The Stata Journal

The *Stata Journal* publishes reviewed papers together with shorter notes or comments, regular columns, book reviews, and other material of interest to Stata users. Examples of the types of papers include 1) expository papers that link the use of Stata commands or programs to associated principles, such as those that will serve as tutorials for users first encountering a new field of statistics or a major new technique; 2) papers that go "beyond the Stata manual" in explaining key features or uses of Stata that are of interest to intermediate or advanced users of Stata; 3) papers that discuss new commands or Stata programs of interest either to a wide spectrum of users (e.g., in data management or graphics) or to some large segment of Stata users (e.g., in survey statistics, survival analysis, panel analysis, or limited dependent variable modeling); 4) papers analyzing the statistical properties of new or existing estimators and tests in Stata; 5) papers that could be of interest or usefulness to researchers, especially in fields that are of practical importance but are not often included in texts or other journals, such as the use of Stata in managing datasets, especially large datasets, with advice from hard-won experience; and 6) papers of interest to those who teach, including Stata with topics such as extended examples of techniques and interpretation of results, simulations of statistical concepts, and overviews of subject areas.

The *Stata Journal* is indexed and abstracted by *CompuMath Citation Index*, *Current Contents/Social and Behavioral Sciences*, *RePEc: Research Papers in Economics*, *Science Citation Index Expanded* (also known as *SciSearch*, *Scopus*, and *Social Sciences Citation Index*.

For more information on the *Stata Journal*, including information for authors, see the webpage

http://www.stata-journal.com

Subscriptions are available from StataCorp, 4905 Lakeway Drive, College Station, Texas 77845, telephone 979-696-4600 or 800-STATA-PC, fax 979-696-4601, or online at

http://www.stata.com/bookstore/sj.html

Subscription rates listed below include both a printed and an electronic copy unless otherwise mentioned.

U.S. and Canada		Elsewhere	
Printed & electronic		**Printed & electronic**	
1-year subscription	$ 98	1-year subscription	$138
2-year subscription	$165	2-year subscription	$245
3-year subscription	$225	3-year subscription	$345
1-year student subscription	$ 75	1-year student subscription	$ 99
1-year university library subscription	$125	1-year university library subscription	$165
2-year university library subscription	$215	2-year university library subscription	$295
3-year university library subscription	$315	3-year university library subscription	$435
1-year institutional subscription	$245	1-year institutional subscription	$285
2-year institutional subscription	$445	2-year institutional subscription	$525
3-year institutional subscription	$645	3-year institutional subscription	$765
Electronic only		**Electronic only**	
1-year subscription	$ 75	1-year subscription	$ 75
2-year subscription	$125	2-year subscription	$125
3-year subscription	$165	3-year subscription	$165
1-year student subscription	$ 45	1-year student subscription	$ 45

Back issues of the *Stata Journal* may be ordered online at

http://www.stata.com/bookstore/sjj.html

Individual articles three or more years old may be accessed online without charge. More recent articles may be ordered online.

http://www.stata-journal.com/archives.html

The *Stata Journal* is published quarterly by the Stata Press, College Station, Texas, USA.

Address changes should be sent to the *Stata Journal*, StataCorp, 4905 Lakeway Drive, College Station, TX 77845, USA, or emailed to sj@stata.com.

Users of any of the software, ideas, data, or other materials published in the *Stata Journal* or the supporting files understand that such use is made without warranty of any kind, by either the *Stata Journal*, the author, or StataCorp. In particular, there is no warranty of fitness of purpose or merchantability, nor for special, incidental, or consequential damages such as loss of profits. The purpose of the *Stata Journal* is to promote free communication among Stata users.

The *Stata Journal* (ISSN 1536-867X) is a publication of Stata Press. Stata, stata, Stata Press, Mata, mata, and NetCourse are registered trademarks of StataCorp LP.

The Stata Journal (2013)
13, Number 3, pp. 492–509

Computing adjusted risk ratios and risk differences in Stata

Edward C. Norton
Departments of Health Management & Policy and Economics
University of Michigan
Ann Arbor, MI
and National Bureau of Economic Research
ecnorton@umich.edu

Morgen M. Miller
Departments of Health Management & Policy and Economics
University of Michigan
Ann Arbor, MI
mmmill@umich.edu

Lawrence C. Kleinman
Departments of Health Evidence & Policy and Pediatrics
Icahn School of Medicine at Mount Sinai
New York, NY
lawrence.kleinman@mssm.edu

Abstract. In this article, we explain how to calculate adjusted risk ratios and risk differences when reporting results from logit, probit, and related nonlinear models. Building on Stata's `margins` command, we create a new postestimation command, `adjrr`, that calculates adjusted risk ratios and adjusted risk differences after running a logit or probit model with a binary, a multinomial, or an ordered outcome. `adjrr` reports the point estimates, delta-method standard errors, and 95% confidence intervals and can compute these for specific values of the variable of interest. It automatically adjusts for complex survey design as in the fit model. Data from the Medical Expenditure Panel Survey and the National Health and Nutrition Examination Survey are used to illustrate multiple applications of the command.

