REGRESSION MODELLING AND OTHER METHODS TO CONTROL CONFOUNDING

R McNamee

Occup Environ Med 2005;**62**:500–506. doi: 10.1136/oem.2002.001115

Confounding is a major concern in causal studies because it results in biased estimation of exposure effects. In the extreme, this can mean that a causal effect is suggested where none exists, or that a true effect is hidden. Typically, confounding occurs when there are differences between the exposed and unexposed groups in respect of independent risk factors for the disease of interest, for example, age or smoking habit; these independent factors are called confounders. Confounding can be reduced by matching in the study design but this can be difficult and/or wasteful of resources. Another possible approach—assuming data on the confounder(s) have been gathered—is to apply a statistical ''correction'' method during analysis. Such methods produce ''adjusted'' or ''corrected'' estimates of the effect of exposure; in theory, these estimates are no longer biased by the erstwhile confounders.

Given the importance of confounding in epidemiology, statistical methods said to remove it deserve scrutiny. Many such methods involve strong assumptions about data relationships and their validity may depend on whether these assumptions are justified. Historically, the most common statistical approach for dealing with confounding in epidemiology was based on *stratification*; the standardised mortality ratio is a well known statistic using this method to remove confounding by age. Increasingly, this approach is being replaced by methods based on *regression models*. This article is a simple introduction to the latter methods with the emphasis on showing how they work, their assumptions, and how they compare with other methods.

Before applying a statistical correction method, one has to decide which factors are confounders. This sometimes[1–4] complex issue is not discussed in detail and for the most part the examples will assume that age is a confounder. However, the use of automated statistical procedures for choosing variables to include in a regression model is discussed in the context of confounding.

REGRESSION MODELS

As a means of studying influences on a outcome

Most introductions to regression discuss the simple case of two variables measured on continuous scales, where the aim is to investigate the influence of one variable on another. It is useful to begin with this familiar application before discussing confounder control.

Suppose we are interested in describing the decline with age of forced expiratory volume in one second (FEV_1) in non-smokers and that data on both variables has been gathered from a cross-sectional sample of a population. A statistical analysis might begin with a scatter plot of the data (see fig 1); then a *model* of the relationship in the population would be proposed, where the model is specified by a *model form* or *model equation*. The choice of model form should ideally be dictated by subject matter knowledge, biological plausibility, and the data. Suppose a linear relationship is proposed; then the model would have the general form:

$$FEV_1 = a + b.age + r \qquad \text{Model 1}$$

The three unknown quantities in this model—a, b, r—would then be estimated or quantified in the analysis. The model ignoring r (by setting it equal to zero) is a description of the relationship between age and the *mean* FEV_1 among people of a given age. The term r is a *random* component assumed to vary from person to person. Inclusion of this term in the model allows for the fact that people of the same age are not all the same: their individual FEV_1 values will vary about the mean for that age. Random variation is unpredictable but, overall, it can be described by a statistical distribution. With continuous variables such as FEV_1, the random component is often assumed to have a Normal distribution with a mean of zero.

Statistical formulae or software can be used to estimate the *regression coefficients*, a and b, and SD(r), the standard deviation of r, from a data sample. In the ''least squares'' estimation method, the rationale is to choose values for a and b which minimise SD(r) in the data set. Application to

Correspondence to:
Dr R McNamee, Biostatistics Group, Division of Epidemiology and Health Sciences, Faculty of Medical and Human Sciences, The University of Manchester, Oxford Road, Manchester M13 9PT, UK; rmcnamee@man.ac.uk

Received 22 November 2004

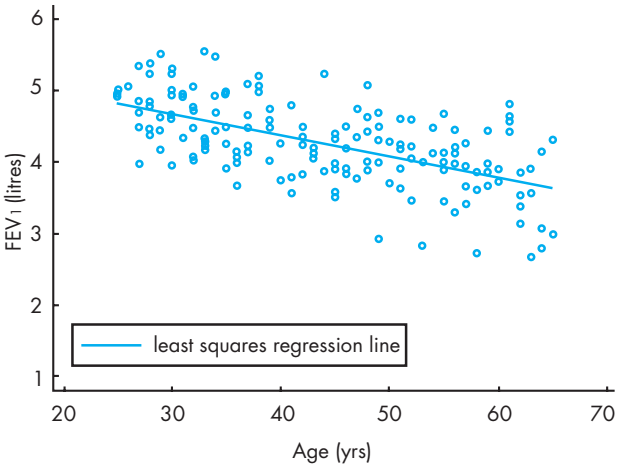

Figure 1 Relationship between FEV_1 and age in 160 male non-smokers.

the data in fig 1 gives estimates of 5.58 litres for a, −0.03 litres/y for b, and 0.46 litres for SD(r). Therefore the "fitted" model is: $FEV_1 = 5.58 - 0.03.age + r$.

