows that all eigenvalues of $I + M_z^{-1} M_\tau$ are greater than or equal to 1, so all eigenvalues of $(I + M_z^{-1} M_\tau)^{-1}$ are less than or equal to 1.

This means that $(I + M_z^{-1} M_\tau)^{-1}$ is a contraction, so $\|\beta_0\| \leq \|\beta\|$.

Therefore the sum of squares of fitted slopes is smaller on average than the sum of squares of actual slopes, due to measurement error.

Types of measurement error

The "classical" measurement error model

$$x = z + \tau,$$

where z is the true value and x is the observed value, is the one most commonly considered.

Alternatively, in the case of an experiment it may make more sense to use the Berkson error model:

$$z = x + \tau.$$

For example, suppose we aim to study a chemical reaction when a given concentration x of substrate is present. However, due to our inability to completely control the process, the actual concentration of substrate z differs randomly from x, by an unknown amount τ.

Types of measurement error

You cannot simply rearrange $z = x + \tau$ to $x = z - \tau$ and claim that the two situations are equivalent.

In the first case, τ is independent of z but dependent with x. In the second case, τ is independent of x but dependent with z.