The Stata Journal (2007)
7, Number 2, pp. 183–196

Two postestimation commands for assessing confounding effects in epidemiological studies

Zhiqiang Wang
School of Medicine and School of Population Health
University of Queensland
Brisbane, Queensland, Australia
z.wang@uq.edu.au

Abstract. Confounding is a major issue in observational epidemiological studies. This paper describes two postestimation commands for assessing confounding effects. One command (`confall`) displays and plots all possible effect estimates against one of *p*-value, Akaike information criterion, or Bayesian information criterion. This computing-intensive procedure allows researchers to inspect the variability of the effect estimates from various possible models. Another command (`chest`) uses a stepwise approach to identify variables that have substantially changed the effect estimate. Both commands can be used after most common estimation commands in epidemiological studies, such as logistic regression, conditional logistic regression, Poisson regression, linear regression, and Cox proportional hazards models.

Keywords: st0124, confall, confgr, chest, epidemiological methods, confounding, all possible effects, change in estimate

1 Introduction

Confounding is a bias introduced by the imbalanced distribution of extraneous risk factors among comparison groups. The issue of assessing confounding effects has been discussed in several papers, and various methods of controlling for confounding effects have been proposed (Miettinen 1974, Schlesselman 1978, Greenland and Robins 1985, Debanne and Sokol 1986, Greenland and Robins 1986, Grayson 1987, Weinberg 1993, Schwartz and Coull 2003, Steenland and Greenland 2004, Sturmer et al. 2005).

A common practice for assessing confounding is to use either stratification or multiple regression methods to compare the crude with adjusted effect estimates. However, such comparisons, if possible, can be labor intensive because of many possible combinations of potential confounders. It has been suggested that one should adjust for all variables believed a priori to be potential confounders, regardless of their properties in the current data. However, this approach sometimes can be subjective and lacks transparency. Particularly, when the number of potential confounders is large and the sample size is small, this approach can result in an estimate with poor precision. Like most model searching methods, a stepwise regression procedure focuses on identifying the predictors of the dependent variable. Some true confounders might not be identified, especially when sample size is small. This paper presents two practical tools, Stata postestimation commands, to help researchers better understand the presence and direction of possible confounding effects in their data.

One is the all-possible-estimates method (`confall`) and the other is the change-in-estimates method (`chest`). The all-possible-estimates method allows the user to inspect possible estimates with many different adjustments. It is useful in understanding the nature of confounding, but this method should not be used to select the final model in a particular dataset. The change-in-estimates method selects variables in a stepwise fashion according to the magnitude of the differences between adjusted and unadjusted effect estimates. Inspecting changes in estimates, as more variables are adjusted, is useful in understanding the nature of confounding and the joint confounding by multiple variables.

2 The confall command displays and plots all possible effect estimates

`confall` is a postestimation command that calculates and displays all possible effect estimates ($2^n - 1$ adjusted estimates plus one crude estimate, where n is the number of total potential confounders). It plots all effect estimates against one set of the following values: p-value, Akaike information criterion (AIC; Akaike 1974), Bayesian information criterion (BIC; Schwarz 1978), confidence interval range, and the number of confounders R^2 or pseudo-R^2 values from the corresponding models. The `confgr` command produces the same plots as `confall` does, directly using the saved results.

This `confall` command can be used after most commonly used estimation commands in epidemiological studies, such as logistic regression, conditional logistic regression, Poisson regression, linear regression, and the Cox proportional hazards model.

2.1 Syntax

`confall` *varname* [, `eform`[(*string*)] `xis(`*string*`)` `table` `showvar(`*string*`)`
`lockterms(`*varlist*`)` `addaic` `addbic` `savefile(`*string*`)` `level(`#`)` `format(%`*fmt*`)`
`xformat(%`*fmt*`)` `mostn(`#`)` `nograph` *graph_options*]

The *varname* is the exposure of interest, and all other independent variables in the original model are potential confounders.