Table 1 shows typical software output from fitting model 1 to the data. It includes 95% confidence intervals (CI) for a and b, and p values from significance tests. In each test, the null hypothesis is that the true value of the coefficient is zero. If b were zero, then age would have no effect on FEV_1. Here, the test and the 95% CI strongly suggest that b is negative.

Model 1 is an example of a linear model: it assumes that mean FEV_1 declines by a fixed amount (estimated as 30 ml) for every year of age. It is important to realise that linearity was assumed, not proven: *the statistical analysis merely estimates the coefficients of an assumed model*. We could have proposed a more complicated model equation, for example, quadratic or exponential, and then estimated its coefficients. The process of estimation does not tell us which model form, if any, is right. However there are a range of post-estimation, regression "diagnostic methods" to help with this task, for example, "analysis of residuals" and "leverage" statistics, which highlight discrepancies between the data and the assumed model form. A relatively non-technical account of regression diagnostics can be found in Armitage and colleagues.[5]

Extending the basic model

Other factors besides age are known to affect FEV_1, for example, height and number of cigarettes smoked per day. Regression models can be easily extended to include these and any other determinants of lung function. Model 2 includes height and cigarettes. It *assumes* that each has a linear relationship with FEV_1 and *assumes* that the *joint effect* of the three factors together is the sum of their separate effects:

$$FEV_1 = a + b.age + c.height + d.cigarettes + r$$

Model 2

A standard statistical analysis based on this model and data would produce estimates of a, b, c, d, and SD(r), as well as 95% CIs and "null hypothesis" tests for each coefficient.

It might be argued that, since FEV_1 measures volume, one would expect it to increase proportional to the *cube* of height. To reflect this, we could postulate an alternative equation which includes height^3 among the variables on the right hand side. Regression diagnostic methods can help decide which model form—linear or cubic—is the better fit. Another development would be to consider whether the magnitude of the effect of smoking varies with age. (The phenomenon whereby the effect of one factor is modified or changed by another is known as "effect modification"[6]). Again, one proceeds by proposing a model equation with additional variables and coefficients on the right hand side, followed by an analysis to estimate the coefficients.

How far we go in building up the complexity of the model depends very much on the purpose of the study, for example, prediction, causal analysis, or description. For the present purposes, the important things to note so far are: (1) that regression modelling offers a way of investigating the joint effects of *several* risk factors on health; and (2) that a regression model equation makes strong assumptions about the form of these effects.

Models for disease outcomes

In models 1 and 2 the "dependent" variable, FEV_1, is measured on a continuous scale. A model of this type is not suitable for investigations of disease incidence or prevalence. In the latter case, the dependent variable would have a dichotomous scale—disease present or absent. Two models often used for disease data—the logistic regression model and the Poisson regression model—are discussed briefly later. These models have a different form for the *left* hand side of the model equation. The right hand side of the equation—which specifies the factors we think affect disease risk—remains as flexible as before, both in terms of number of factors and form.

Regression as a means of confounder control

An occupational or environmental epidemiologist recognises that there are multiple risk factors for the disease of interest but typically wants to focus on the casual effect of only one factor, for example, an occupational exposure; hereafter this factor is called "the exposure". In this setting, other risk factors for the disease are considered only because they might be confounders, rather than being of direct interest. The goal is then to study the effect of the exposure on disease, "controlling" or "adjusting" for the others. No new regression methodology is needed to do this. As before, we begin with a model equation of which the right hand side includes terms representing the exposure *and* the potential confounders. The

Table 1 Estimation of model 1 coefficients from data in fig 1: typical software output

FEV_1	Coefficient	Std error	*t* statistic	Probability	95% CI
Age	−0.0301	0.0032	−9.52	<0.001	−0.0363 to −0.0238
Constant	5.5803	0.1440	38.75	<0.001	5.2960 to 5.8647

Mean Square Error, that is, SD(r) = 0.464 litres.

formulae or software used to fit the model are exactly the same as before. In other words, the fact that the epidemiologist labels several of the factors as "confounders" and one as "the exposure", is of no consequence to the fitting process.