Keywords: st0306, adjrr, risk ratio, adjusted risk ratio, risk difference, adjusted risk difference, odds ratio, logistic, logit, probit, multinomial, ordered

1 Introduction

Researchers often fit logit models when the dependent variable is dichotomous. Because the coefficients from logit models are, on their own, hard to interpret, researchers frequently report their results using statistics generated from those coefficients, often odds ratios. It is well known, however, that most people misinterpret odds ratios as risk ratios (Klaidman 1990; Teuber 1990; Altman, Deeks, and Sackett 1998; Bier 2001). When the risk of the outcome is high, these two measures diverge with the odds ratio being further from 1.0 than the risk ratio. For these and other reasons, many people have called for researchers to report risk ratios instead of odds ratios (for example, Greenland [1987]; Spiegelman and Hertzmark [2005]; Cummings [2009a]).

The search for the best way to estimate risk ratios has shown that these statistics can be estimated in a number of ways from different kinds of models (for example, Flanders and Rhodes [1987]; Greenland and Holland [1991]; Greenland [2004]). Kleinman and Norton (2009) propose a simple and intuitive formula for the risk ratio, adjusted for covariates. For models with categorical covariates, the adjusted risk ratio (ARR) reproduces Mantel–Haenszel results. Kleinman and Norton (2009) also demonstrate that their method is correct given the distribution of covariates, including complex specifications with continuous variables, and is robust in many cases.

This article makes several contributions. First, we show how to compute an ARR and an adjusted risk difference (ARD) in Stata not only for logit models but also for other related models. These other models include the multinomial logit, ordered logit, probit, multinomial probit, and ordered probit models. This shows that this approach applies generally to models with binary or categorical dependent variables. In addition, it is easy and fast to calculate these statistics in Stata because our command builds on the `margins` command. However, our command makes it much easier than `margins`, especially for multinomial and ordered models. Second, in addition to the ARR, we compute the ARD. This statistic can be useful because it shows the predicted difference in percentage point (or absolute) terms, which is sometimes of interest. Third, because it is always important to report the level of uncertainty along with any estimated statistic, our new command estimates delta-method standard errors. Again, because we build on the `margins` command, estimating delta-method standard errors is fast and easy. Fourth, the command will compute the two statistics for any two values of the variable of interest, not only 0 and 1 (the default values). Although the variable of interest is often binary, one could be interested in comparing probabilities for two different values of, say, age. Fifth, we show how to compute all of these when the researcher wants to control for complex survey design or robust standard errors. Large, representative datasets often have sampling weights, clustered observations, and stratification. These can be taken into account when computing ARRs and ARDs.

2 Methods

2.1 Estimating ARRs and ARDs

The ARR and ARD are two ways to express the relationship between two predicted probabilities based on the fit model and a set of observations. One is the predicted probability when the variable of interest equals 1; the other is the predicted probability when the variable of interest equals 0 (more generally, pick any two values of the variable). These predicted probabilities are then averaged over the entire dataset (or perhaps an interesting subset of the data). The ARR is the ratio of the mean predicted probabilities, and the ARD is the difference of the mean predicted probabilities. The ARD is sometimes called the average treatment effect because it compares the effect of a change in the variable of interest (the treatment) for all observations. All of these probabilities, and functions of probabilities, are easily calculated from logit or probit models through simple algebraic manipulations.

For example, consider the probability of mortality within a year of treatment for a population of patients, some of whom were randomly given a new drug. After fitting the model, compute two predicted probabilities of mortality for each observation, one assuming the patient did get the drug and the other assuming the patient did not. The key is to hold all other covariates at their original values so that the only difference in predicted probability is attributable to the new drug.

We begin with the simplest case, where the variable of interest is binary, the population of interest is the entire sample, and the model (logit or probit) is for a dichotomous outcome. Let P_1 be the mean of the predicted probabilities that the dependent variable y equals 1, computed over the whole sample, with the variable of interest x set equal to 1 and all other covariates X (including the constant term) equal to their original values. Therefore, the probability is a function of the linear index $\beta_x x + X\beta$. Let P_0 be defined in a corresponding way, but with x set to 0.

$$P_1 = \frac{1}{N}\sum_{i=1}^{N}\Pr(y_i = 1|X, x = 1) \tag{1}$$

$$P_0 = \frac{1}{N}\sum_{i=1}^{N}\Pr(y_i = 1|X, x = 0) \tag{2}$$

Then the ARR is the ratio P_1/P_0 and the ARD is the difference $P_1 - P_0$.