`confgr using` *filename* [, `eform(`*string*`)` `xis(`*string*`)` `format(%`*fmt*`)` `addaic`
`addbic` *graph_options*]

The *filename* is the result file saved using the `savefile()` option in the previous `confall` command.

2.2 Options

`eform`[`(`*string*`)`] reports exponentiated coefficients. *string* can be used to label these exponentiated coefficients as odds ratios, hazard ratios, or relative-risk ratios, depending on the estimation command. Confidence intervals are similarly transformed.

`xis(`*string*`)` specifies the x axis. The default is `xis(pr)`, representing p-values from likelihood-ratio tests. For `xis(pr)`, the default x axis is rescaled according to $X = p^{\{\ln(0.5)/\ln(0.05)\}}$, although the original p-values are labeled on the axis. If the user wants to plot the original scale, use option `xis(pr)`. Alternatives include `aic` (AIC), `bic` (BIC), `r2` (R^2 or pseudo-R^2), `n` (number of confounders), and `civ` (confidence interval range).

`table` shows a table of all effect estimates.

`showvar(`*string*`)` specifies a variable so that the effect estimates from models including this particular variable are shown in a different symbol color (or pattern) from others. This option is not available in `confgr`.

`lockterms(`*varlist*`)` specifies variables to be included in all models.

`addaic` and `addbic` mark the effect estimates from models with the minimum AIC and BIC, respectively.

`savefile(`*filename* [`, replace`]`)` saves a dataset of all effect estimates as *filename*; use `replace` to overwrite an existing filename.

`level(`*#*`)` specifies the confidence level, as a percentage, for confidence intervals. The default is `level(95)` or as set by `set level`; see [U] **20.6 Specifying the width of confidence intervals**.

`format(%`*fmt*`)` specifies the display format for presenting effect estimates in the graph and table. The default is `format(%9.0g)`.

`xformat(%`*fmt*`)` specifies the display format for presenting values on the x axis. The default is `xformat(%9.0g)`.

`mostn(`*#*`)` specifies the maximum number of potential confounders allowed in a model. This option can be useful in situations with many potential confounders.

`nograph` suppresses the graph.

graph_options refers to any of the options documented in [G] **graph twoway scatter**.

2.3 Example 1: A positive association with some confounding

To examine the association between body mass index (BMI) and type 2 diabetes with an example dataset, we performed logistic regressions with diabetes as the dependent variable (`1` for yes and `0` for no) and BMI as the exposure of interest. Variables considered potential confounders were age, sex, impaired glucose tolerance (IGT), serum total cholesterol, diastolic blood pressure, C-reactive protein, gamma glutamyltransferase (GGT), albuminuria, smoking, and drinking.

First, we run an estimation command (`logistic`).

```
. use diabdata

. logistic diabetes BMI Age Sex IGT ACR CRP Cholesterol DiastolicBP GGT
> Smoking Drinking
  (output omitted)
```

Then we run the `confall` command, which specified BMI as the exposure of interest. The lockterms specified include `Age` and `Sex` in all models. The x axis is p-value, which is rescaled using the `xis(pr)` option. Because we have little interest in large p-values, especially those values larger than 0.5, a relatively smaller space is allocated to the same distance in larger p-values on the x axis. The original p-values are labeled on the axis, but the actual axis is rescaled according to $x = p^{.23137821}$ (or $p^{\ln(0.5)/\ln(0.05)}$). We use options `addaic` and `addbic` to distinguish the effect estimates from "best" models according to AIC and BIC, respectively.

```
. confall BMI, eform(Odds ratio between diabetes and BMI) addaic addbic xis(pr)
> lockterms(Age Sex) ylabel(0.8 1(.5)3) yline(1, lp(dash)) savefile(bmiresults)
fitting models........
drawing graph ...

 Age Sex in all models
256 sets of confounders
Outcome variable: diabetes      Exposure: BMI
file bmiresults.dta.saved
```