Although the technical process is identical, the *practice* of regression modelling can—and should[5 7]—vary with the purpose of the research. Consider the development of a regression model in a different situation, for example, where the purpose is to find a simple method of *predicting* disease status on the basis of symptoms, tests, and personal characteristics of patients. There will be a preference for a simple model that will be easy to translate into simple rules for clinic use. Apart from their ability to predict, the only preference for having a particular factor in the model will be on the ground that it is cheap or easily measured. The issue of the casual relationship between the factors and disease is not central—if the model predicts well, it has done its job. In contrast, the present problem—to address the causal effect of an exposure by control of confounding—means that the exposure variable and sufficient measured confounders *must* be included in the model. A simple model is not required: the final choice of model form should be dictated by the need to eliminate confounding,[7–9] not parsimony.

Automated selection methods in regression

When there are a lot of potential confounders, the time taken to consider if, and how they should be included in models may be substantial. Given that we are not really interested in these relationships per se, it may be tempting to use a statistical algorithm to make the decisions. A number of automated procedures for selecting variables for regression models are included in most regression software. These include *forward entry (or step-up)*, *backward elimination (or step-down)*, and *best subset* method. In all cases the guiding principle is statistical significance and linear relationships are automatically assumed. For example, the forward entry, sequential method starts with an "empty" model—that is, no variables. At step 1, the variable with the most "statistically significant" relationship with the health outcome is entered into the model. At step 2, a second variable is selected from the remainder on the basis that it adds most, in terms of "significance", to the model of step 1. The procedure continues in this way, stopping when no variables outside the model add "significantly" to it.

Automated selection procedures should not[8 9] be relied on to make decisions about confounders as they may result in inappropriate exclusions or inclusions in a model. If a variable is a confounder, it ought to be controlled—regardless of statistical significance. Also, certain variables may *not* be confounders, and therefore should not be controlled,[2] even though they have a statistically significant relationship with disease, for example, factors on the causal pathway between exposure and disease. The fact that different types of automated procedure can give rise to different selections in the same data should also serve as a warning. There is no statistical algorithm that can succeed in identifying all confounders in a study—the decision process will always require subject matter knowledge and judgement,[1 3 8 9] as well as statistical information.

Interpretation of model coefficients

In general, a regression coefficient for a factor in a model estimates the effect of an increase of one unit in that factor, if all other factors in the model stay unchanged *and assuming the model assumptions are correct*. Consider a study of the acute effect on FEV_1 of occupational exposure to flour as measured in mg/m^3 for each individual. To study this relationship, while allowing for the effects of age, height, and smoking habit, the following model with the usual assumptions about r is proposed:

$$FEV_1 = a + b.age + c.height + d.cigarettes + e.flour + r \qquad \text{Model 3}$$

Suppose an analysis estimates e as -0.15 litres/mg/m^3. Then, *if* the model assumptions are correct, this measures the effect, on FEV_1, of an increase in exposure of 1 mg/m^3, but *no change* in age, height, or smoking. In other words, it is a measure of the effect of exposure *unconfounded* by age, height, or smoking. In reality, this claim depends on the appropriateness of the model, the quality of the measurements, and of course there being no omitted confounders.

In practice one can never be sure that all the assumptions of a model are correct. Statisticians refer to methods which tend to give "the right answer", even if assumptions on which they are built turn out to be wrong, as being *robust*, in contrast to methods which are *sensitive* to mis-specification. In the context of confounder control, one should therefore consider the sensitivity of regression analysis to any inbuilt assumptions about the effects of confounders. For example, would it make much difference to our conclusion about exposure if we assumed a linear relationship between disease risk and the confounder age, when really it is exponential? If in doubt, a "sensitivity analysis"—which compares results from different sets of assumptions—is recommended.[8] An example where assumptions do make a difference is given below.

