Stratification and Statistical Modelling in Epidemiology

Michael Hills and David Clayton

May 11, 2001

Summary

The continuing popularity of stratification as a way of controlling for confounders is due to the fact that the process is simple and intuitively appealing (Rothman, 1986[1]). Statistical modelling is a more powerful technique, but it is neither simple nor intuitive and the fact that the exposure variable and the stratifying variable have the same logical status means that most of the results are either irrelevant or not in an immediately useful form. There is no reason for this divide. It is possible, as we show in this article, to cast the input and output for an analysis in the style of stratification, but to use statistical models for the computations, and thus reap the benefits of both approaches. We also describe a graphical interface written for the Stata statistical package to do this.

1 Introduction

A central problem in epidemiology is to control the effect of an exposure on the outcome for one or more confounding variables. Using stratification, this is done by forming strata within which the confounding variables are approximately constant, estimating stratum–specific effects of exposure, checking to see whether these stratum–specific effects are roughly the same, and finally combining them to form a single estimate of their common value. This single estimate is then called the effect of exposure controlled for the confounding variables. If the effect varies appreciably from one stratum to another it is said to be modified by strata, and is usually reported separately for each stratum.

The word control is often used to describe both the act of estimating stratum–specific effects and the act of combining them. This can lead to confusion, and we have adopted the following convention:

- A variable is selected as a *modifying* variable if the effects of exposure are to be estimated at different values of the variable in order to see whether they differ.

- A variable is selected as a *control* variable if the effects of exposure are to be estimated at different values of the variable and then combined to give a single estimate.

Although stratification is intuitively appealing it does have some limitations. In order to form strata the variable which is being controlled for must be categorical: it is not possible to control for a metric variable without first grouping its values and converting it to a categorical variable. Controlling for two or more variables involves forming strata for every combination of values of the variables, which can lead to very small strata and consequent loss of information.

The situation with statistical models is the mirror image of this. It is now possible to control for many confounders without losing information, and the confounders can be metric as well as categorical. But statistical modelling, in its conventional form, also has some limitations. Instead of stratum specific effects there are interaction terms which are not easy to interpret. In addition the sprawling output, 90% of which is usually irrelevant to the aims of the study, makes it easy to lose the clarity of purpose which comes with stratification (Vandenbroucke, 1987 [2]).

These limitations arise because statistical modelling grew out of multiple regression models in which the mean (μ) of a Gaussian distribution is related to other variables by

$$\mu = \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

where the X's are measurements on a continuous scale. The introduction of indicators allowed variables to be categorical, product terms allowed interactions, and link functions and families allowed other distributions, but the interpretation of the regression coefficients relies on understanding how the multiple regression model was manipulated into fulfilling its new roles. Of course this has become second nature to experienced users of statistical modelling, but for those with less experience the use of these models is something of a minefield.

In this paper we show how the framework of exposure, modifier, and control variables can be used to guide the way the statistical model is set up, and to present the results without reference to the underlying X variables. It is our hope that this approach, which could be implemented in any package with window facilities, will encourage the sensible and fruitful use of statistical modelling by analysts who do not necessarily have much insight into how statistical models are estimated by computer packages. In section 6 we introduce a particular implementation using the statistical package Stata.

2 A simple example

The example refers to a study of 337 subjects followed after recording their weighed diet over two weeks (Morris, Marr, and Clayton (1977) [3]; Clayton and Hills (1993) [4]). The outcome of interest is in the variable `d`, coded 1 if the subject developed coronary heart disease, 0 otherwise. The length of follow–up in years is in the variable `y`. The exposure of interest in this analysis is the total energy consumption per day, averaged over the two

week period. For the purposes of this example the effect of the variable `energy` is studied in two ways:

1. By using a categorical variable `hieng` with two levels, formed by cutting `energy` into two groups: < 2500 and 2500+ Kcals per day.

2. By using `energy` itself, as a metric variable.

The subjects in the study were obtained from three different occupations (bus drivers, bus conductors, and bank workers), coded in the variable `job`, so `job` was treated as a potential confounder. Another potential confounder was `height`, and of course age, but only `job` and `height` are used in this example.