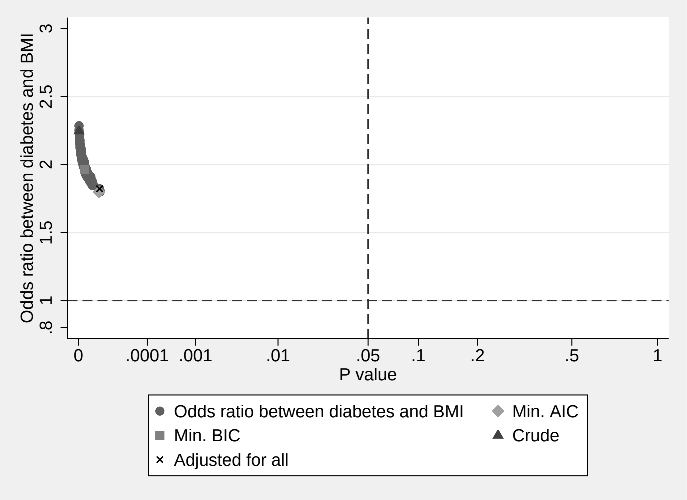

Figure 1: All possible odds ratios and corresponding p-values from likelihood-ratio tests

After running the logistic regression with BMI and other potential confounding factors in the model, `confall` calculated all possible effect estimates and plotted those estimates against p-values of the BMI variable in the corresponding models, as shown in figure 1. All odds ratios between BMI and diabetes were significantly higher than the null effect, above the horizontal null effect line (odds ratio = 1) and on the left side of the vertical α value line ($\alpha = 0.05$). The odds ratios range from 1.8 to 2.28 with the adjustments for different subsets of potential confounding factors. Although the differences among estimates indicated the presence of some degree of confounding, such confounding would not alter the conclusion that higher BMI level is associated with higher risk of diabetes. Even if the same data were analyzed by different researchers using different model selection methods, their findings would probably be consistent.

Similarly, all possible effect estimates can be plotted against one of the other values. Figure 2, produced using the following command, is an example of using AIC as the x axis.

```
. confall BMI, eform(Odds ratio) addaic addbic xis(aic) lockterms(Age Sex)
> ylabel(0.8 1(.5)3) yline(1, lp(dash))
fitting models........
drawing graph ...
 Age Sex in all models
256 sets of confounders
Outcome variable: diabetes      Exposure: BMI
```

