

SENSITIVITY ANALYSIS

Jon Wakefield,

Departments of Statistics and Biostatistics, Box 357232, University of
Washington, Seattle, WA 98195–7232.

jonno@u.washington.edu

Keywords: Confounding; Measurement Error; Observational studies; Randomization; Selection bias.

Abstract

In an observational study there is always the possibility of bias due to unmeasured variables, measurement error in explanatory variables and/or selection bias; and randomized studies are not immune to at least some of these biases also. A sensitivity analysis provides a means of asking a series of “what if?” questions, in order to determine the robustness of the observed association to sources of bias.

Introduction

Consider an experiment in which varying dosage levels of a drug are randomly assigned to groups of individuals. If the randomization is successfully implemented, the groups of subjects will be balanced with respect to all variables other than the dosage levels of interest, at least on average (and if the groups are sufficiently large, effectively in practice also). The beauty of randomization is that the groups are balanced not only with respect to measured variables, but also with respect to unmeasured variables. In a non-randomized situation, although one may control for *observed* explanatory variables, one can never guarantee that observed associations are not due to unmeasured variables. Selection bias, in which the chances of observing a particular individual depend on the values of their responses and explanatory variables, is another potential source of bias. A further source of bias is due to measurement error in the explanatory variable(s) (in the randomized example this could correspond to inaccurate measurement of the dosage received, though it could also correspond to other explanatory variable). This problem is sometimes referred to as *errors-in-variables* and is discussed in detail in [1]. Many other types of sensitivity analysis are possible (for

example, with respect to prior distributions in a Bayesian analysis) but we consider confounding, measurement error and selection bias only. For more discussion of these topics in an epidemiological context, see Chapter 19 of [4].

A general approach to sensitivity analyses is to first write down a plausible model for the response in terms of accurately measured explanatory variables (some of which may be unobserved), and with respect to a particular selection model. One may then derive the induced form of the model in terms of observed variables and the selection mechanism assumed in the analysis. The parameters of the derived model can then be compared with the parameters of interest in the “true” model, to reveal the extent of bias. We follow this approach but note that it should only be pursued only when the sample size in the original study is large, so that sampling variability is negligible; references in the discussion consider more general situations.

In the following we assume that data are not available to control for bias. So in the next section we consider the potential effects of *unmeasured* confounding. In the errors-in-variables context we assume that we do observe “gold standard” data in which a subset of individuals provide an accurate measure of the explanatory variable, along with the inaccurate measure. Similarly with respect to selection bias we assume that the sampling probabilities for study individuals are unknown and cannot be controlled for (as can be done in matched case-control studies, see Chapter 16 of [4] for example), or that supplementary data on the selection probabilities of individuals are not available, as in two-phase methods (e.g., [6]); in both of these examples, the selection mechanism is known from the design (and would lead to bias if ignored since the analysis must respect the sampling scheme).

Sensitivity to unmeasured confounding

Let Y denote a univariate response and X a univariate explanatory variable, and suppose that we are interested in the association between Y and X , but Y also potentially depends on U , an unmeasured variable. The discussion in [2] provided an early and clear account of the sensitivity of an observed association to unmeasured confounding, in the context of lung cancer and smoking. For simplicity we assume that the “true” model is linear and given by

$$E[Y|X, U] = \alpha^* + X\beta^* + U\gamma^*.$$

Further assume that the linear association between U and X is $E[X|U] = a + bU$. Roughly speaking, a variable U is a *confounder* if it is associated with both the response, Y , and the explanatory variable, X , but is not be caused by Y or on the causal pathway between X and Y . For a more precise definition of confounding, and an extended discussion see Chapter 8 of [4]. We wish to derive the implied linear association between Y and X since these are the variables that are observed. We use iterated expectation to average over the unmeasured variable, given X :

$$\begin{aligned} E[Y|X] &= E_{U|X} \{E[Y|X, U]\} = E_{U|X} \{\alpha^* + X\beta^* + U\gamma^*\} \\ &= \alpha^* + X\beta^* + E[U|X]\gamma^* = \alpha^* + X\beta^* + (a + bX)\gamma^* = \alpha + X\beta, \end{aligned}$$

where $\alpha = \alpha^* + \gamma^*a$ and, of more interest,

$$\beta = \beta^* + \gamma^*b. \tag{1}$$

Here the “true” association parameter, β^* , that we would like to estimate is represented with a * superscript, while the association parameter that we can estimate, β , does not have a superscript. Equation (1) shows that the bias $\beta - \beta^*$ is a function of the level of association between X and U (via the parameter b) and the association between Y and U (via γ^*). Equation (1) can be used to assess the effects of an unmeasured confounding using plausible values of b and γ^* , as we now demonstrate through a simple example.