There are three ways to generalize the above approach. First, allow the variable of interest x to take on any two policy relevant values, not just 0 and 1. For a continuous variable, 0 and 1 may not be appropriate values of comparison. It might be of policy interest, for example, to compare predicted outcomes for persons aged 85 and persons aged 65, holding all else constant. Our new Stata command allows the user to specify the range of values for the variable of interest.

Second, compute the statistic for a subset of the analysis data. Then the average probabilities would not be computed over the entire sample of size N but for a subset of interest. Our new Stata command allows the user to compute the ARR and ARD for a subset of the data, for example, just for women or just for those with comorbidities.

Third, (1) and (2) can be modified to allow weights, as is often the case for complex survey design. Instead of a simple average, one would compute a weighted average. Our new Stata command automatically incorporates weights from a complex survey design into the formulas for the ARR and ARD. Furthermore, the estimated standard errors also automatically take into account stratification and clustering.

We can incorporate all three of these generalizations into (1) and (2) by conditioning on general values of x, averaging over a subset of the data, and allowing weights.

$$P_A = \frac{1}{n}\sum_{i=1}^{n} \Pr(y_i = 1|X, x = A)\omega_i \quad (3)$$

$$P_B = \frac{1}{n}\sum_{i=1}^{n} \Pr(y_i = 1|X, x = B)\omega_i \quad (4)$$

In (3) and (4), A and B represent any two values at which to estimate the predicted probabilities, n represents the sample size of the subsample of interest, and ω_i represents the weights associated with the complex survey design.

Finally, we can adjust the definition of probability to be appropriate for models with more than two outcomes such that the dependent variable y equals 1 (as opposed to 0). While the above formulas work well for dichotomous outcomes (logit and probit), they need to be modified for multinomial and ordered models. Our new Stata command allows the user to compute the ARR and ARD for binary, multinomial, and ordered outcomes. The following subsections show the specific formulas for these models that extend the basic framework.

2.2 Logit model

The computation of the probabilities in (1) and (2) depends on the specific model. In the logit model, the estimated coefficients are transformed to probabilities through the logistic function. For the logit model, the formula for the probability that y equals 1 is the logistic cumulative distribution function:

$$\Pr(y = 1|X, x) = \frac{1}{1 + e^{-(\beta_x x + X\beta)}}$$

2.3 Probit model

The probit model is a common alternative to the logit for binary outcomes. For the probit model, the formula for the probability that y equals 1 is the normal cumulative distribution function:

$$\Pr(y = 1|X, x) = \Phi(\beta_x x + X\beta)$$

There is no substantive difference between simple logit and probit models; the choice between them is largely a matter of personal preference. The magnitude of the coefficients is quite different. The coefficients in the probit are predictably smaller by a factor of about 0.6. However, predicted probabilities—and therefore statistics like the ARR and ARD—are always nearly identical.

2.4 Multinomial models

Multinomial models have three or more outcomes that are discrete and not ordered (for example, the choice of mode of transportation or choice of major in college). For the multinomial logit, the formula for the probabilities of each possible outcome j, for $j = 1$ to $J - 1$ (the Jth category has its coefficient normalized to 0), is as follows:

$$\Pr(y = j, j \neq J|X, x) = \frac{e^{(\beta_x^j x + X\beta^j)}}{\sum e^{(\beta_x^j x + X\beta^j)} + 1}$$

For the Jth category, the predicted probability is

$$\Pr(y = J|X, x) = \frac{1}{\sum e^{(\beta_x^J x + X\beta^J)} + 1}$$

The formulas are similar for the multinomial probit model, but the cumulative normal replaces the cumulative logistic function.

2.5 Ordered models

Ordered models have three or more outcomes that are ordered. Examples include self-reported health status (excellent, good, fair, or poor) and body mass index categories (underweight, normal weight, overweight, or obese). For the ordered probit model, the formula for the probabilities of each possible middle outcome ($j \in 2, \dots, J - 1$) is as follows:

$$\Pr(y = j, j \neq 1 \text{ or } J|X, x) = \Phi(\beta_0^j + \beta_x x + X\beta) - \Phi(\beta_0^j - 1 + \beta_x x + X\beta)$$

For the first category, it is

$$\Pr(y = 1|X, x) = \Phi(\beta_0^1 + \beta_x x + X\beta)$$

and for the last (highest, Jth) category, it is

$$\Pr(y = J|X, x) = 1 - \Phi(\beta_0^J - 1 + \beta_x x + X\beta)$$

Again, for the ordered logit model, the cumulative logistic function would replace the cumulative normal function in the above equations.