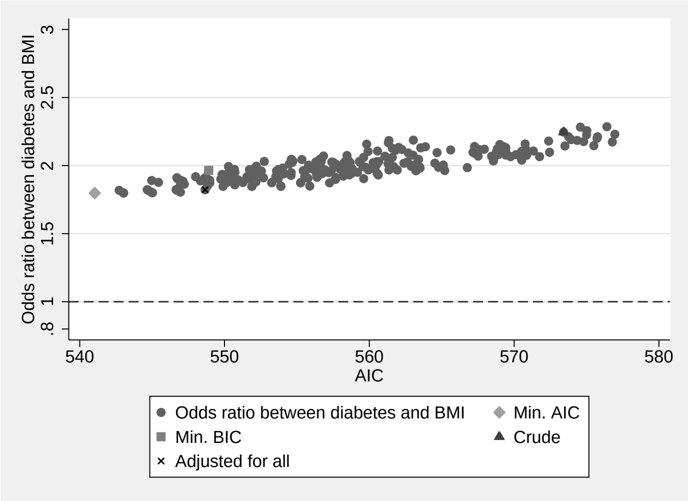

Figure 2: All possible odds ratios and their corresponding AIC

Other features: all possible effect estimates can be displayed as a table in the Stata Results window by using the `table` option. Those estimates can also be saved into a Stata file by using the `savefile(`*filename*`)` option. We can use the `showvar(`*string*`)` option to distinguish the effect estimates from the models with the specified variable in them. In the above example, if the variable `weight` had mistakenly been added as a potential confounder, we could identify large p-values from those models with variable weight, as shown in figure 3. Since body weight is a part of the BMI calculation (weight/height^2), and to a certain degree measures the same construct (obesity) as BMI, it should not be taken as a confounder.

```
. logistic diabetes BMI Age Sex IGT ACR CRP Cholesterol DiastolicBP GGT
> Smoking Drinking weight
  (output omitted)
. confall BMI, eform(Odds ratio) ylabel(0.8 1(.5)3) yline(1, lp(dash)) xis(pr)
> lockterms(Age Sex) showvar(weight)
fitting models.........
drawing graph ...
 Age Sex in all models
512 sets of confounders
Outcome variable: diabetes      Exposure: BMI
```

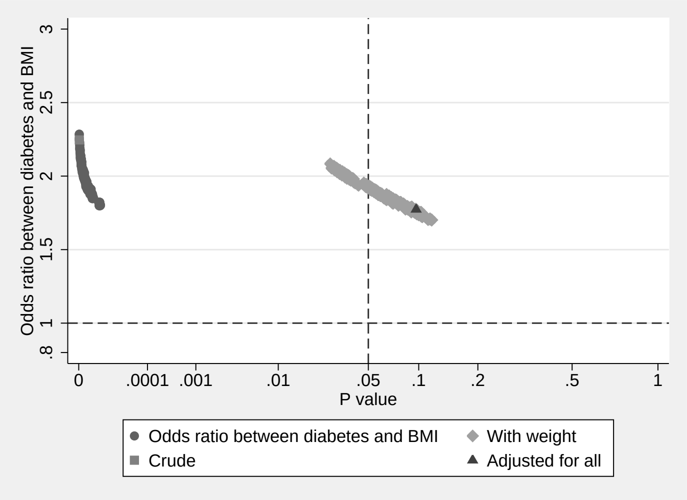

Figure 3: All possible odds ratios and corresponding p-values

2.4 Example 2: No association with little confounding

Using the same dataset, we generated a random variable—`RandomVar`, which was taken as the exposure of interest. We can use the `confall` command to produce a graph, figure 4, of the odds ratios that are close to the noneffect line, and the p-values are large. With a plot like this, researchers can confidently report their findings that there is no evidence of association between this variable and diabetes.

```
. logistic diabetes RandomVar BMI Age Sex IGT ACR CRP Cholesterol DiastolicBP
> GGT Smoking Drinking
. confall RandomVar, eform(Odds ratio) ylabel(0.8(.2)1.6) yline(1, lp(dash))
> xis(pr) lockterms(Age Sex)
fitting models.........
drawing graph ...
 Age Sex in all models
512 sets of confounders
Outcome variable: diabetes      Exposure: RandomVar
```

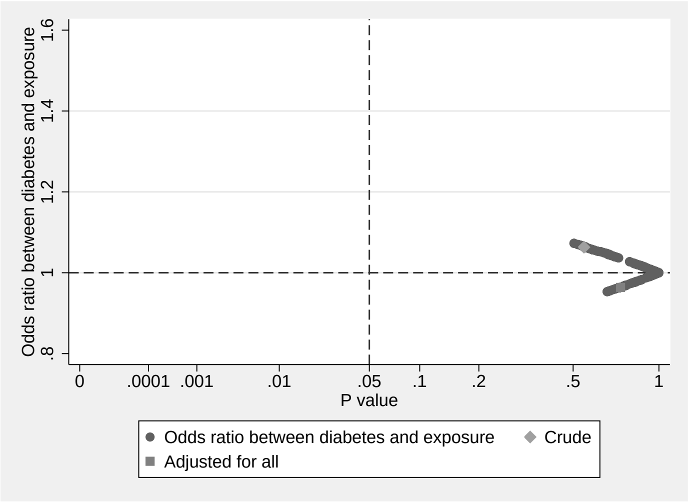

Figure 4: All possible odds ratios and corresponding p-values

2.5 Example 3: Crude estimate shows significant association with possible confounding

In this example, we used the same data as in that for the previous examples; however, the variable GGT was taken as the exposure of interest. `confall` calculated and plotted all possible odds ratios for diabetes corresponding to a 1−standard deviation increase in GGT. As shown in figure 5, the crude and some adjusted odds ratios were significantly higher than the null effect 1, whereas other adjusted odds ratios are closer to 1 with large p-values without significant associations. The observed pattern in figure 5 indicates the imbalanced distribution of other risk factors among participants with different GGT values, as well as the possible presence of confounding.

