Summary. In a presentation of various methods for assessing the sensitivity of regression results to unmeasured confounding, Lin et al. (1998) use a conditional independence assumption to derive algebraic relationships between the true exposure effect and the apparent exposure effect in a reduced model which does not control for the unmeasured confounding variable. However, Hernán and Robins (1999) have noted that if the measured covariates and the unmeasured confounder both affect the exposure of interest then the principal conditional independence assumption which is used to derive these algebraic relationships cannot hold. One particular result of Lin et al. does not rely on the conditional independence assumption but only on assumptions concerning additivity. It can be shown that this assumption is satisfied for an entire family of distributions even if both the measured covariates and the unmeasured confounder affect the exposure of interest. These considerations clarify the appropriate context in which relevant sensitivity analysis techniques can be applied.

Keywords. Causal inference; Conditional independence; Regression; Sensitivity analysis; Unmeasured confounding.
1. Introduction

Lin, Psaty, and Kronmal (1998) present methods for sensitivity analysis in observational studies which assess the potential influence of unmeasured confounding variables. Their results are of particular interest because in certain circumstances there exists a simple algebraic relationship between (1) the true exposure effect and (2) the exposure effect in the model which does not control for the unmeasured confounding variable. One can then use these results to produce sensitivity bounds for the true exposure effect by making simple adjustments to point and interval estimates obtained from the model which does not control for the unmeasured confounding variable. Lin et al. (1998) present results for linear, log-linear and logistic regression and also for the Cox proportional hazards model (Cox, 1972). In all cases the exposure is assumed to be binary. Progress is made possible principally by assuming that the measured covariates $Z$ and the unmeasured confounding variable $U$ are conditionally independent given the exposure variable $X$. However, Hernán and Robins (1999) have noted that this conditional independence relation cannot hold if in fact $Z$ itself contains variables that confound the relationship between the exposure $X$ and the outcome $Y$. In this paper we present the algebraic results of Lin et al. (1998); we contrast the conditional independence assumption which cannot hold if the measured covariates $Z$ contain confounding variables with an assumption concerning additivity which can hold even if both the measured covariates and the unmeasured confounder affect the exposure of interest; we show how the assumption can be used even with continuous exposures; we provide guidelines for assessing the plausibility of the additivity assumption; we discuss the implications of these findings and we clarify which of the results of Lin et al. (1998) can reasonably be used for sensitivity analysis.

2. Algebraic Adjustments

We first present the algebraic relationships derived by Lin et al. (1998). As noted above, all the cases that Lin et al. (1998) consider assume that the exposure $X$ is binary. We
begin with the case of a binary unmeasured confounding variable $U$. Consider a log-linear regression for a binary outcome $Y$,

$$\Pr(Y = 1|X, Z, U) = \exp(\alpha + \beta X + \gamma_X U + \theta'Z)$$

where $\gamma_X$ denotes the regression coefficient for $U$ on $Y$ for a particular exposure level $X$. Note that $\gamma_X$ may vary with $X$. Lin et al. (1998) show that if the model being fit is

$$\Pr(Y = 1|X, Z) = \exp(\alpha^* + \beta^* X + \theta''Z)$$

then if $U$ is independent of $Z$ conditional on $X$ then $\beta$ and $\beta^*$ are related by:

$$\beta = \beta^* - \log \frac{e^{\gamma_1 P_1} + (1 - P_1)}{e^{\gamma_0 P_0} + (1 - P_0)}$$ (1)

where $P_1$ and $P_0$ are $\Pr(U = 1|X = 1)$ and $\Pr(U = 1|X = 0)$ respectively. Equation (1) is of interest for the purposes of sensitivity analysis because it relates, by a simple algebraic formula the true effect of $X$, $\beta$, and the apparent effect of $X$, $\beta^*$. Furthermore, the quantities in the equation relating $\beta$ and $\beta^*$, namely $\gamma_0$, $\gamma_1$, $P_0$ and $P_1$, all have fairly straightforward and intuitive interpretations. The researcher may thus be able to propose a plausible range of values for these parameters to be used in sensitivity analysis.