Using strata based on `job` we obtain the following stratum–specific estimates of the rate ratio comparing level 2 with level 1 of `hieng`:

```
Effects of hieng on the ratio scale

Level or value
of job                          Effect  95% Confidence Interval

driver                          0.410   [ 0.124 , 1.362 ]
conductor                       0.655   [ 0.227 , 1.888 ]
bank                            0.518   [ 0.212 , 1.267 ]
```

Thus a high energy diet reduces the risk of coronary heart disease in all three strata, presumably becuase it is associated with an active life-style. To control for `job`, the stratum–specific estimates are combined to give

```
Level 2 versus level 1          Effect  95% Confidence Interval

                                0.525   [ 0.290 , 0.949 ]
```

To do the same thing with statistical modelling requires a model which includes `hieng`, `job` and their interaction. Although both `hieng` and `job` are included in the model, no distinction is drawn between them – the model is symmetric in the two variables. Output from a statistical modelling program would look something like this.

```
Effect                          Coeff        s.e.

hieng==1                        0.410        0.251
job==1                          1.137        0.568
job==2                          0.813        0.371
hieng==1 & job==1               1.597        1.304
hieng==1 & job==2               1.262        0.964
_cons                           0.014**      0.005
```

The effect of `hieng` which is reported (0.410) refers to the effect of `hieng` at the first level of `job`. The interactions contrast this with the effect of `hieng` at the other levels of `job`. Thus the effect of `hieng` at the second level of `job` is $0.410 \times 1.597 = 0.655$ and $0.410 \times 1.262 = 0.518$ at the third level of `job`.

The main difference between these two approaches is that with stratification the effects of exposure are reported for each stratum, while in statistical modelling the effects of exposure are reported only for the first stratum. The effects in the other strata can be recovered using the interaction terms, but to do this it is necessary to provide further information about the functions of the variables in the analysis.

Another difference is that with stratification the effects of the `job` are not reported because they are irrelevant to the analysis, but because the statistical modelling program treats `hieng` and `job` symmetrically, it also reports the effects of `job` at the first level of `hieng`. The interactions contrast these with the effects of `job` at the other levels of `hieng`. To reproduce this with stratification it would be necessary to think of `job` as the exposure and form strata using `hieng`.

3 The functions and attributes of the explanatory variables

Explanatory variables are either metric (ie a measurement with units) or categorical. In this example exposure is measured both as a categorical variable, `hieng` with 2 levels, and as a metric variable, `energy` which takes values on a continuous scale with units 1 Kcal. The variable `job` is categorical, and the variable `height` is metric with units 1cm.

The first choice to be made in an analysis is the exposure variable. There may be several exposure variables of interest, but these should be studied separately, so for any particular analysis there is only one exposure variable (we return to this point in the discussion). When the exposure variable is categorical the effects of exposure are reported relative to one of the levels (usually the first) called the base level. When the exposure variable is metric the effects of exposure are reported per unit of the variable, or some multiple like per 100 units or per 0.1 units. These are important attributes of the exposure variable.

The next choice is whether to include a modifying variable, and finally there may be one or more control variables. Modifying variables and control variables can be either categorical or metric.

4 Parametrizing the statistical model

To study the extent to which the variable `job` modifies the effects of `hieng` in the conventional formulation, `hieng`, `job` and the interactions between them, are entered in the model. When the exposure is categorical it is entered in the form of indicator variables which pick out the levels; for two levels there are two indicator variables, but the one

corresponding to the base level is omitted, leaving one. Similarly for the variable `job`. The extent to which `job` modifies the effect of `hieng` is measured by the interactions which are included as products made up from the indicator for `hieng` multiplied by each indicator for `job`. Thus the model will contain the variables

$$A_2,\ B_2,\ B_3,\ A_2 \times B_2,\ A_2 \times B_3$$

where A_i is the indicator for level i of `hieng`, B_j is the indicator for level j of `job`. The coefficient of A_2 is the effect of `hieng` (level 2 compared to level 1) when `job` is at level 1, and the coefficient of $A_2 \times B_2$ measures the extent to which the effect of `hieng` differs when `job` changes from level 1 to level 2.