(*Continued on next page*)

```
. logistic diabetes BMI Age Sex IGT ACR CRP Cholesterol DiastolicBP GGT
> Smoking Drinking
. confall GGT, eform(Odds ratio)  ylab(0.8(.2)1.6) yline(1, lp(dash)) xis(pr)
> lockterms(Age Sex) addaic addbic
fitting models........
drawing graph ...

 Age Sex in all models
256 sets of confounders
Outcome variable: diabetes      Exposure: GGT
```

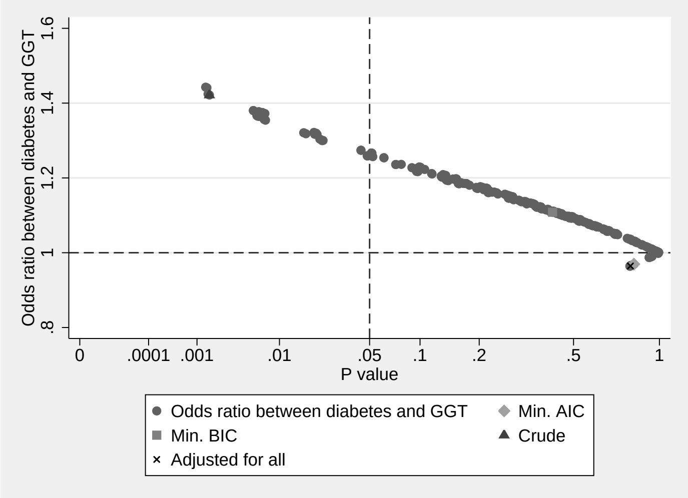

Figure 5: All possible odds ratios and corresponding p-values

`confgr` generates plots from a saved file. To save computing time, users can explore different plots directly by using the saved file. For example, when generating figure 1, we saved a file named `bmiresults` with the `savefile(bmiresults)` option. To reproduce figure 1, we used the following command:

```
. confgr using bmiresults, eform(Odds ratio between diabetes and BMI) addaic
> addbic ylabel(0.8 1(.5)3) yline(1, lp(dash))
```

Using the same saved file, we can also reproduce figure 2:

```
. confgr using bmiresults, xis(aic) eform(Odds ratio between diabetes and BMI)
> addaic addbic ylabel(0.8 1(.5)3) yline(1, lp(dash))
```

3 The chest command displays and plots the change-in-effect estimates

The change-in-estimates method has been suggested in several epidemiology textbooks (Kleinbaum, Kupper, and Morgenstern 1982; Rothman and Greenland 1998). I previously wrote a program to perform this task (Wang 2000), particularly for assessing multiple confounders. `chest` is a similar postestimation command that is easier to use than its previous version. It selects variables in a stepwise fashion. One potential confounder at a time is included in the model by using either a forward or backward strategy. At step 1, using the forward strategy, the variable is included in the model because its adjustment causes the largest change in the effect measurement. At step 2, the variable that causes the largest change among the remaining variables is included. This process continues until all variables are added in the model. Therefore, the `chest` command takes much less computing time than the `confall` command does. Instead of fitting 2^n models in `confall`, `chest` needs only to fit $1+2+3+\cdots+n$ models.

3.1 Syntax

`chest` *varname* [, `eform`[`(`*string*`)`] `lockterms(`*varlist*`)` `format(%`*fmt*`)` `mostn(`#`)` `backward` `vertic` `level(`#`)` `nograph` *graph_options*]

3.2 Options

`eform`[`(`*string*`)`] reports exponentiated coefficients. *string* can be used to label these exponentiated coefficients as odds ratios, hazard ratios, or relative-risk ratios, depending on the estimation command. Confidence intervals are similarly transformed.