2.6 Survey commands

When a model is fit with `svy` commands, information to adjust for weights, clustering, or stratification is automatically passed along to our new Stata command and is used in `margins` to compute the ARR and ARD, adjusted for complex survey design. Stata computes linearized standard errors, the default for survey data, which replace the variance–covariance matrix of the estimated coefficients (which is conditional on the covariates) with an estimator that is unconditional on the covariates. Our command designates the variance estimation type as "unconditional" for models with survey data, generating linearized standard errors. For all other models, `margins` will calculate delta-method standard errors using the variance estimation type designated in the previously run model (bootstrap, jackknife, clustered standard errors, etc.).

3 The adjrr command

3.1 Mechanics of the adjrr command

The `adjrr` command uses the `margins` command to calculate ARRs and ARDs after running a logit or probit model with a binary, a multinomial, or an ordered outcome. The `margins` command is versatile and estimates marginal effects for complex, nonlinear models, including those with interactions and survey data.

Within each type of nonlinear model, the `adjrr` command uses the `margins` command with the `at()` option. The `at()` option directs Stata to calculate the two individual predicted probabilities that construct the ARR and ARD at specified values. For multinomial and ordered outcome variables, the particular outcome value is selected, and the code loops over each value the outcome can take. For example, in an ordered model with five possible outcomes, the `adjrr` command computes the ARR and ARD for each of the five outcomes. By using `nlcom` after `margins`, `adjrr` manipulates the predicted probabilities to calculate the ARR and ARD.

Because the `margins` command takes into account the variance structure of the previously run model and estimates delta-method standard errors, the `adjrr` command also incorporates these various variance structures. When the original model is fit controlling for complex survey design, the default is to compute linearized standard errors, again taking into account the complex survey design.

3.2 Syntax

The syntax for calculating the ARR and ARD for a particular covariate after running a specific model is

`adjrr` *varname* [*if*] [, `x0(`*value0*`)` `x1(`*value1*`)` `at(`*atspec*`)`]

where *varname* represents the covariate of interest. The default of this command is to calculate the ARR and ARD of a binary variable, setting the baseline value (`x0()`) equal to 0 and the resulting value (`x1()`) equal to 1. Therefore, when users evaluate a continuous covariate such as age, simply typing

```
. adjrr age
```

will calculate the ARR and ARD, comparing observations at age 1 with observations at age 0, all else equal. In addition, users can specify other values at which to evaluate a particular covariate by inputting specific values for `x0()` and `x1()`. For example, suppose the desired comparison is between observations at age 65 and age 85. The user will then input

```
. adjrr age, x0(65) x1(85)
```

to estimate the ARR and ARD of interest. Further options for this command include designating a particular subsample over which to calculate the ARR and ARD by using an `if` statement. For example, if the user wants to investigate the subsample of women and compare observations at age 20 and age 30, the user will input

```
. adjrr age if female == 1, x0(20) x1(30)
```

The user may also specify values for other covariates in the model by using the `at()` option. If the relevant comparison is between observations at age 20 and age 30, treating all observations in the full sample as if they were women, including those who are actually men, the user will input

```
. adjrr age, x0(20) x1(30) at(female == 1)
```

3.3 Output

When the user runs `adjrr` for a particular covariate, estimates of the ARR and ARD are displayed on separate lines along with their delta-method standard errors and 95% confidence intervals. In models where the outcome is multinomial or ordered, ARRs, ARDs, standard errors, and confidence intervals are estimated for each outcome. `adjrr` also reports the predicted probabilities that compose the elements of the ARR and ARD formulas, their standard errors, and 95% confidence intervals. These elements can be thought of as the baseline risk and the exposed risk. Additionally, two *p*-values are reported. One *p*-value is from a linear test of equivalence between the baseline and exposed risks. The second *p*-value is from a nonlinear test that the natural log of the ARR is equal to 0.

`adjrr` stores results in `r()`. The 95% confidence interval for the ARR is estimated first on the log scale before the endpoints are exponentiated. This transform-the-endpoints method (previously discussed in Cummings [2011]; StataCorp [2011, 1330–1332]) results in an asymmetric confidence interval for the ARR that is asymptotically equivalent to a traditionally constructed confidence interval. This approach to estimating confidence intervals performs better with small sample sizes.