Once the functions of the variables have been specified it is a simple job to reparametrize the statistical model so that the terms refer to the effects of `hieng` for each level of `job`. This is done by changing the list of variables to be included to

$$B_2,\ B_3,\ A_2 \times B_1,\ A_2 \times B_2,\ A_2 \times B_3$$

Now the coefficients of $A_2 \times B_1, A_2 \times B_2, A_2 \times B_3$ refer to the effect of `hieng` when `job` is at levels 1, 2, 3 respectively.

The test to see whether the effects of `hieng` are modified by `job` is most easily carried out using the model in its conventional form. A test for no effect modification is then the same as the test for no interaction. When this test (plus inspection of the stratum–specific effects) suggests that `job` does not modify the effect of `hieng`, the next step is to control for `job` by excluding the interaction terms and fitting the model which includes

$$A_2,\ B_2,\ B_3$$

The coefficient of A_2 is the effect of `hieng` controlled for `job`.

To do the same thing with `energy` as a metric exposure requires an additional step in modelling, namely the assumption that the effect of a unit change in `energy` is the same at each value of `energy` (ie a linear relationship between the log rate and energy). The effect per Kcal will be extremely small, so it would be more sensible to change this to (say) per 100 Kcals. The model will include

$$B_2,\ B_3,\ \texttt{energy} \times B_1,\ \texttt{energy} \times B_2,\ \texttt{energy} \times B_3$$

and the ccoefficients of the last three terms are the stratum–specific effects of `energy` per 100 Kcals. The results look like this:

```
Effect per 100 unit(s) of energy

Level or value
of job                  Effect   95% Confidence Interval

driver                  0.9078   [ 0.793 , 1.039 ]
conductor               0.9028   [ 0.794 , 1.027 ]
bank                    0.8739   [ 0.783 , 0.975 ]
```

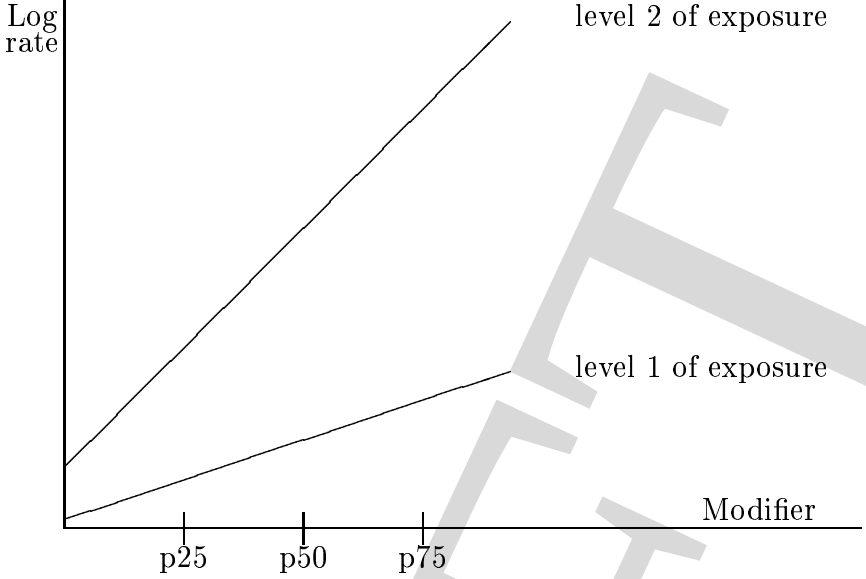

Figure 1: Displaying effects when the modifier is metric

The variable `job` does not seem to modify the effect of `energy` so the effect of `energy` can be controlled for `job` by including the variables

$$B_2,\ B_3,\ \texttt{energy}$$

The coefficient of `energy` is the effect of `energy` controlled for `job`.