Lin et al. (1998) also present similar results for a normally distributed unmeasured confounding variable $U$. In the case of log-linear regression, if $U$ is normally distributed conditional on $X$ and $Z$ with mean $\mu_{X,Z}$ and variance 1 and if $U$ is independent of $Z$ conditional on $X$ so that $\mu_{X,Z} = \mu_X$ then $\beta$ and $\beta^*$ are related by:

$$\beta = \beta^* - \{(\gamma_1 \mu_1 - \gamma_0 \mu_0) + 0.5(\gamma_1^2 - \gamma_0^2)\}.$$ (2)
If $\gamma_1 = \gamma_0 = \gamma$ then (2) reduces to

$$\beta = \beta^* - \gamma(\mu_1 - \mu_0).$$

(3)

Lin et al. (1998) note that (3) holds even if $U$ and $Z$ are not conditionally independent given $X$ provided that the mean of $U$ conditional on $X$ and $Z$ is additive in $X$ and $Z$ i.e. $\mu_{X,Z} = \mu_X + q(Z)$. Lin et al. (1998) also show how these relationships hold approximately for logistic regression when the outcome is rare.

It is furthermore shown that equations (1)-(3) hold approximately under the same assumptions concerning $U$ when comparing the Cox proportional hazards model with the unmeasured confounding variable

$$\lambda(t|X, Z, U) = \lambda_0(t) \exp(\beta X + \gamma_X U + \theta' Z)$$

to that without the unmeasured confounding variable

$$\lambda(t|X, Z) = \lambda_0^*(t) \exp(\beta^* X + \theta'^* Z)$$

where $\lambda_0(t)$ and $\lambda_0^*(t)$ are arbitrary baseline hazard functions. Equation (1) holds approximately if $U$ is binary and $U$ is independent of $Z$ conditional on $X$; equation (2) holds approximately provided $U$ is normally distributed conditional on $X$ and $Z$ and provided that $U$ is independent of $Z$ conditional on $X$; equation (3) holds approximately if $\gamma_1 = \gamma_0 = \gamma$ and if $U$ is normally distributed conditional on $X$ and $Z$ with mean that is additive in $X$ and $Z$.

Finally, for the linear regressions

$$Y = \alpha + \beta X + \gamma_X U + \theta' Z + \epsilon$$
and

\[ Y = \alpha^* + \beta^* X + \theta^* Z + \epsilon^* \]

Lin et al. (1998) show that

\[ \beta = \beta^* - (\gamma_1 \mu_1 - \gamma_0 \mu_0) \]

for both binary and normally distributed \( U \) when \( U \) and \( Z \) are conditionally independent given \( X \). For linear regression, if \( U \) and \( Z \) are not conditionally independent given \( X \) but it is the case that \( \gamma_1 = \gamma_0 = \gamma \) then equation (3) holds for both binary and conditionally normally distributed \( U \) provided that the conditional mean of \( U \) is additive in \( X \) and \( Z \).

Lin et al. (1998) also devote further discussion to sensitivity analysis when the conditional independence of \( U \) and \( Z \) does not hold; in this case, no simple algebraic relationship holds between \( \beta \) and \( \beta^* \) and sensitivity analysis entails considerable computational difficulties.

3. Distributional Assumptions

Consider now the principal assumption required to derive equations (1) and (2), that \( U \) is independent of \( Z \) conditional on \( X \). We elaborate on the discussion of Hernán and Robins (1999) concerning why this assumption cannot hold and then contrast this assumption with one concerning additivity. Suppose that \( Z \) contains confounding variables so that \( Z \) affects both \( X \) and \( Y \). Recall that \( U \) is assumed to confound the \( X-Y \) relationships by the context of the sensitivity analysis. Then the relationships between \( U, Z, X \) and \( Y \) can be given as in Figure 1.