3.4 Alternative approaches to calculating ARRs in Stata

There are alternative ways to calculate ARRs in Stata. For alternatives, see Cummings (2009b, 2011) and Localio, Margolis, and Berlin (2007). However, we feel that our new Stata command, `adjrr`, is the easiest to use and has the most features.

4 Calculating ARRs and ARDs after running nonlinear models

4.1 Medical Expenditure Panel Survey data

We illustrate the application of `adjrr` with data from the 2004 Medical Expenditure Panel Survey (MEPS). The data in these examples were drawn from the Household Component, one of four components. The Household Component contains data on a sample of families and individuals drawn from a nationally representative subsample of households that participated in the 2003 National Health Interview Survey. We used a subset of the MEPS 2004 annual file that included all adults aged 18 and older who had no missing data on the main variables of interest. The resulting dataset has 6 variables and 19,386 observations.

We provide examples for each family of nonlinear models (binary, multinomial, and ordered outcomes) for which our command can calculate ARRs and ARDs. Within each example, the dependent variable of interest is health insurance status. Regression risk analysis will be conducted treating this variable as a binary, a multinomial, and an ordered outcome. In this MEPS dataset, health insurance is divided into three mutually exclusive categories. About 29% are covered by public insurance, 53% by private insurance, and 18% are uninsured.

```
. use meps2004_adjrr.dta
(MEPS04 date with edits)
. summarize
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
      female |     19386    .5495719    .4975494          0          1
         age |     19386    45.36088      17.387         18         85
     race_bl |     19386    .1382441    .3451649          0          1
    race_oth |     19386    .0653564    .2471601          0          1
     insured |     19386    .8223976    .3821875          0          1
-------------+--------------------------------------------------------
   ins_group |     19386    1.646652     .764021          1          3
```

The explanatory variables in each model were age, sex, and race. Race is divided simply into black and other nonwhite, with white as the omitted group.

```
. tab ins_group
  Insurance |
      group |      Freq.     Percent        Cum.
------------+-----------------------------------
  1 Private |     10,293       53.10       53.10
   2 Public |      5,650       29.14       82.24
3 Uninsured |      3,443       17.76      100.00
------------+-----------------------------------
      Total |     19,386      100.00
```

The models used in this article are for illustrative purposes only, and readers should not infer causality or focus on the substantive findings.

4.2 Logit model

Using a binary measure of health insurance, we estimated the probability of having any insurance versus having none based on a few demographics.

```
. logit insured female age race_bl race_oth, nolog
Logistic regression                               Number of obs   =      19386
                                                  LR chi2(4)      =    1132.62
                                                  Prob > chi2     =     0.0000
Log likelihood = -8501.2678                       Pseudo R2       =     0.0625

------------------------------------------------------------------------------
     insured |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |   .3009373   .0387592     7.76   0.000     .2249706    .3769039
         age |   .0391182   .0013055    29.96   0.000     .0365595    .0416769
     race_bl |   .0002987   .0556061     0.01   0.996    -.1086872    .1092846
    race_oth |   .3806466   .0848323     4.49   0.000     .2143784    .5469148
       _cons |  -.2819593   .0583701    -4.83   0.000    -.3963626   -.1675559
------------------------------------------------------------------------------
```

After running a logit model, the `adjrr` command calculates and displays estimates of the ARR and ARD with delta-method standard errors.

```
. adjrr female
R1  = 0.8415 (0.0035)    95% CI  (0.8347, 0.8483)
R0  = 0.7997 (0.0041)    95% CI  (0.7916, 0.8078)
ARR = 1.0522 (0.0070)    95% CI  (1.0387, 1.0660)
ARD = 0.0418 (0.0054)    95% CI  (0.0312, 0.0524)
p-value (R0 = R1):  0.0000
p-value (ln(R1/R0) = 0):  0.0000
. adjrr age, x0(20) x1(30)
R1  = 0.7454 (0.0042)    95% CI  (0.7371, 0.7537)
R0  = 0.6650 (0.0069)    95% CI  (0.6515, 0.6784)
ARR = 1.1210 (0.0061)    95% CI  (1.1090, 1.1330)
ARD = 0.0804 (0.0033)    95% CI  (0.0739, 0.0869)
p-value (R0 = R1):  0.0000
p-value (ln(R1/R0) = 0):  0.0000
```

The ARR estimate on the variable `female` can be interpreted as women being 1.0522 times more likely (5.22% more likely) to have insurance than men, on average, holding all else constant. The ARD represents an absolute risk measure and can be interpreted as women having insurance 4.18 percentage points more often than men, on average.

Because insurance was common (82% in the study sample), the adjusted odds ratio of 1.35 [exp(0.3001)] was much further from 1 than the ARR of 1.05.