5 A metric modifier

To see whether the metric variable `height` modifies the effect of `hieng` it is again necessary to make the assumption that a change of 1 unit in `height` is the same at each value of `height` (ie a linear relationship between the log rate and `height`), but rather than display the different slopes of the lines, we choose to display predicted effects of `hieng` at the 25, 50, and 75 percentiles of `height` (see Figure 1). In this way the results are made to look the same as for a categorical modifier, and are easier to interpret:

```
Effects of hieng on the ratio scale

Level or value
of height                    Effect   95% Confidence Interval

p25                          0.6409   [ 0.348 , 1.181 ]
p50                          0.5673   [ 0.295 , 1.091 ]
p75                          0.4916   [ 0.193 , 1.249 ]
```

With `energy` (a metric exposure) and `height` (a metric modifier), the model contains the terms

$$\texttt{energy},\ \texttt{height},\ \texttt{energy} \times \texttt{height}$$

and is displayed as

```
Effect per 100 unit(s) of energy

Level or value
of height                      Effect   95% Confidence Interval

p25                            0.9256   [ 0.862 , 0.994 ]
p50                            0.8931   [ 0.825 , 0.967 ]
p75                            0.8564   [ 0.762 , 0.962 ]
```

Thus metric exposures and modifiers can be included in the statistical model but the results can be made to follow the style of stratification.

6 The graphical interface

Stata is a command line driven package, and in some ways the use of a menu is contrary to the spirit of the package. Although we agree that serious data analysis requires the use of a reproducible file of commands which serves as a record of the analysis which has been carried out, the use of a menu can greatly simplify interactive work. The information from the menu is available in the form of global macros, so it is easy to include these in the reproducible file of commands, and to by pass the menu, when a record is required.

There are several things which must be done before using statistical modelling in the way we have suggested. The first is to specify which of the variables are categorical, and which are metric; the second is to declare the exposure, modifying, and control variables, together with their attributes; and finally the appropriate model must be specified. For this we have set up a menu, called by the command `efmenu`. Once the information from the menu is available the model can be fitted and the results arranged in the style of stratification, using the command `effects`. The menu is in three parts, the first of which is shown in Figure 2. This is used to declare which variables are categorical (metric is the default). Pressing OK produces the second part, shown in Figure 3. The first box in this menu refers to the statistical model which will be used – in this case it is `poisson` because the data are concerned with events in time. The outcome variable is `d` and `e(y)` is entered in the model options box as this is how Stata provides the follow–up time. To display the effects on a ratio scale, the exponential box is checked. The exposure and modifying variables are selected in the next two boxes, and in the final part the control variables can either be selected from a list of variables or given in the form of a model formula. Because the stata windowing commands do not respond immediately to the information which is entered, it is necessary to press OK again, and this causes further boxes to be shown (see Figure 4), depending on the nature of the exposure and mofiying variables. For example, if the exposure is metric a "per" box appears, and if the modifier is metric a "showat" box appears. There is also a description of the model as currently specified, and if this corresponds to what the user wants a further press of OK will close the menu, and the `effects` command can be run. The results are shown below.

```
model      :      poisson d, e(y)
exposure   :      hieng     (categorical)
modifier   :      job       (categorical)
```

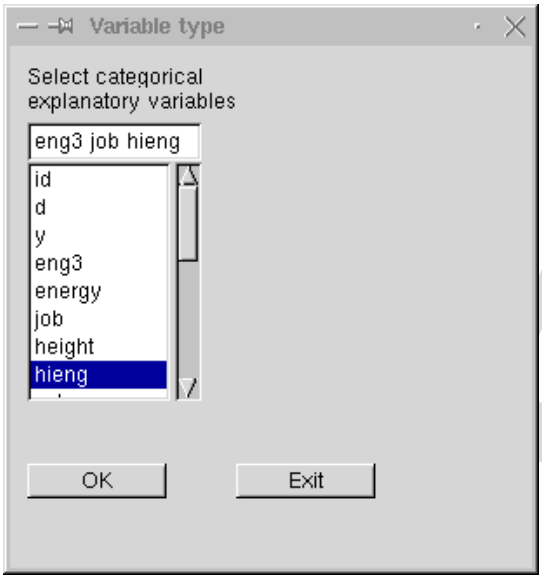

Figure 2: The first part of the menu.

```
Number of records used in the fit :  337

Effects of hieng on the ratio scale

Level 2 versus level 1

Level or value
of job                          Effect   95% Confidence Interval

driver                          0.4103   [ 0.124 , 1.362 ]
conductor                       0.6551   [ 0.227 , 1.888 ]
bank                            0.5177   [ 0.212 , 1.267 ]

Overall test for effect modification
chi2(  2)    =    0.331   P-value      =   0.847
```