`lockterms(`*varlist*`)` specifies variables to be included in all models.

`format(%`*fmt*`)` specifies the display format for presenting effect estimates in the graph and table. The default is `format(%9.0g)`.

`mostn(`#`)` specifies the maximum number of potential confounders allowed in a model. This option is not available with the `backward` option.

`backward` specifies a backward selection method. The default is a forward selection method.

`vertic` specifies a vertical spike plot. A horizontal spike plot is the default.

`level(`#`)` specifies the confidence level, as a percentage, for confidence intervals. The default is `level(95)` or as set by `set level`; see [U] **20.6 Specifying the width of confidence intervals**.

`nograph` suppresses the graph.

graph_options refers to any of the options documented in [G] **graph twoway scatter**.

(*Continued on next page*)

3.3 Example 4: A positive association with some confounding

We can use the variable name `BMI` after `chest` to specify the exposure of interest. The forward approach starts in a model without any potential confounding variables. In the following example, the initial model includes only the exposure of interest (`BMI`) and the variables (`Age` and `Sex`) specified by `lockterms(`*varlist*`)`.

```
. logistic diabetes BMI Age Sex IGT ACR CRP Cholesterol DiastolicBP GGT
> Smoking Drinking
  (output omitted)
. chest BMI, eform("Odds ratio between diabetes and BMI") lockterms(Age Sex)
> format(%6.2f) xline(1, lp(dash)) xlabel(.8 1(.5)3)
Change-in-estimate
logistic regression.                 Outcome:  diabetes
number of obs = 714                  Exposure: BMI

Variables
added          Odds ratio b~   [95% Conf.     Interval]      Change, %

Initial model       2.25          1.82           2.78
         +ACR       2.02          1.62           2.52           -9.95
         +CRP       1.93          1.55           2.40           -4.71
 +Cholesterol       1.85          1.48           2.31           -4.16
         +IGT       1.80          1.43           2.26           -2.65
     +Smoking       1.82          1.44           2.30            1.15
         +GGT       1.82          1.44           2.30            0.23
 +DiastolicBP       1.82          1.44           2.30           -0.05
    +Drinking       1.82          1.44           2.31            0.03

 Age Sex in all models
```

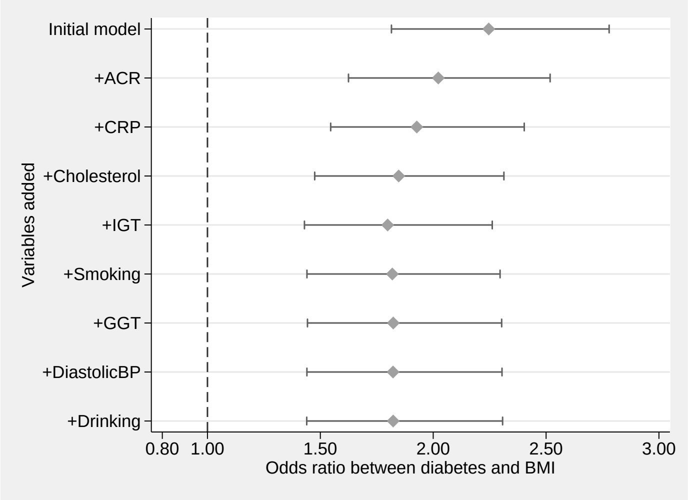

Figure 6: Odds ratio for diabetes corresponding to a 1–standard deviation increase in BMI

One variable that causes the largest change in the effect estimate among candidate variables was added into the model until all potential confounding variables were entered. On the other hand, the backward approach would have started with a full model, and the least important confounder for the change in estimate would have been removed from the model at each step until all potential confounders had been removed. We used the forward selection method in this example to examine the association between BMI and diabetes. After adding `ACR` and `CRP`, the odds ratio between BMI and diabetes became smaller. Adding other variables made few changes in the effect estimate. Again, the odds ratio between diabetes and BMI remains high at all steps (figure 6). Therefore, the confounding effects from those variables would not alter the conclusion that higher BMI is associated with a higher risk of diabetes.