Example: Consider a study in which we wish to estimate the association between the rate of oral cancer and alcohol intake in men over 60 years of age. Let Y represent the natural logarithm of the rate of oral cancer and suppose we have a two-level alcohol variable X with $X = 0$ corresponding to zero intake and $X = 1$ non-zero intake. A regression of Y on X gives an estimate $\hat{\beta} = 1.20$ so that the rate of oral cancer is $e^{1.20} = 3.32$ higher in the $X = 1$ population when compared to the $X = 0$ population.

The rate of oral cancer also increases with tobacco consumption (which we suppose is unmeasured in our study), however, and the latter is also positively associated with alcohol intake. We let $U = 0/1$ represent no tobacco/tobacco consumption. Since U is a binary variable $E[U|X] = \Pr(U = 1|X)$. Suppose that the probability of tobacco consumption is 0.05 and 0.45 in those with zero and non-zero alcohol consumption, respectively; that is

$\Pr(U = 1|X = 1) = 0.05$ and $\Pr(U = 1|X = 0) = 0.45$ so that $a = 0.05$ and $a + b = 0.45$ to give $b = 0.40$. Suppose further that the log rate of oral cancer increases by $\gamma^* = \log 2.0 = 0.693$ for those who use tobacco (in both alcohol groups). Under these circumstances, from (1), the true association is

$$\hat{\beta}^* = \hat{\beta} - \gamma^*b = 1.20 - 0.693 \times 0.40 = 0.92,$$

so that the increase in the rate associated with alcohol intake is reduced from 3.32 to $\exp(0.92) = 2.51$.

In a real application the sensitivity of the association would be explored with respect to a range of values of b and γ^* .

Sensitivity to measurement errors

In a similar way we may examine the potential effects of measurement errors in the regressor X . As an example consider a simple linear regression and suppose the true model is

$$Y = E[Y|X] + \epsilon^* = \alpha^* + \beta^*X + \epsilon^* \quad (2)$$

where $E[\epsilon^*] = 0$, $\text{var}(\epsilon^*) = \sigma_{\epsilon^*}^2$. Rather than measure X we measure a surrogate W where

$$W = X + \delta \quad (3)$$

with $E[\delta] = 0$, $\text{var}(\delta) = \sigma_{\delta}^2$, and $\text{cov}(\delta, \epsilon^*) = 0$. The least squares estimator of β^* in model (2), from a sample (X_i, Y_i) , $i = 1, \dots, n$, has the form

$$\hat{\beta}^* = \frac{\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2} = \frac{\text{cov}(X, Y)}{\text{var}(X)}. \quad (4)$$

In the measurement error situation we fit the model

$$Y = E[Y|W] + \epsilon = \alpha + \beta W + \epsilon \quad (5)$$

where $E[\epsilon] = 0$, $\text{var}(\epsilon) = \sigma_{\epsilon}^2$. The least squares estimator of β in model (5), from a sample (W_i, Y_i) , $i = 1, \dots, n$, has the form

$$\hat{\beta} = \frac{\frac{1}{n} \sum_{i=1}^n (W_i - \bar{W})(Y_i - \bar{Y})}{\frac{1}{n} \sum_{i=1}^n (W_i - \bar{W})^2} = \frac{\text{cov}(W, Y)}{\text{var}(W)}, \quad (6)$$

and to assess the extent of bias we need to compare $E[\hat{\beta}]$ with $\hat{\beta}^*$. From (6) we have

$$\begin{aligned}\hat{\beta} &= \frac{\frac{1}{n} \sum_{i=1}^n (X_i + \delta_i - \bar{X} - \bar{\delta})(Y_i - \bar{Y})}{\frac{1}{n} \sum_{i=1}^n (X_i + \delta_i - \bar{X} - \bar{\delta})^2} \\ &= \frac{\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y}) + \frac{1}{n} \sum_{i=1}^n (\delta_i - \bar{\delta})(Y_i - \bar{Y})}{\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 + \frac{2}{n} \sum_{i=1}^n (X_i - \bar{X})(\delta_i - \bar{\delta}) + \frac{1}{n} \sum_{i=1}^n (\delta_i - \bar{\delta})^2} \\ &= \frac{\text{cov}(X, Y) + \text{cov}(\delta, Y)}{\text{var}(X) + 2\text{cov}(X, \delta) + \text{var}(\delta)}.\end{aligned}$$