To control the effect of `hieng` for `job` the variable `job` must be removed from the modifier box and added to the control box. The results look like this:

```
model      :       poisson d, e(y)
exposure   :       hieng     (categorical)

controlled :       job       (categorical)

Number of records used in the fit :  337
```

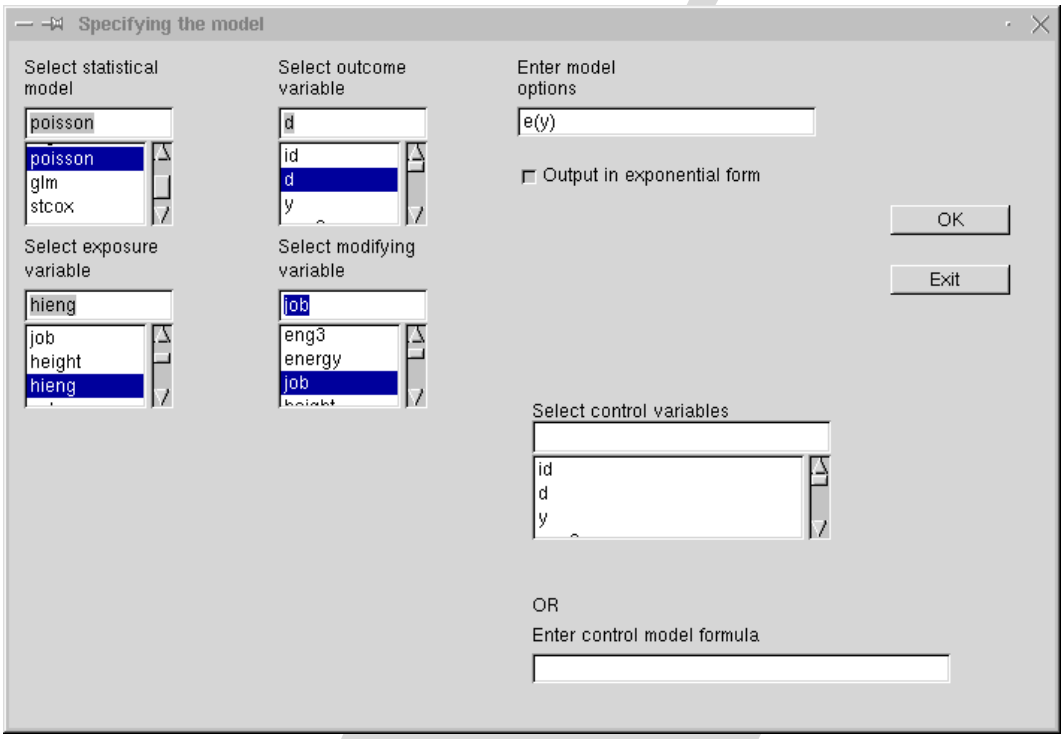

Figure 3: The second part of the menu.

```
Effects of hieng on the ratio scale

Level 2 versus level 1       Effect    95% Confidence Interval

                              0.5248   [ 0.290 , 0.949 ]

Test for no effects of exposure

Poisson:  likelihood-ratio test   chi2(1)     =       4.69
```

The only new thing here is that a test for no effects of exposure has been carried using a likelihood ratio test.

A metric modifier is displayed at its quartiles, by default, but it is possible to select values at which to display the predicted using the showat box. For example, to display the effects at values 155, 165, 175, and 185 for height, enter 155(10)185 in the showat box.

When controlling for both `job` and `height` they can be selected in the control variables box. The additive model is selected by default, so the effects of a unit change in `height` is assumed to be the same at each value of height and for each level of `job`. The same result would be achieved by entering `job + height` in the control formula box. It would also be possible to fit the model with different linear effects of `height` by entering `job*height` in the control formula box.

