

# **Model mis-specification and confounding**

Suppose we have a data generating model of the form

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

The usual assumptions  $E(\epsilon|X, Z) = 0$  and  $\text{var}(\epsilon|X, Z) = \sigma^2$  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 reflect  $E(Y|X)$  and  $\text{var}(Y|X)$ ?

The simplest case is where  $X$  and  $Z$  are independent. 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,

$$E(\epsilon|X) = E_{Z|X}E(\epsilon|X, Z) = 0$$

and since  $Z$  and  $X$  are independent

$$\sum_i E(Z_i(X_i - \bar{X})|X) = \sum_i (X_i - \bar{X})E(Z_i|X) = EZ \sum_i (X_i - \bar{X}) = 0.$$

Therefore  $\hat{\beta}$  remains unbiased if there is an unmeasured covariate  $Z$  that is independent of  $X$ .

What about  $\hat{\sigma}^2$ ?

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

$$\begin{aligned} EY'(I - P)Y &= \gamma^2 \text{var}(Z) \text{rank}(I - P) + \sigma^2 \text{rank}(I - P) \\ &= (\gamma^2 \text{var}(Z) + \sigma^2)(n - 2). \end{aligned}$$

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

Are our inferences correct?

Since

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

we can view  $\gamma Z + \epsilon$  as being the error term and all the results for correctly specified models hold.

## Confounding

Continue to assume the data generating model

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

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.”

## Interpretation of non-orthogonal covariate effects

Suppose  $X$  and  $Z$  are standardized, and  $\text{cor}(X, Z) = r$ .

If  $X$  increases by one unit and  $Z$  remains fixed, the expected response increases by  $\beta$  units.

If  $Z$  increases by one unit and  $X$  remains fixed, the expected response increases by  $\gamma$  units.

This interpretation is relevant if we can control the value of  $X$  or  $Z$  without changing the other value.

However, if we select from our sample a pair of individuals with  $X$  values differing by one unit, their  $Z$  values will tend to differ by  $r$  units. Therefore these two individuals will have expected responses differing by  $\beta + r\gamma$  units.

Suppose our interest lies mainly with  $X$ . A known confounder  $Z$  that can be measured and included in a regression model does not generally pose a problem unless it is highly collinear with  $X$ .

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

The best way to be sure that confounding is not an issue is to randomly assign the levels of  $X$ . In this case, there can be no association between  $X$  and  $Z$ .

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 independent of  $X$ ), then  $\beta$  is estimated correctly.

Since

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

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

$$\hat{Y} \approx \alpha + \beta X + \gamma r X.$$

How does the fitted model relate to  $E(Y|X)$ ? It is easy to get

$$E(Y|X) = \alpha + \beta X + \gamma E(Z|X),$$

so the fitted regression model agrees with  $E(Y|X)$  as long as

$$E(Z|X) = rX.$$

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  $\text{var}(Y|X)$ .

By the law of total variation,

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

So for  $\hat{\sigma}^2$  to estimate  $\text{var}(Y|X)$  we need

$$\text{var}(Z|X) = 1 - r^2.$$

## 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  $\mu$  and  $\Sigma$  as

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

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  $\Sigma_{21} = \Sigma'_{12}$ .

## Gaussian case (continued)

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

$$\text{cov}(A|B) = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}.$$

Now we apply these results to our model

$$X = \theta Z + \eta,$$

assuming  $Z$  and  $\eta$  to be jointly Gaussian.

The mean vector and covariance matrix are

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

so we get

$$E(Z|X) = rX \qquad \text{cov}(Z|X) = 1 - r^2.$$

These are exactly the conditions stated above that guarantee the fitted mean model converges to  $E(Y|X)$  and the fitted variance model converges to  $\text{var}(Y|X)$ .

## **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 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  $E(Z | X) = rX$ .

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  $\text{var}(Z|X) = 1 - r^2$ .

## 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 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 agree with the estimates obtained by using least squares to analyze the observable  $Y, X$  data. Depending on the sign of  $\gamma\theta$ , we may either overstate or understate the individual treatment effect  $\beta$ , and the population variance of the treatment effect will always be overstated.

## Errors in Variables 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  $\tau$  is a random vector of covariate measurement errors with  $E\tau = 0$ . Assuming  $X_1 = 1$  is the intercept, it is natural to set  $\tau_1 \equiv 0$ .

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

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.

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

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

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

$$Z'\tau/n \rightarrow 0,$$

and that the covariate measurement error  $\tau$  and observation error  $\epsilon$  are uncorrelated, so

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

Under these circumstances,

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

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

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

and hence the limiting bias is

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

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.

## 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  $\beta$ ), it becomes possible to extrapolate back to  $\hat{\beta}_{-1}$ .