Under the assumptions of our model, $\text{cov}(\delta, Y) = 0$ and $\text{cov}(X, \delta) = 0$ and so, from (4)

$$E[\hat{\beta}] \approx \frac{\text{cov}(X, Y)}{\text{var}(X) + \text{var}(\delta)} = \frac{\text{cov}(X, Y)/\text{var}(X)}{1 + \text{var}(\delta)/\text{var}(X)} = \beta^* r$$

where the *attenuation factor*

$$r = \frac{\text{var}(X)}{\text{var}(X) + \sigma_\delta^2}$$

describes the amount of bias by which the estimate is attenuated toward zero. Note that with no measurement error ($\sigma_\delta^2 = 0$) $r = 1$ and no bias results, and also that the attenuation will be smaller in a well-designed study in which a large range of X is available. Hence to carry out a sensitivity analysis we can examine, for an observed estimate $\hat{\beta}$, the increase in the true coefficient for different values of σ_δ^2 via

$$\hat{\beta}^* = \frac{\hat{\beta}}{r}.$$

It is important to emphasize that the above derivation was based on a number of strong assumptions such as independence between errors in Y and in W , and constant variance for errors in both Y and W . Care is required in more complex situations, including those in which we have more than one explanatory variable. For example, if we regress Y on both X (which is measured without error) and a second explanatory variable measured with error, then we will see bias in our estimator of the coefficient associated with X , if there is a non-zero correlation between X and the second variable (see [1] for more details).

Example: Let Y represent systolic blood pressure (in mm Hg) and X sodium intake (in mmol/day) and suppose that a linear regression of Y on X produces an estimate of $\hat{\beta} = 0.1$ mm Hg so that an increase in daily sodium of 100 mmol/day is associated with an increase in blood pressure of 10 mm Hg. Suppose also that $\text{var}(X) = 4$ mm Hg. Table 1 shows the sensitivity of the coefficient associated with X , β^* , to different levels of measurement error; as expected the estimate increases with increasing measurement error.

| Measurement Error σ_δ^2 | Attenuation Factor r | True Estimate $\hat{\beta}^*$ |
|--|---------------------------|----------------------------------|
| 0 | 0 | 0.1 |
| 1 | 0.8 | 0.125 |
| 2 | 0.67 | 0.15 |
| 4 | 0.5 | 0.2 |

Table 1: The effect of measurement error when $\text{var}(X) = 4$.

Sensitivity to selection bias

This section concerns the assessment of the bias that is induced when the probability of observing the data of a particular individual depends on the data of that individual. We consider a slightly different scenario to those considered in the last two sections and assume we have a binary outcome variable, Y , and a binary exposure, X , and let $p_x^* = \Pr(Y = 1|X = x)$, $x = 0, 1$, be the “true” probability of a $Y = 1$ outcome given exposure x , $x = 0, 1$. We take as parameter of interest the odds ratio:

$$\text{OR}^* = \frac{\Pr(Y = 1|X = 1)/\Pr(Y = 0|X = 1)}{\Pr(Y = 1|X = 0)/\Pr(Y = 0|X = 0)} = \frac{p_1^*/(1 - p_1^*)}{p_0^*/(1 - p_0^*)}, \quad (7)$$

which is the ratio of the odds of a $Y = 1$ outcome given exposed ($X = 1$), to the odds of such an outcome given unexposed ($X = 0$).

We now consider the situation in which we do not have constant probabilities of responding (being observed) across the population of individuals under study and let $R = 0/1$ correspond to the event non-response/response, with response probabilities:

$$\Pr(R = 1|X = x, Y = y) = q_{xy},$$

for $x = 0, 1$, $y = 0, 1$; and we assume that we do not know these response rates. We do observe estimates of $p_x = \Pr(Y = 1|X = x, R = 1)$, the probability of a $Y = 1$ outcome given both values of x and *given* response. The estimate of the odds ratio for the *observed* responders is then given by:

$$\text{OR} = \frac{\Pr(Y = 1|X = 1, R = 1)/\Pr(Y = 0|X = 1, R = 1)}{\Pr(Y = 1|X = 0, R = 1)/\Pr(Y = 0|X = 0, R = 1)} = \frac{p_1/(1 - p_1)}{p_0/(1 - p_0)}. \quad (8)$$