If in addition, it is assumed that \( U \) and \( Z \) are unconditionally independent then Figure 1 gives a causal directed acyclic graph (Pearl, 1995). The assumption that \( U \) and \( Z \) are unconditionally independent might plausibly hold for example when \( Z \) consists of environmental factors and \( U \) is a genetic trait (Lin et al., 1998; Hernán and Robins, 1999). It can be shown that it then immediately follows from the d-separation criterion (Pearl, 1995) that \( U \) is not independent of \( Z \) conditional on \( X \) because, conditional on \( X \) there is an unblocked
path between $U$ and $Z$, namely $U - X - Z$. Note that again by the d-separation criterion, $U$ and $Z$ will in general be conditionally associated given $X$ even when $U$ and $Z$ are not unconditionally independent e.g. if there were an arrow from $U$ to $Z$ or from $Z$ to $U$ or if $U$ and $Z$ had a common cause. However, the theory of directed acyclic graph is not needed to see that $U$ will not be independent of $Z$ conditional on $X$ when both $Z$ and $U$ confound the $X - Y$ relationship. As noted by Hernán and Robins (1999), this relationship can be understood in a very intuitive manner. Suppose that both $U$ and $Z$ have an effect on $X$ and that higher values of $U$ and higher values of $Z$ render $X$ more likely. If $X = 1$ because of a high value of $Z$ then it will in general be less likely that $U$ will take on a high value. Similarly if $X = 1$ because of a high value of $U$ then it will in general be less likely that $Z$ will take on a high value. Essentially, $Z$ and $U$ will not in general be independent within strata of $X$ if they both affect $X$. Indeed it can be shown that for binary $X$, if the probability of $X$ depends on both $Z$ and $U$ then $U$ cannot be independent of $Z$ conditional on $X$; there must be at least one stratum of $X$ such that $U$ and $Z$ correlated (Hernán and Robins, 1999).

If $Z$ affected only the outcome $Y$ and not the exposure $X$ then $U$ and $Z$ would be conditionally independent given $X$; but in this case $Z$ would not be a confounding variable. The setting described by Lin et. al (1998) included both measured confounding variables $Z$ and an unmeasured confounding variable $U$. Thus the central assumption used to derive the algebraic relationships (1) and (2) used in sensitivity analysis is violated by the very context in which the sensitivity analysis will be useful.

Consider now instead the assumption concerning additivity needed to derive equation (3) that the mean of $U$ conditional on $X$ and $Z$ is additive in $X$ and $Z$ i.e. $\mu_{X,Z} = \mu_X + q(Z)$. We will show that this assumption is satisfied by an entire family of distributions even when both $U$ and $Z$ involve confounding variables. To do so we will construct a family of distributions constituting data generating processes for $P_{Y,X,U,Z}(y, x, u, z)$ such the additivity assumption is satisfied and such that $X$, $U$ and $Z$ affect $Y$ directly and such that $U$ and $Z$ affect $X$ directly. We first consider the case of a conditionally normally distributed confounding

6
variable $U$. Let $P_{Y|X,U,Z}(y|x,u,z)$ denote the conditional distribution of $Y$ given $X$, $U$ and $Z$. This conditional distribution is given by the regression model under consideration. We have $P_{Y,X,U,Z}(y,x,u,z) = P_{Y|X,U,Z}(y|x,u,z)P_{X,U,Z}(x,u,z)$. Now since $U$ is conditionally normally distributed given $X$ and $Z$ we have by the additivity assumption,

$$U|X,Z \sim N(\mu_X + q(Z), 1).$$

Since $X$ is binary, the conditional distribution of $X$ given $Z$ must be given by,

$$X|Z \sim Ber\{w(Z)\}$$

for some function $w$. Let $P_Z(z) = h(Z)$ denote the marginal distribution of $Z$. Then,

$$P_{X,U,Z}(x,u,z) = P_{U,X,Z}(u|x,z)P_X(x|z)P_Z(z) = \phi[u - \{\mu_X + q(z)\}]w(z)^x(1 - w(z))^{1-x}h(z)$$

(4)

where $\phi$ denotes the probability density function for a standard normal random variable. 