The problem of measurement error afflicts all types of analysis, including regression methods. The impact of error in confounder measurement is perhaps not as widely appreciated as it should be. Poor measurement of confounders reduces the ability to control for their effects by statistical methods: there will be *residual confounding*,[10] despite the claim to have "adjusted" for the confounder(s). Consider a study of exposure and respiratory disease where pack-years of cigarettes is an important risk factor and confounder. If the regression analysis relies on current smoking—yes/no—to "adjust for smoking", then it may not fully reflect the impact of tobacco or its confounding impact; the result may be residual confounding.

REGRESSION VERSUS STRATIFICATION

Control of confounding by stratification

Stratification methods have a long history of use in epidemiology to make age adjusted comparisons of the mortality experience of different groups; for example, to compare an exposed occupational cohort and the general population. Typically in such studies, the crude comparison of rates is confounded by differences in the age distributions of the cohort and general population: "age adjusted" statistics are needed. The steps of the stratification method, illustrated in table 2 for a comparison of lung cancer mortality rates among male pottery workers with all men in England and Wales, are as follows:

- S1: Stratify each group—cohort and population—into a number of subgroups, based on levels of the confounder.

Table 2 Stratification as a means of adjusting rate comparisons for age: death rates among male pottery workers versus all males in England and Wales, 1986–90

	Male pottery workers			England and Wales male population		
Age (y)	P-years	Ca lung deaths	Death rate /1000/y (1)	Death rate/1000/y (2)	Rate ratio (1)/(2)	Weight*
55–59	4334	13	3.0	1.3	2.3	0.346
60–64	2200	11	5.0	2.8	1.8	0.378
65–69	1096	8	7.3	4.1	1.8	0.276
Total	7630	32				

*Weights were chosen to minimise the standard error of the weighted mean (see text).

Here the subgroups (''strata'') are age bands, each spanning five years.

- S2: Compare the cohort and population death rates *in each stratum* by calculating a simple comparative statistic in each. For example, in table 2, the *rate ratio* (cohort rate ÷ population rate) is calculated for each age band.
- S3 (optional): For brevity, combine the comparative statistics from each stratum into a single summary figure. Usually a *weighted mean* of the statistics is calculated; a weighted mean tends to give more ''weight'' to bigger strata. Details of how to choose weights need not concern us here.

Suppose there are no other confounders besides age, and consider the first age band in table 2. If it is reasonable to assume that mortality rates do not vary appreciably within the range 55–59 years, then everyone *within* this band can be treated as having the same effective age. It follows that the rate ratio (RR) for this band, 2.3, is not biased by age differences. The same will be true of the RRs for the other two bands. The three rate ratios are quite close in value and it therefore seems reasonable to combine them (S3) into a single figure. Based on the weights in table 2, their weighted mean is 1.88. As its three components are unbiased, this weighted mean is also unbiased. Hence 1.88 is an estimate of the RR among pottery workers, which is not confounded by age.

Usually an SMR is calculated for such data. If this is done here, using the usual formula,[11] it is found that SMR = 188 = 1.88×100. This is not a coincidence; steps S1–S3 are an alternative way of computing an SMR. The method explained here has the advantage of illustrating clearly the logic of ''adjustment by stratification'', which can also be used in other settings.

Another statistic based on the stratification approach is the Mantel-Haenszel (MH) adjusted odds ratio, often used in case-control studies;[12] in this case, the statistic used at S2 is an odds ratio and the MH statistic is a weighted average of odds ratios from different strata. In theory, stratification can be applied to any study design and any choice of statistic. For example, to compare mean FEV_1 between an exposed and unexposed group while adjusting for age, we could stratify both groups by age (S1), calculate the difference in exposed and unexposed FEV_1 means in each stratum (S2), and calculate a weighted average of the differences across strata (S3).

Stratification versus regression based methods

Assumptions

Stratification based methods are not free of assumptions but, in general, they require fewer assumptions than regression. In particular, using stratification, it is not necessary to make formal assumptions about the relationship between the confounder(s) and the disease/health measure. Consider the lung cancer example above: we did not have to define the form of the relationship between lung cancer risk and age, and hence potentially difficult questions as to the shape of this relationship were avoided. In contrast, in regression we have to formally model these relationships and consider the consequences if we get this wrong.