The ARR estimate on the continuous variable `age` can be interpreted similarly. On average, 30-year-olds are 12.1% more likely to have insurance than 20-year-olds. The ARD shows that 30-year-olds, on average, have health insurance 8.04 percentage points more often than 20-year-olds, holding all else constant.

4.3 Probit model

A comparable probit model can be fit predicting insurance status as a function of the same demographic variables. Although the probit coefficients are typically smaller, the substantive results are essentially the same as those of the logit model.

```
. probit insured female age race_bl race_oth, nolog
Probit regression                                 Number of obs   =      19386
                                                  LR chi2(4)      =    1161.97
                                                  Prob > chi2     =     0.0000
Log likelihood = -8486.5894                       Pseudo R2       =     0.0641

------------------------------------------------------------------------------
     insured |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |   .1617922   .0218733     7.40   0.000     .1189214    .2046629
         age |   .0223198   .0007202    30.99   0.000     .0209082    .0237313
     race_bl |  -.0052502   .0313762    -0.17   0.867    -.0667466    .0562461
    race_oth |   .2019072   .0462274     4.37   0.000     .1113032    .2925112
       _cons |  -.1157799   .0335045    -3.46   0.001    -.1814475   -.0501123
------------------------------------------------------------------------------
```

Calculating the ARR and ARD separately when sex and age are the variables of interest, the command generates the following results. The estimates are similar to those reported after running the logit model.

```
. adjrr female
R1  = 0.8406 (0.0035)    95% CI  (0.8339, 0.8474)
R0  = 0.8011 (0.0041)    95% CI  (0.7930, 0.8091)
ARR = 1.0494 (0.0069)    95% CI  (1.0360, 1.0631)
ARD = 0.0396 (0.0054)    95% CI  (0.0290, 0.0501)
p-value (R0 = R1):  0.0000
p-value (ln(R1/R0) = 0):  0.0000
. adjrr age, x0(20) x1(30)
R1  = 0.7429 (0.0043)    95% CI  (0.7345, 0.7512)
R0  = 0.6664 (0.0066)    95% CI  (0.6534, 0.6794)
ARR = 1.1147 (0.0055)    95% CI  (1.1041, 1.1255)
ARD = 0.0765 (0.0030)    95% CI  (0.0707, 0.0822)
p-value (R0 = R1):  0.0000
p-value (ln(R1/R0) = 0):  0.0000
```

Alternatively, we can calculate the ARR and ARD for a subgroup. Restricting the sample to individuals who report their race as black, we can recalculate the ARR and ARD for the variable `female` as follows:

```
. adjrr female if race_bl == 1
R1  = 0.8296 (0.0072)    95% CI  (0.8154, 0.8438)
R0  = 0.7882 (0.0086)    95% CI  (0.7714, 0.8051)
ARR = 1.0525 (0.0076)    95% CI  (1.0377, 1.0676)
ARD = 0.0414 (0.0057)    95% CI  (0.0302, 0.0527)
p-value (R0 = R1):  0.0000
p-value (ln(R1/R0) = 0):  0.0000
```

These estimates show that among blacks, on average, women are 1.0525 times more likely (5.25% more likely) to be insured than men. Black women are also 4.14 percentage points more likely to be insured than black men, on average.

4.4 Multinomial models

When users run either a logit or a probit model with a multinomial outcome variable, ARRs and ARDs can be calculated for each outcome by using the `adjrr` command. The multinomial health insurance variable, `ins_group`, captures whether an individual has private (`ins_group` = 1), public (`ins_group` = 2), or no insurance (`ins_group` = 3). The following simple multinomial logistic model is fit.

```
. mlogit ins_group female age race_bl race_oth, nolog
Multinomial logistic regression                 Number of obs   =      19386
                                                LR chi2(8)      =    5015.31
                                                Prob > chi2     =     0.0000
Log likelihood = -16924.793                     Pseudo R2       =     0.1290

------------------------------------------------------------------------------
   ins_group |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
1_Private    |  (base outcome)
-------------+----------------------------------------------------------------
2_Public     |
      female |   .4801498   .0379564    12.65   0.000     .4057567    .5545429
         age |   .0640783   .0012046    53.20   0.000     .0617174    .0664392
     race_bl |   .6329817   .0527671    12.00   0.000     .5295601    .7364032
    race_oth |  -.0808391   .0772467    -1.05   0.295    -.2322397    .0705616
       _cons |  -4.127384   .0707389   -58.35   0.000     -4.26603   -3.988739
-------------+----------------------------------------------------------------
3_Uninsured  |
      female |   -.181544   .0397481    -4.57   0.000    -.2594489   -.1036392
         age |   -.020199   .0014168   -14.26   0.000    -.0229759    -.017422
     race_bl |   .1675143   .0581024     2.88   0.004     .0536356    .2813929
    race_oth |  -.3975927   .0858646    -4.63   0.000    -.5658842   -.2293012
       _cons |  -.2113852   .0622577    -3.40   0.001    -.3334081   -.0893622
------------------------------------------------------------------------------
```