We thus have

$$P_{X|U,Z}(x|u,z) = \frac{P_{X,U,Z}(x,u,z)}{P_{U,Z}(u,z)} = \frac{\phi[u - \{\mu_X + q(z)\}]w(z)^x(1 - w(z))^{1-x}h(z)}{\int x \phi[u - \{\mu_X + q(z)\}]w(z)^x(1 - w(z))^{1-x}h(z)dx}$$

and so

$$X|U,Z \sim Ber\left(\frac{\phi[u - \{\mu_1 + q(z)\}]w(z)}{\phi[u - \{\mu_1 + q(z)\}]w(z) + \phi[u - \{\mu_0 + q(z)\}]\{1 - w(z)\}}\right).$$

We see then that unlike the conditional independence assumption, the additivity assumption can hold for a family of distributions even if both $U$ and $Z$ affect $X$. If $U$ and $Z$ are unconditionally independent then the data generating process for $(Y, X, U, Z)$ is given
by \( P_{Y,X,U,Z}(y, x, u, z) = P_{Y|X,U,Z}(y|x, u, z)P_{X|U,Z}(x|u, z)P_Z(z)P_U(u) \) where \( P_U(u) \) is easily derived from (4) as

\[
P_U(u) = \int_{x,z} P_{X,U,Z}(x, u, z) dx dz
= \int_z \phi[u - \{\mu_1 + q(z)\}]w(z)h(z)dz + \int_z \phi[u - \{\mu_0 + q(z)\}]\{1 - w(z)\}h(z)dz. \tag{5}
\]

If \( U \) and \( Z \) are not unconditionally independent and \( U \) directly affects \( Z \) then the data generating process for \((Y, X, U, Z)\) is \( P_{Y,X,U,Z}(y, x, u, z) = P_{Y|X,U,Z}(y|x, u, z)P_{X|U,Z}(x|u, z)P_{Z|U}(z|u)P_U(u) \).

If \( Z \) directly affects \( U \) then the data generating process for \((Y, X, U, Z)\) is \( P_{Y,X,U,Z}(y, x, u, z) = P_{Y|X,U,Z}(y|x, u, z)P_{X|U,Z}(x|u, z)P_U(Z|u)P_Z(z) \). Expressions for \( P_{Z|U}(z|u) \) and \( P_U(Z|u) \) are also easily derived from (4). What is clear from the discussion, however, is that data generating processes exist for \((Y, X, U, Z)\) such that the additivity assumption needed for sensitivity analysis is satisfied and such that \( X, U \) and \( Z \) affect \( Y \) directly and \( U \) and \( Z \) affect \( X \) directly. Note that in the derivation of the family of distributions that satisfy the additivity condition, we made no assumption other than that \( U \) was conditionally normally distribution given \( X \) and \( Z \) with a mean that was additive in \( X \) and \( Z \). There is thus a one-to-one correspondence between the distributions which satisfy the additivity condition and the distributions derived in this section.

We note that with linear regression models and a binary confounding variable \( U \) with a mean that is additive in \( X \) and \( Z \) i.e. \( \mu_{X,Z} = \mu_X + q(Z) \) it can be shown that

\[
X|U,Z \sim \text{Ber}(\frac{\{\mu_1 + q(z)\}^u[\{1 - \{\mu_1 + q(z)\}\}^{1-u}w(z)]}{\{\mu_1 + q(z)\}^u[\{1 - \{\mu_1 + q(z)\}\}^{1-u}w(z)] + \{\mu_0 + q(z)\}^u[\{1 - \{\mu_0 + q(z)\}\}^{1-u}1-w(z)]})
\]