Measurement accuracy

The stratification method only works if the stratification is sufficiently fine to eliminate the relationship between the disease and the confounder *within* strata. If it is not, then there will be residual confounding. This point is closely connected to the issue of poor measurement of confounders mentioned above. With stratification, we effectively replace the original scale of measurement (for example, year of age) of a confounder by a less accurate, categorical version (for example, 5-year age bands). This may not matter too much with narrow categories, but wide categories (20-year age bands) would lead to *residual confounding*.

Extension to several confounders

In theory the stratification approach can be extended to *several* confounders. Suppose we define eight age bands and five smoking bands. To adjust for both variables may mean that subjects have to be divided up into all 40 (= 5×8) bands formed by the age and smoking combinations. As the number of strata increases, the size of each decreases and the within-stratum estimation process—step S2—becomes unstable. Rothman and Greenland[8] refer to the point ''when stratification has exceeded the limits of the data'', and note that then it can give extreme results which are biased. This problem therefore places a practical limit on the number of variables that can be adjusted for by stratification methods. In contrast the number of variables in a regression model can be quite large—especially when the outcome has a continuous scale—without any major effect. Hence this method tends to be favoured when there are many confounders to be controlled.

PROPENSITY SCORE METHODS TO ADJUST FOR CONFOUNDING

To date, this more recent approach[13 14] has not been used much in occupational or environmental epidemiology, but this may change. It is explained here for the simple case where there are two exposure groups—exposed and unexposed—and the health outcome is disease status. A propensity score analysis has two stages:

- P1: Build a regression model to *predict exposure group*. In this model the *left* hand side of the equation is exposure

group, while the right hand side includes variables which influence exposure group membership. This analysis can be viewed as an attempt to model the selection process which leads to some people being exposed and others not, and thereby to understand how exposed and unexposed differ *before* exposure took place. Research authors often address this issue informally by a table which compares characteristics—for example, gender, age, year of birth, smoking—of exposed and unexposed. The difference here is that these characteristics are considered jointly in a regression model. Furthermore, this model is used to give each person a *propensity score* which measures the *propensity (probability) to be exposed* given their individual characteristics.

- P2: In the second stage, disease rates are compared between exposed and unexposed after adjustment for propensity scores. The adjustment method can be based on regression or on stratification. In these analyses, the propensity score—a single measure—takes the place usually occupied by several confounders in traditional analyses.

There has been little research into the advantages and disadvantages of propensity methods for confounder control compared to traditional regression methods. Since propensity methods use regression in P1, they involve the usual assumptions and issues of variable selection. As elsewhere, reliance on purely statistical criteria to choose variables correctly at this step is not guaranteed to take care of confounders.

FURTHER TOPICS IN REGRESSION

Regression models for comparing groups

If we are only able to classify subjects into exposure *groups*, for example, low, medium, high, the exposure variable is said to be categorical. Categorical variables can be included as predictors on the right hand side of a model equation although special rules are needed when there are three or more categories. We concentrate here on the case of two categories—exposed and unexposed.

Consider a study comparing FEV_1 in two groups where we want to adjust the comparison for age, height, and smoking. The data set contains a variable "group" which takes the value 1 if the subject is exposed and 0 if unexposed. Now consider the following model:

$$FEV_1 = a + b.age + c.height + d.cigarettes + e.group + r \qquad \text{Model 4}$$

The coefficient e is the one of primary interest. Its interpretation follows from the general definition given above: it is the effect of increasing "group" by 1 unit, *assuming* all other variables are fixed and the model is correct. In other words, it is the mean difference in FEV_1 between exposed (group = 1) and unexposed (group = 0) if age, height, and cigarettes were the same in each group.

Some models for disease outcomes: logistic and Poisson regression

Suppose that the outcome measure is presence or absence of respiratory disease at a point in time. This is a dichotomous variable with values 1 or 0 according to disease status. To model the probability (p) of disease as a function of age, cumulative pack-years of cigarettes smoking and exposure concentration, a possible model would be:

$$\ln\left(\frac{p}{1-p}\right) = a + b.age + c.pack\text{-}yrs + d.conc$$

This is called a logistic regression model; the term logistic denotes the form of the *left* hand side of the equation. In this model the random component is assumed to have a binomial distribution. Data from prevalence studies, and longitudinal incidence studies with a fixed duration of follow up, can be analysed using logistic models.