3.4 Example 5: Crude estimate showing a significant association with possible confounding

In this example, examining the association between `GGT` and diabetes, the odds ratio changed substantially after adding two variables: `Cholesterol` and `BMI` (figure 7), suggesting the crude association is more likely because of confounding from those variables.

```
. logistic diabetes BMI Age Sex IGT ACR CRP Cholesterol DiastolicBP GGT
> Smoking Drinking
  (output omitted)
. chest GGT, forward eform("Odds ratio between diabetes and GGT")
> lockterms(Age Sex) format(%6.2f) xline(1, lp(dash)) xlabel(.8(.2)1.8)
Change-in-estimate
logistic regression.                 Outcome:  diabetes
number of obs = 714                  Exposure: GGT

Variables      
added          Odds ratio b~    [95% Conf.     Interval]      Change, %

Initial model        1.42          1.14          1.77
 +Cholesterol        1.20          0.95          1.53         -15.44
         +BMI        1.09          0.85          1.41          -9.03
         +ACR        1.03          0.79          1.34          -5.59
         +IGT        0.99          0.76          1.30          -4.05
         +CRP        0.97          0.74          1.27          -2.14
    +Drinking        0.96          0.73          1.28          -0.56
     +Smoking        0.96          0.73          1.28           0.09
 +DiastolicBP        0.96          0.73          1.28          -0.04

 Age Sex in all models
```

(*Continued on next page*)

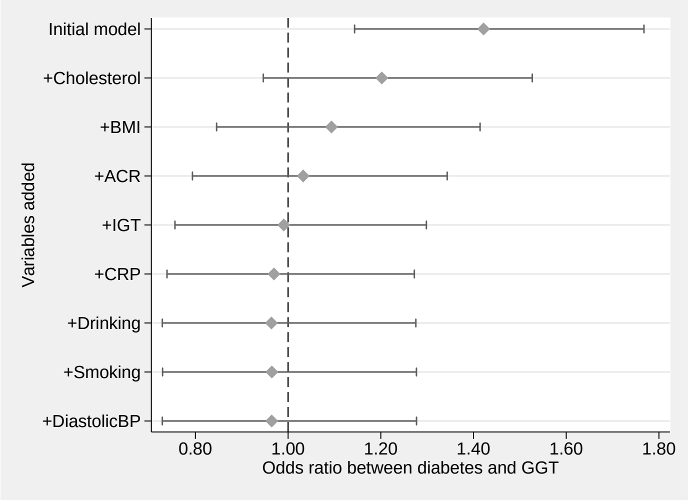

Figure 7: Odds ratio for diabetes corresponding to a 1−standard deviation increase in GGT

The `chest` command also displays a table, containing effect estimates, their 95% confidence intervals, and the changes in percentage at all steps. Both the table and the graph provide us information on the presence and direction of confounding as well as the important confounding variables.

4 Comments

The two commands `confall` and `chest` are designed to help researchers understand the uncertainty of effect estimates. When all possible effect estimates are similar, researchers are confident with their conclusions regardless of the methods used for selecting a satisfactory model. On the other hand, when effect estimates differ substantially, a careful examination and identification of confounding factors are warranted.

Those programs are only tools to examine effect estimates from many possible models. Such examination can be time consuming if each model is fitted individually. However, those tools are not a substitute for carefully incorporating available knowledge to select confounding factors at the design stage or for careful data analysis to identify an appropriate nonlinear relationship.

Other important aspects such as chance and information (measurement) bias can also influence the patterns of all-possible-estimates and change-in-estimates plots. The commands can examine confounding effects only of variables that have been collected. Confounding may exist from unmeasured variables even though both `confall` and `chest` suggest no confounding from the variables included in the analysis.