where \( X|Z \sim \text{Ber}\{w(Z)\} \) as above. We note further that equation (3) can be applied under the additivity assumption to linear, log-linear, logistic and proportional hazards models in contexts which are more general than indicated by Lin et al. (1998). The discussion of Lin et al. concerns only binary exposures; however, for a continuous exposure \( X \), where the
mean of the unmeasured confounding variable $U$ is given by $\mu_{X,Z} = \tau X + q(Z)$, equation (3) may be used by replacing $(\mu_x - \mu_0)$ with $\tau$. In this case, if the conditional probability density function of $X$ given $Z$ is $g(x,z)$ then the conditional distribution of $X$ given $U$ and $Z$ is given by

$$P_{X|U,Z}(x|u,z) = \frac{P_{X,U,Z}(x,u,z)}{P_{U,Z}(u,z)} = \frac{\phi[u - \{\tau x + q(z)\}]g(x,z)}{\int \phi[u - \{\tau v + q(z)\}]g(v,z)dv}$$

when $U$ is conditionally normally distributed and

$$P_{X|U,Z}(x|u,z) = \frac{P_{X,U,Z}(x,u,z)}{P_{U,Z}(u,z)} = \frac{\{\tau x + q(z)\}^u[1 - \{\tau x + q(z)\}]^{1-u}g(x,z)}{\int \{\tau v + q(z)\}^u[1 - \{\tau v + q(z)\}]^{1-u}g(v,z)dv}$$

for linear regression when $U$ is binary.

4. Assessing the Additivity Assumption

Sensitivity analyses for confounding can be classified into analyses in which the researcher assumes a generic unmeasured confounder and analyses in which the researcher has a clear sense as to what the unmeasured variable is that confounds the exposure-outcome relationship. In analyses of the latter type it is often possible to assess the assumptions used in sensitivity analysis that are made about the unmeasured confounding variable. In such analyses, it is also often possible to use the researcher’s knowledge of the unmeasured confounding variable to model more directly the relationship between the unmeasured confounding variable and the exposure, covariates and outcome. In analyses of the former type, without specific subject-matter knowledge, it will be more difficult to assess whether assumptions about the unmeasured confounding variable are reasonable. Nevertheless such analyses can still be helpful in providing rough guidelines as to how substantial the influence of a confounding factor would need to be in order to eliminate or reverse the effect observed without controlling for the unmeasured confounding variable.

In the previous section we have shown that there exist data generating processes for
\((Y, X, U, Z)\) such that the additivity assumption needed for the approach of Lin et al. (1998) is satisfied and such that \(X\), \(U\) and \(Z\) affect \(Y\) directly and \(U\) and \(Z\) affect \(X\) directly. Whether these data generating processes are plausible is another matter. We will comment explicitly on the case of a conditionally normally distributed unmeasured confounding variable \(U\); analogous comments hold for the case of linear regression with a binary unmeasured confounding variable. We have shown that if the additivity assumption of Lin et al. (1998) is to hold then the marginal distribution for the unmeasured confounding variable \(U\) must be given by (5). The distribution represented in equation (5) is a mixture of normally distributed random variables; the functions \(q\), \(w\) and \(h\) are arbitrary - no assumptions have been made with regard to their form. While the mixture of normally distributed random variables in (5) is perhaps not a priori unlikely, the functions \(w\) and \(h\) also appear in the distributions for \(Z\) and \(X\) given \(Z\) respectively. There thus must exist a relation, in functional form, between the distribution of \(U\) and of \(Z\) even when the \(U\) and \(Z\) are marginally independent. The relation between the functional form of these distributions may be quite complicated. Even if \(Z\) consists of a single binary variable, the marginal distribution of \(U\) given in equation (5) becomes

\[
P_U(u) = \phi[u - \{\mu_1 + q(0)\}]w(0)h(0) + \phi[u - \{\mu_0 + q(0)\}]{1 - w(0)}h(0) + \phi[u - \{\mu_1 + q(1)\}]w(1)h(1) + \phi[u - \{\mu_0 + q(1)\}]{1 - w(1)}h(1).
\]

The mixture probabilities \(w(0)h(0), \{1 - w(0)\}h(0), w(1)h(1)\) and \(\{1 - w(1)\}h(1)\) must correspond to the products of probabilities of \(Z = 1\) and of \(X = 1\) given \(Z = 0\) or \(Z = 1\). In general it seems unlikely that these assumptions will be met by realistic data generating mechanisms.