A logistic regression model is not suitable for incidence studies where the length of follow up varies among subjects. In such studies the basic measure of interest is the incidence density rate = number of cases of disease (Y say) divided by person-time of observation (T say). Poisson regression models are a natural extension for investigating factors which affect these rates. Imagine an occupational cohort study with two exposure groups monitored over many years for lung cancer. A possible Poisson model, with predictors age, cigarette pack-years, and exposure group is:

$$\ln Y = \ln(T) + a + b.age + c.pack\text{-}yrs + d.group + r$$

The term "Poisson" stems from the assumption of a Poisson distribution for the random component of the model. Further discussion of these and other possible models for disease data is beyond the scope of this article, but can be found in Checkoway and colleagues[11] and Jewell.[12]

Minimising assumptions: semi-parametric regression

Regression models are attractive because of their flexibility in dealing with several influences on disease. But this flexibility comes with a price—reliance on strong assumptions about relationships, for example, linear, quadratic, exponential assumptions. Here we consider two ways of avoiding these "parametric" assumptions, which can be applied to all or just some of the variables in a model. The resulting regressions are then labelled "semi-parametric".

Consider a model of the effects of pack-years of cigarettes on disease risk. The first method, based on categorisation, requires very little specialist knowledge to implement. The continuous variable, pack-years, is categorised—into k categories say—and all but one of these categories (that is, k−1) is represented in the model separately. The omitted category is a "baseline" category, often the category with the lowest risk, for example, never smokers. In this model, risks in different categories are not constrained to follow a set pattern. We can visualise the relationship between risk (y axis) and smoking category (x axis) in such models as a series of steps: as we move from one smoking category to another, risk jumps up (or down) to another step but the steps can be of different heights. One difficulty with this method is that it is susceptible to "gerrymandering",[8] that is, changing the boundaries of the categories until one gets a result that one likes. Nevertheless this method is recommended as an exploratory first step, even if a parametric model is the eventual goal, since it can help in choosing the parametric model form.

In the above method, risk "jumps" at the category boundaries but is flat in between. In reality, risk functions are likely to be smooth. A way of achieving a smooth relationship—without parametric assumptions—is to use a

Table 3 Estimation of coefficients of linear model fitted to data in fig 2

SBP	Coefficient	Std error	*t* statistic	Probability	95% CI
Age	1.81	0.0963	18.75	<0.001	1.62 to 2.00
Group	−6.93	2.3973	−2.89	0.004	−11.66 to −2.19
Constant	53.28	3.1976	16.66	<0.001	46.97 to 59.59

Mean Square Error (that is, SD(r)) = 8.3 mm.

scatterplot smoothing technique. Spline methods[15] are an example. In these, the predictor variable x (pack-years) is again divided into a number of categories; a smooth curve linking risk and x is fitted within each category but the curves are made to join smoothly at the category boundaries (now called "knots"). The number of categories can be made very large so that the overall curve is quite free to bend and flex as needed. Spline methods are often used in studies of environmental pollution and daily mortality to remove confounding by time related factors such as season and influenza epidemics. Such methods may also be used to explore the exposure-disease relationship itself, as in a recent study[16] of the relationship between lung cancer risk and exposure to silica. An attraction of such methods over parametric regression is their ability to reveal threshold effects.

When wrong assumptions lead to false relationships: an example

The following example illustrates a situation where the wrong parametric assumption makes a substantial difference to the results of a causal analysis. The data—on systolic blood pressure and age—were generated artificially from a known model form, but analysed assuming a different form. Suppose the true relationship, among all adult men in a certain community, is given by the solid curve in fig 2A; this non-linear relationship was generated by the equation: SBP = 99 + 0.1 × age + exp(age/15). Also assume that individual SBP values vary randomly about the curve, with the random component having a Normal distribution with SD of 5 mm Hg. The dots in fig 2A are blood pressures for a fictional sample of 171 men from this community, their values having been generated for present purposes using the above assumptions and a random number generator. For simplicity we assume there is no diurnal variation in SBP and that it can be measured without error.

