

Specification Errors, Measurement Errors, Confounding

Kerby Shedden

Department of Statistics, University of Michigan

October 21, 2019

An unobserved covariate

Suppose we have a data generating model of the form

$$Y = \alpha + \beta X + \gamma Z + \epsilon.$$

The usual conditions $E[\epsilon|X = x, Z = z] = 0$ and $\text{var}[\epsilon|X = x, Z = z] = \sigma^2$ 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 $E[Y|X = x]$, and does the MSE $\hat{\sigma}^2$ estimate $\text{var}[Y|X = x]$?

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,

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

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 .

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

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

Hence the MSE of $\hat{\sigma}^2$ has expected value $\gamma^2 \text{var}(Z) + \sigma^2$.

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

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

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**.

Confounding

Suppose X and Z are standardized, and $\text{cor}[X, Z] = r$. 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 these expected responses for these two cases differ by $\beta + r\gamma$ units.

Known 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 .

For example, suppose we are studying the health effects of second-hand smoke exposure. We measure the health outcome (Y) directly. Subjects who smoke are of course 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.

Unmeasured confounders

An **unmeasured confounder** is a variable that we know about, and for which we may have some knowledge of its statistical properties, but is not measured in a 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.

For example, there may be genetic factors that produce similar health effects as second-hand smoke, but we have no knowledge of which specific genes are involved.

Randomization 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 , and in large enough samples the actual (sample-level) association between X and Z will be very low, so very little confounding is possible.

In small samples, randomization cannot guarantee approximate orthogonality against unmeasured confounders.

Confounding

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.

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

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.

Confounding

Since

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

and it is easy to show that $\hat{\alpha} \rightarrow \alpha$, 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** $E[Y|X]$? It is easy to get

$$E[Y|X = x] = \alpha + \beta x + \gamma E[Z|X = x],$$

so the fitted regression model agrees with $E[Y|X = x]$ as long as

$$E[Z|X = x] = rx.$$

Confounding

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

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

Ideally this should estimate the variance function of the **marginal model** $\text{var}[Y|X]$.

Confounding

By the law of total variation,

$$\begin{aligned}\text{var}[Y|X = x] &= E_{Z|X} \text{var}[Y|X = x, Z] + \text{var}_{Z|X=x} E[Y|X = x, Z] \\ &= \sigma^2 + \text{var}_{Z|X=x}(\alpha + \beta x + \gamma Z) \\ &= \sigma^2 + \gamma^2 \text{var}[Z|X = x].\end{aligned}$$

So for $\hat{\sigma}^2$ to estimate $\text{var}[Y|X = x]$ we need

$$\text{var}[Z|X = x] = 1 - r^2.$$

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 $\Sigma = \text{cov}[Y]$. 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

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

and covariance matrix

$$\text{cov}[A|B = b] = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}.$$

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

$$E \begin{bmatrix} Z \\ X \end{bmatrix} = 0 \qquad \text{cov} \begin{bmatrix} Z \\ X \end{bmatrix} = \begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix}.$$

so we get

$$E[Z|X = x] = rx \qquad \text{cov}[Z|X = x] = 1 - r^2.$$

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 $\text{var}[Y|X]$.

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 $E[Z | X = x] = rx$.

Population average covariate effect

The variance of $y_i - y_j$ is

$$2\sigma^2 + 2\gamma^2 \text{var}[Z|X],$$

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

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

$$\text{var}[Y_i - Y_j | X_i = x_* + 1, X_j = x_*, Z_i = z_j] = 2\sigma^2.$$

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.

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 \mathbf{X} and \mathbf{Z} now represent the $n \times p + 1$ observed and ideal design matrices. The least squares estimate of the model coefficients is

$$\begin{aligned}\hat{\beta} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \\ &= (\mathbf{Z}'\mathbf{Z} + \mathbf{Z}'\boldsymbol{\tau} + \boldsymbol{\tau}'\mathbf{Z} + \boldsymbol{\tau}'\boldsymbol{\tau})^{-1}(\mathbf{Z}'\mathbf{Y} + \boldsymbol{\tau}'\mathbf{Y}) \\ &= (\mathbf{Z}'\mathbf{Z}/n + \mathbf{Z}'\boldsymbol{\tau}/n + \boldsymbol{\tau}'\mathbf{Z}/n + \boldsymbol{\tau}'\boldsymbol{\tau}/n)^{-1}(\mathbf{Z}'\mathbf{Y}/n + \boldsymbol{\tau}'\mathbf{Z}\beta/n + \boldsymbol{\tau}'\epsilon/n).\end{aligned}$$

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

$$\mathbf{Z}'\boldsymbol{\tau}/n \rightarrow 0,$$

and that the covariate measurement error $\boldsymbol{\tau}$ and observation error ϵ are uncorrelated, so

$$\boldsymbol{\tau}'\epsilon/n \rightarrow 0.$$

Measurement error for linear models

Under these circumstances,

$$\hat{\beta} \rightarrow (\mathbf{Z}'\mathbf{Z}/n + \tau'\tau/n)^{-1}\mathbf{Z}'\mathbf{Y}/n.$$

Let M_z be the limiting value of $\mathbf{Z}'\mathbf{Z}/n$, and let M_τ be the limiting value of $\tau'\tau/n$. Thus the limit of $\hat{\beta}$ is

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

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 .

SIMEX

SIMEX (simulation-extrapolation) is a relatively straightforward way to adjust for the effects of measurement error – if the variances and covariances among the measurement errors can be considered known.

Regress Y on $X + \lambda E$, where E is simulated noise having the same variance as the assumed measurement error. Denote the coefficient vector of this fit as $\hat{\beta}_\lambda$.

Repeat this for several values of $\lambda \geq 0$, leading to a set of $\hat{\beta}_\lambda$ vectors.

Ideally, $\hat{\beta}_{-1}$ would approximate the coefficient estimates under no measurement error.

By fitting a line or smooth curve to the $\hat{\beta}_\lambda$ values (separately for each component of β), it becomes possible to extrapolate back to $\hat{\beta}_{-1}$.