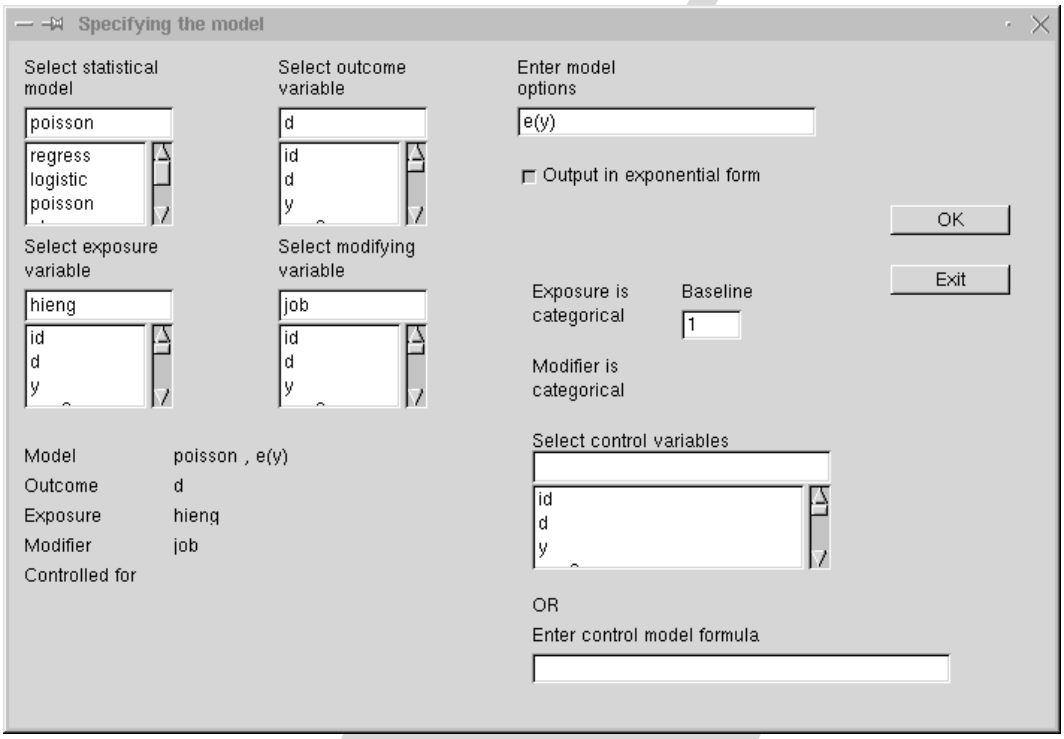

Figure 4: The third part of the menu

As a final example Figure 5 shows the menu for a metric modifier (`height`) controlled for `job` using the level 2 of `hieng` as the base.

The results look like this.

```
model      :      poisson d, e(y)
exposure   :      hieng     (categorical)
modifier   :      height    (metric)

controlled :      job       (categorical)

Number of records used in the fit :  337

Effects of hieng on the ratio scale

Level 1 versus level 2

Level or value
of height                   Effect   95% Confidence Interval

155                         0.9444   [ 0.197 , 4.529 ]
165                         1.3368   [ 0.612 , 2.921 ]
175                         1.8923   [ 0.895 , 3.999 ]
```