Now suppose all these men work in a factory where 87 of them are exposed to a factor wrongly suspected of increasing blood pressure; the remaining 84 are unexposed. An investigator measures their SBP and age. From fig 2B, one can see that exposed men are generally older than the unexposed. To estimate the effect of exposure, "adjusted" for differences in age, the investigator proposes a linear regression model as follows: SBP = a + b.age + c.group + r, where r is a Normal random component. The results of the regression analysis are shown in table 3. Of most interest is the coefficient for c which estimates the effect of exposure, "adjusted" for age: it is −7 mm Hg (95% CI −12 to −2, p = 0.004). Thus this analysis, which "controls for age", has found a statistically difference in SBP between exposure groups, yet we know there is no exposure effect.

What has gone wrong? The linear model wrongly assumes a constant age gradient across the whole age range. Since the exposed men lie at one end of this range and the unexposed

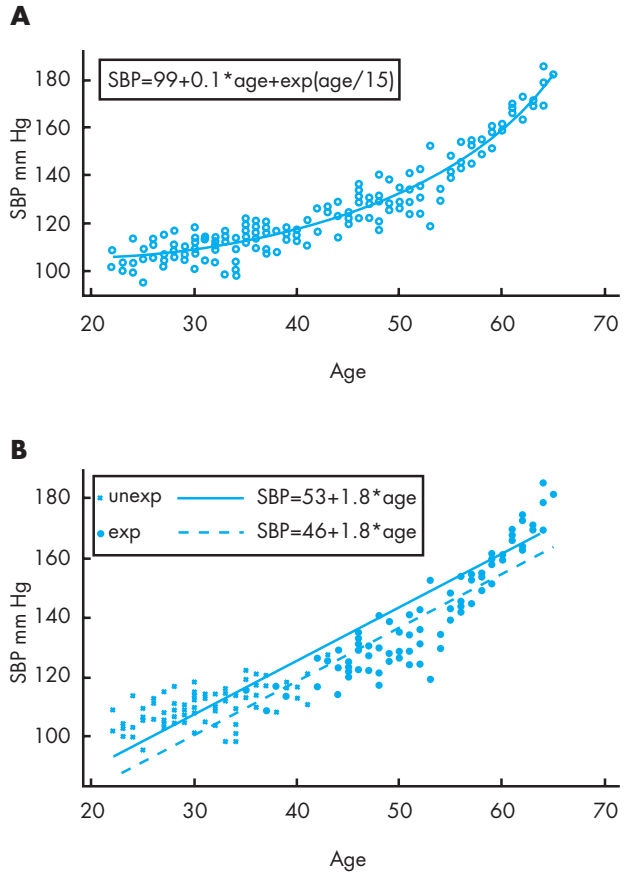

Figure 2 Relationship between SBP and age in 171 men. (A) True relationship. (B) Fitted relationship, assuming linear model.

Main messages

- Regression models are a flexible way of investigating the separate or joint effects of several risk factors for disease or ill health. These factors may include exposures and confounders of the exposure-disease relationship.
- Usual (parametric) regression requires strong assumptions to be made about the form of the relationship between disease risk and each risk factor in the model.
- Claims of having "controlled for confounders" depend to some extent on the validity of such assumptions. Hence sensitivity analysis and regression diagnostic methods are recommended.
- Other methods for dealing with confounders, some of which require fewer assumptions, include non-parametric regression, stratification-based methods, and the propensity score approach.
- Automated statistical algorithms based on statistical significance, should *not* be relied on for selecting confounders to include in regression models.

at the other, the implications of this false assumption are different for the two groups. This differential gets translated into an ''exposure effect''. It might be argued that no sensible person would try to make a comparison between two groups when their age distributions are so different. But the point is, whether sensible or not, regression methods enable an ''adjusted'' comparison to be made—provided we make some assumptions about how SBP changes with age. It is instructive to consider what would happen in an analysis based on stratification into 5-year age groups. The stratification method makes no assumptions about how SBP varies with age. All men under 35 and over 49 would be dropped from the comparison and no exposure effect would be found in the remainder.