Regardless of the reference outcome chosen for the above regression, the `adjrr` command can estimate the ARR and ARD with standard errors for each outcome category. The syntax of this command is equivalent to the logit case. If we isolate `female` as the variable of interest, the output of the `adjrr` command is as follows:

```
. adjrr female
R1(outcome 1)  = 0.5134 (0.0046)    95% CI  (0.5044, 0.5225)
R0(outcome 1)  = 0.5535 (0.0052)    95% CI  (0.5434, 0.5636)
ARR(outcome 1) = 0.9277 (0.0120)    95% CI  (0.9044, 0.9515)
ARD(outcome 1) = -0.0400 (0.0069)    95% CI  (-0.0536, -0.0264)
p-value (R0 = R1)(outcome 1):  0.0000
p-value (ln(R1/R0) = 0)(outcome 1):  0.0000
R1(outcome 2)  = 0.3277 (0.0040)    95% CI  (0.3199, 0.3355)
R0(outcome 2)  = 0.2464 (0.0041)    95% CI  (0.2383, 0.2544)
ARR(outcome 2) = 1.3301 (0.0274)    95% CI  (1.2774, 1.3850)
ARD(outcome 2) = 0.0813 (0.0057)    95% CI  (0.0701, 0.0925)
p-value (R0 = R1)(outcome 2):  0.0000
p-value (ln(R1/R0) = 0)(outcome 2):  0.0000
R1(outcome 3)  = 0.1588 (0.0035)    95% CI  (0.1520, 0.1657)
R0(outcome 3)  = 0.2001 (0.0041)    95% CI  (0.1920, 0.2083)
ARR(outcome 3) = 0.7936 (0.0239)    95% CI  (0.7481, 0.8419)
ARD(outcome 3) = -0.0413 (0.0054)    95% CI  (-0.0519, -0.0307)
p-value (R0 = R1)(outcome 3):  0.0000
p-value (ln(R1/R0) = 0)(outcome 3):  0.0000
```

In modeling health insurance as a categorical variable, we can create a richer understanding of how different types of health insurance vary between men and women. We find women are more likely to have public insurance than men. In contrast, men are more likely than women to have private insurance or to be uninsured. Interpreting the ARR estimates for outcome 1 (private insurance), we find that women are 1.0723 times less likely (7.23% less likely, where $0.0723 = 1 - 0.9277$) to have private insurance than men, on average. In terms of absolute differences in insurance coverage, women, on average, have private insurance four percentage points less often than men.

4.5 Ordered models

We can extend our simple model to an ordered probit model by considering the health insurance categories as representing different levels of coverage that can be ordered. We use the same outcome variable, `ins_group`, for this model, and the insurance groups are now considered ordered (for the purpose of this illustration only) according to this assumption: that private insurance represents a higher level of coverage in comparison with public insurance. Thus being uninsured represents the lowest level of coverage. After running the simple ordered model, we calculate ARRs and ARDs with standard errors for each insurance category. All ARRs and ARDs will be correctly estimated regardless of the reference category chosen when running the regression.

```
. oprobit ins_group female age race_bl race_oth, nolog
Ordered probit regression                         Number of obs   =      19386
                                                  LR chi2(4)      =     188.76
                                                  Prob > chi2     =     0.0000
Log likelihood = -19338.067                       Pseudo R2       =     0.0049

------------------------------------------------------------------------------
   ins_group |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |   .0147683    .016844     0.88   0.381    -.0182454    .0477819
         age |   .0049324   .0004587    10.75   0.000     .0040335    .0058314
     race_bl |   .1533977   .0239079     6.42   0.000     .1065391    .2002562
    race_oth |  -.1688425   .0350418    -4.82   0.000     -.237523   -.1001619
-------------+----------------------------------------------------------------
       /cut1 |   .3269836   .0253022                      .2773923    .3765749
       /cut2 |   1.175754   .0258825                      1.125026    1.226483
------------------------------------------------------------------------------

. adjrr female
R1(outcome 1)  = 0.5310 (0.0046)    95% CI  (0.5219, 0.5401)
R0(outcome 1)  = 0.5369 (0.0052)    95% CI  (0.5268, 0.5470)
ARR(outcome 1) = 0.9891 (0.0123)    95% CI  (0.9653, 1.0136)
ARD(outcome 1) = -0.0058 (0.0067)    95% CI  (-0.0189, 0.0072)
p-value (R0 = R1)(outcome 1):  0.3806
p-value (ln(R1/R0) = 0)(outcome 1):  0.3803
R1(outcome 2)  = 0.2906 (0.0034)    95% CI  (0.2839, 0.2973)
R0(outcome 2)  = 0.2885 (0.0035)    95% CI  (0.2817, 0.2954)
ARR(outcome 2) = 1.0070 (0.0081)    95% CI  (0.9913, 1.0230)
ARD(outcome 2) = 0.0020 (0.0023)    95% CI  (-0.0025, 0.0066)
p-value (R0 = R1)(outcome 2):  0.3813
p-value (ln(R1/R0) = 0)(outcome 2):  0.3815
R1(outcome 3)  = 0.1784 (0.0033)    95% CI  (0.1719, 0.1850)
R0(outcome 3)  = 0.1746 (0.0036)    95% CI  (0.1675, 0.1817)
ARR(outcome 3) = 1.0218 (0.0251)    95% CI  (0.9737, 1.0722)
ARD(outcome 3) = 0.0038 (0.0043)    95% CI  (-0.0047, 0.0123)
p-value (R0 = R1)(outcome 3):  0.3803
p-value (ln(R1/R0) = 0)(outcome 3):  0.3809
```