Adding a variable that measures the same construct as the exposure of interest or is an intermediate between exposure and disease outcome can substantially distort the true association. The `confall` or `chest` command will show a similar pattern as that when confounding is present. However, this design issue should be addressed before analysis.

Although this report used only logistic regression, the programs can be applied with most commonly used estimation commands in epidemiological studies such as those for the Cox proportional hazards model, Poisson, and conditional logistic regressions. For both `confall` and `chest`, the exposure of interest can be either a continuous or dichotomous variable, and potential confounding variables can be any type of variable. Categorical potential confounding variables need to be identified using the Stata `xi` command in the estimation command. When an interaction term is used, the interaction term and main effect term are treated as separate terms. Programs for systematically assessing interaction terms (effect modification) should be developed separately.

5 Acknowledgments

This work was funded by the National Health and Medical Research Council (NH&MRC) of Australia (grant numbers 301024 and 320860).

I received suggestions at the 12th UK Stata Users Group Meeting from Ian White (MRC Biostatistics Unit, Cambridge) to add the `savefile()` and `showvar()` options; Nick Cox (Durham University) to rescale the x axis; and Patrick Royston (MRC Clinical Trial Unit, London) to assess nonlinear relations, which may be incorporated in future versions. The reviewer suggested several improvements. I thank Wendy Hoy (University of Queensland) for her support.

6 References

Akaike, H. 1974. A new look at statistical model identification. *IEEE Transactions on Automatic Control* AC-19: 716–723.

Debanne, S. M., and R. J. Sokol. 1986. Adjusting for confounding with statistical software packages. *American Journal of Perinatology* 3: 133–134.

Grayson, D. A. 1987. Confounding confounding. *American Journal of Epidemiology* 126: 546–553.

Greenland, S., and J. M. Robins. 1985. Confounding and misclassification. *American Journal of Epidemiology* 122: 495–506.

———. 1986. Identifiability, exchangeability, and epidemiological confounding. *International Journal of Epidemiology* 15: 413–419.

Kleinbaum, D. G., L. L. Kupper, and H. Morgenstern. 1982. *Epidemiologic Research: Principles and Quantitative Methods.* New York: Nostrand Reinhold.

Miettinen, O. 1974. Confounding and effect-modification. *American Journal of Epidemiology* 100: 350–353.

Rothman, K., and S. Greenland. 1998. *Modern Epidemiology*. 2nd ed. Philadelphia: Lippincott–Raven.

Schlesselman, J. J. 1978. Assessing effects of confounding variables. *American Journal of Epidemiology* 108: 3–8.

Schwartz, J., and B. A. Coull. 2003. Control for confounding in the presence of measurement error in hierarchical models. *Biostatistics* 4: 539–553.

Schwarz, G. 1978. Estimating the dimension of a model. *Annals of Statistics* 6: 461–464.

Steenland, K., and S. Greenland. 2004. Monte Carlo sensitivity analysis and Bayesian analysis of smoking as an unmeasured confounder in a study of silica and lung cancer. *American Journal of Epidemiology* 160: 384–392.

Sturmer, T., S. Schneeweiss, J. Avorn, and R. J. Glynn. 2005. Adjusting effect estimates for unmeasured confounding with validation data using propensity score calibration. *American Journal of Epidemiology* 162: 279–289.

Wang, Z. 2000. sbe27: Assessing confounding effects in epidemiological studies. *Stata Technical Bulletin* 49: 12–15. Reprinted in *Stata Technical Bulletin Reprints*, vol. 9, pp. 134–138. College Station, TX: Stata Press.

Weinberg, C. R. 1993. Toward a clearer definition of confounding. *American Journal of Epidemiology* 137: 1–8.

About the author

Zhiqiang Wang is a senior research fellow in the School of Medicine and School of Population Health at the University of Queensland, Brisbane, Australia. His research interests are mainly in the epidemiology and prevention of noncommunicable diseases.