The sensitivity analysis assumptions and results of Lin et al. (1998) are thus perhaps best viewed as a set of simplifying assumptions to obtain algebraic adjustment formulas which provide rough guidelines as to how substantial the influence of a confounding factor
would need to be in order to eliminate or reverse the effect observed without controlling for the unmeasured confounding variable. When specific subject-matter knowledge about the unmeasured confounding variable and its relation to the exposure, covariates and outcome is available, it would be better to incorporate this knowledge more directly into the sensitivity analysis approach.

5. Implications
As noted in Section 2, if $\gamma_1 = \gamma_0 = \gamma$, equation (3) holds for log-linear regression with a binary outcome and holds approximately for the logistic and proportional hazards model when $U$ is conditionally normally distributed with a mean that is additive in $X$ and $Z$ even when $U$ and $Z$ are not conditionally independent given $X$. Furthermore, equation (3) holds true for linear regression for both binary and conditionally normally distributed $U$, when the mean of $U$ is additive in $X$ and $Z$. Unlike the assumption that $U$ and $Z$ are conditionally independent given $X$, the assumption that the mean of $U$ is additive in $X$ and $Z$ can hold true even when both $U$ and $Z$ affect $X$. Thus when $\gamma_1 = \gamma_0 = \gamma$, the algebraic sensitivity analysis results of Lin et al. (1998) can reasonably be applied for the linear, log-linear, logistic and proportional hazards regression models in the case of a conditionally normally distributed unmeasured confounding variable but only for linear regression in the case of a binary unmeasured confounding variable. For logistic regression and the proportional hazards model, equation (3) only holds approximately and Lin et al. (1998) show through simulations that the approximation is better for binary $U$ than for normally distributed $U$ unless the event is rare. Thus for logistic and proportional hazards models, equation (3) should only be employed in settings with rare events. As noted above, a modification of equation (3) can be applied when the exposure of interest is continuous.

Other sensitivity analysis techniques are of course available (Rosenbaum and Rubin, 1983; Copas and Li, 1997; Lin et al., 1998; Robins, Scharfstein, and Rotnitzky, 2000; Brumback et al., 2004). However, in many cases, the techniques cannot be easily implemented by standard statistical software and may require special programming. It is also possible to
provide bounds for causal effects in the presence of unmeasured confounding (Cornfield et al., 1959; Manski, 1990; Balke and Pearl, 1997; MacLehose et al., 2005; VanderWeele and Robins, 2007) but these bounds are sometimes too wide to be of particular use. The attraction of the results of Lin et al. (1998) is that they may be applied by making simple adjustments to point and interval estimates obtained from a model which does not control for the unmeasured confounding variable. Thus equation (3) of Lin et al. (1998) with a hypothetical conditionally normally distributed unmeasured confounding variable is likely to be of some use in giving the researcher a sense as to how sensitive conclusions are to assumptions about having measured all relevant confounding variables. Its use, unlike the use of equations (1) and (2), does not make assumptions that are violated by the very context from which the need for sensitivity analysis arises. Nevertheless the use of equation (3) does make certain assumptions about the relationship between the functional form of the distributions of the unmeasured confounder and the measured covariates and thus the results obtained by such sensitivity analyses should be interpreted cautiously.

References


List of Figures

Fig. 1. Diagram for the confounding variable relationships: Y - outcome; X - exposure; Z - measured covariates; U - unmeasured confounder