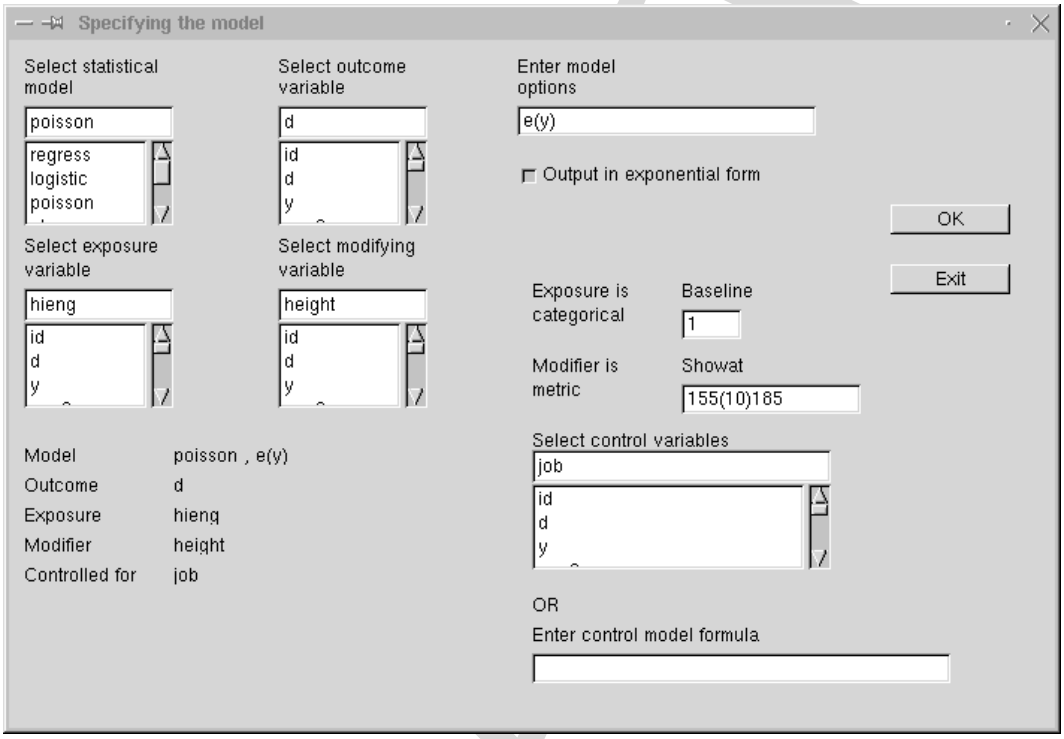

Figure 5: A modifier and a control variable

```
185                          2.6786   [ 0.587 , 12.227 ]

Overall test for effect modification
chi2(  1)    =   0.516
P-value      =   0.472
```

Although the `diet` example uses Poisson regression the ideas apply to any of the commonly used regression models in epidemiology, including cox, logistic, conditional logistic, OLS regression, and many others. In Stata the results of any estimation command are stored in the same way, so a single `effects` program can be used with many different models. A similar program could be preapred for any package which works in the same consistent way.

7 Discussion

The framework allows one exposure variable, one modifier, and many control variables. Some may see this a restriction, but in practice it is a useful constraint. When there are two exposures, such as alcohol and tobacco in oral cancer, we should estimate the effect of alcohol controlled

for tobacco in one analysis, and the effect of tobacco controlled for alcohol in another. These two analyses simulate two quite different quasi-experiments: in the first tobacco is controlled, but alcohol is allowed to vary; while in the second alcohol is controlled but tobacco is allowed to vary. The two analyses thus answer two quite different scientific questions. The conventional regression approach answers both questions in the same analysis, while the approach we have presented in this paper requires two separate analyses for these two different questions. This apparent disadvantage is really an advantage as it encourages clear thought about the purpose of the analysis.

The use of an interface, such as the one we have described, turns the statistical model into a black box: the user can analyse data without knowing or caring what goes on inside the box. Those who have invested time into exploring the contents of this particular black box will see this as a disadvantage, and indeed something is lost by not having the full conventional output. But in our view the advantages stemming from the clarity of purpose imposed by the interface far outweigh this loss, and of course there is nothing to stop the user from returning to a stand–alone mode and fitting a conventional statistical model if this is desired.

References

[1] K. J. Rothman. *Modern Epidemiology.* Little, Brown and Company, Boston/Toronto, 1986.

[2] J.P. Vandenbroucke. Should we abandon statistical modelling altogether? *American Journal of Epidemiology*, 126(1):10–13, 1987.

[3] J.N. Morris, J.W. Marr, and D.G. Clayton. Diet and heart: a postscript. *British Medical Journal*, 19 November(2):1307–14, 1977.

[4] D.G. Clayton and M. Hills. *Statistical Models in Epidemiology.* Oxford University Press, Oxford, 1993.