Competing interests: none declared

REFERENCES

Most statistical textbooks discuss regression modelling and can be consulted for technical details. However, for discussion of their application to confounder control, books aimed at an epidemiological audience are often preferable.

1 **McNamee R**. Confounding and confounders. *Occup Environ Med* 2003;**60**:227–34.
► **Contrasts competing definitions of a confounder, including those based on data and those based on notions of comparability.**
2 **Weinberg CR**. Towards a clearer definition of confounding. *Am J Epidemiol* 1993;**137**:1–8.
► **Highlights the problem of over-adjustment.**
3 **Greenland S**, Morgenstern H. Confounding in health research. *Annu Rev Public Health* 2001;**22**:189–212.
► **A review of current thinking.**
4 **Greenland S**, Pearl J, Robins JM. Causal diagrams for epidemiological research. *Epidemiology* 1999;**10**:37–47.
► **The problem of identifying confounders of an exposure-disease relationship is addressed through causal diagrams.**
5 **Armitage P**, Berry G, Matthews JNS. *Statistical methods in medical research*, 4th edn. Blackwell Publishing, 2002.
► **A very good intermediate level textbook, even though not especially orientated towards epidemiology.**
6 **Last J**, ed. *Dictionary of epidemiology*, 4th edn. Oxford University Press, 2000.
7 **Maldonado G**, Greenland S. Simulation study of cofounder-selection strategies. *Am J Epidemiol* 1993;**138**:923–36.
► **Compares a number of data based strategies for selecting variables to include in regression models when the aim is to control confounding.**
8 **Rothman KJ**, Greenland S. *Modern epidemiology*, 2nd edn. Philadelphia: Lippincott-Raven Publishers, 1998.
► **Insightful on almost all methodological questions in epidemiology and includes two chapters on regression modelling.**
9 **Robins JM**, Greenland S. The role of model selection in causal inference from non-experimental data. *Am J Epidemiol* 1986;**123**:392–402.
10 **Armstrong B**. The effect of measurement error on relative risk regressions. *Am J Epidemiol* 1990;**132**:1176–84.
► **Easy to read introduction to this topic.**
11 **Checkoway H**, Pearce NE, Crawford-Brown DJ. *Research methods in occupational epidemiology*. New York: Oxford University Press, 2004.
► **Useful section on modelling of dose and exposure data as well as being a good all-round textbook.**
12 **Jewell NP**. *Statistics for epidemiology*. Chapman and Hall, 2004.
► **Intermediate text with good coverage of epidemiological statistics.**
13 **Rosenbaum PR**, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika* 1983;**70**:41–55.
14 **Wiles N**, Lunt M, Barrett E, *et al*. Reduced disability at five years with early treatment of inflammatory polyarthritis: results from a large observational cohort, using propensity models to adjust for disease severity. *Arthritis Rheum* 2001;**44**:1033–42.
► **This practical application of the method of propensity score adjustment may be easier to read than reference 13.**
15 **Ruppert D**, Wand M, Carroll R. *Semi-parametric regression*. Cambridge University Press, 2003.
16 **Eisen EA**, Agalliu I, Thurston SW, *et al*. Smoothing in occupational cohort studies: an illustration based on penalised splines. *Occup Environ Med* 2004;**61**:854–60.
17 **Kleinbaum D**, Kupper L, Muller K, *et al*. *Applied regression analysis and multivariate methods*. Duxbury Press, 1998.
► **A thorough exposition of all aspects of parametric regression, including a chapter on confounding and interaction in regression.**
18 **Kleinbaum D**, Kupper L, Morgenstern H. *Epidemiological research. Principles and quantitative methods*. Wiley, 1982.
► **A sophisticated text which examines the problem of control of confounding—whether by stratification or by regression—in detail.**

QUESTIONS (SEE ANSWERS ON P 472)

Indicate whether each of the following statements is true or false.

(1) Parametric regression models for disease require assumptions to be made about the form of the relationships between risk and each variable, including confounders.
(2) Automated statistical selection methods are the best way to decide which variables are confounders.
(3) Wrong assumptions about the form of the relationship between confounder and disease can lead to wrong conclusions about exposure effects.
(4) The SMR is a statistic where the adjustment for age is based on stratification.
(5) There is no alternative to linear regression modelling.