To link the two odds ratios we use Bayes theorem on each of the terms in (8) to give:

$$\begin{aligned}
\text{OR} &= \frac{\frac{\Pr(R=1|X=1,Y=1)\Pr(Y=1|X=1)}{\Pr(R=1|X=1)} / \frac{\Pr(R=1|X=1,Y=0)\Pr(Y=0|X=1)}{\Pr(R=1|X=1)}}{\frac{\Pr(R=1|X=0,Y=1)\Pr(Y=1|X=0)}{\Pr(R=1|X=0)} / \frac{\Pr(R=1|X=0,Y=0)\Pr(Y=0|X=0)}{\Pr(R=1|X=0)}} \\
&= \frac{p_1^*/(1-p_1^*)}{p_0^*/(1-p_0^*)} \times \frac{q_{11}q_{00}}{q_{10}q_{01}} = \text{OR}^* \times s \tag{9}
\end{aligned}$$

where the selection factor s is determined by the probabilities of response in each of the exposure-outcome groups. It is of interest to examine situations in which $s = 1$ and there is no bias. One such situation is when $q_{xy} = u_x \times v_y$, $x = 0, 1; y = 0, 1$, so that there is “no multiplicative interaction” between exposure and outcome in the response model. Note that u_x and v_y are *not* the marginal response probabilities for, respectively, exposure and outcome.

Example: Consider a study carried out to examine the association between childhood asthma and maternal smoking. Let $Y = 0/1$ represent absence/presence of asthma in a child and $X = 0/1$ represent non-exposure/exposure to maternal smoking. Suppose a questionnaire is sent to parents to determine whether their child has asthma and whether the mother smokes. An odds ratio of $\widehat{\text{OR}}=2$ is observed from the data of the responders, indicating that the odds of asthma is doubled if the mother smokes.

To carry out a sensitivity analysis there are a number of ways to proceed. We write

$$s = \frac{\Pr(R = 1|X = 1, Y = 1)/\Pr(R = 1|X = 0, Y = 1)}{\Pr(R = 1|X = 1, Y = 0)/\Pr(R = 1|X = 0, Y = 0)} = \frac{q_{11}/q_{01}}{q_{10}/q_{00}}.$$

Suppose that amongst non-cases the response rate in the exposed group is q times that in the unexposed group (that is $q_{10}/q_{00} = q$), while amongst the cases the response rate in the exposed group is $0.8q$ times that in the unexposed group (i.e. $q_{11}/q_{01} = 0.8q$). In this scenario $s = 0.8$ and

$$\widehat{\text{OR}}^* = \frac{\widehat{\text{OR}}}{0.8} = \frac{2}{0.8} = 2.5,$$

and we have underestimation because exposed cases were under-represented in the original sample.

Discussion

In this article we have considered sensitivity analyses in a number of very simple scenarios. An extension would be to simultaneously consider the combined sensitivity to multiple

sources of bias. We have also considered the sensitivity of point estimates only, and have not considered hypothesis testing or interval estimation. A comprehensive treatment of observational studies and in particular the sensitivity to various forms of bias may be found in [3]. The above derivations can be extended to various different modeling scenarios, for example [5] examines sensitivity to unmeasured confounding in the context of Poisson regression in spatial epidemiology.

References

- [1] Carroll, R.J., Ruppert, D. and Stefanski, L.A. (1995). *Measurement in Nonlinear Models*, Chapman and Hall/CRC Press, London.
- [2] Cornfield, J., Haenszel, W., Hammond, E.C., Lillienfeld, A.M., Shimkin, M.B. and Wynder, E.L. (1959). Smoking and lung cancer: recent evidence and a discussion of some questions. *Journal of the National Cancer Institute*, **22**, 173–203.
- [3] Rosenbaum, P.R. (2002). *Observational Studies, Second Studies*, Springer-Verlag, New York.
- [4] Rothman, K.J. and Greenland, S. (1998). *Modern Epidemiology, Second Edition*, Lippincott-Raven, Philadelphia.
- [5] Wakefield, J.C. (2003). Sensitivity analyses for ecological regression. *Biometrics*, **59**, 9–17.
- [6] White, J.E. (1982). A two-stage design for the study of the relationship between a rare exposure and a rare disease. *American Journal of Epidemiology*, **115**, 119–128.