Modeling health insurance as an ordered categorical variable reveals slightly different conclusions than when modeling insurance as an unordered categorical variable. Women are now estimated to have a higher likelihood of being uninsured than males. Women remain less likely to have private insurance than men, but the predicted relative difference is smaller. Interpreting the coefficients for outcome 3 (uninsured), women are 1.0218 times more likely (2.18% more likely) to be uninsured than men, on average. Alternatively, women, on average, are uninsured 0.38 percentage points more often than men.

4.6 Interactions

We demonstrate two further extensions to the `adjrr` command, models with interaction terms and survey data, using a dataset from the Stata 12 *Survey Data Reference Manual*. This dataset is a selected sample from the National Health and Nutrition Examination Survey.

When users run a model with interaction terms, the interacted variables and the interaction term must be properly identified for the `margins` command to correctly evaluate the model. This means using `#` to show interactions and using the prefixes `i.` and `c.` to indicate indicator and continuous variables. Once the model is appropriately specified, the `adjrr` command can be run as before to estimate the ARR and ARD for the covariate of interest.

We ran a logit model using `nhanes2.dta`, which estimates the outcome of diabetes as a function of sex, age, and race. Our outcome of interest is a binary variable denoting whether an individual has diabetes. The race variables included in the regression are broken down into black, white, and other race. In the notation below, we indicate `female` as an indicator variable, `age` as a continuous variable, and the interaction between age and sex. We would include such an interaction term in our model if we believe age affects the risk of diabetes differently between men and women.

```
. webuse nhanes2
. logit diabetes i.female c.age i.female#c.age i.black i.orace, nolog
Logistic regression                             Number of obs   =     10349
                                                LR chi2(5)      =    380.57
                                                Prob > chi2     =    0.0000
Log likelihood = -1809.4745                     Pseudo R2       =    0.0952

------------------------------------------------------------------------------
    diabetes |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    1.female |   1.352549   .4851081     2.79   0.005     .4017543    2.303343
         age |    .071462   .0063009    11.34   0.000     .0591124    .0838115
             |
female#c.age |
          1  |  -.0197972   .0078278    -2.53   0.011    -.0351395   -.0044549
             |
     1.black |   .7177366    .127091     5.65   0.000     .4686427    .9668304
     1.orace |   .1989662   .3520485     0.57   0.572    -.4910362    .8889686
       _cons |  -7.142681   .3961563   -18.03   0.000    -7.919133   -6.366229
------------------------------------------------------------------------------
. adjrr female
R1  = 0.0515 (0.0029)    95% CI  (0.0457, 0.0572)
R0  = 0.0447 (0.0029)    95% CI  (0.0390, 0.0503)
ARR = 1.1529 (0.0996)    95% CI  (0.9733, 1.3656)
ARD = 0.0068 (0.0041)    95% CI  (-0.0013, 0.0149)
p-value (R0 = R1):  0.0979
p-value (ln(R1/R0) = 0):  0.0996
```

Once interaction terms are incorporated into a model, running the `adjrr` command and interpreting the results are equivalent to the case without interaction terms. As long as Stata's standard `#` notation is used, the `adjrr` command automatically takes into account the interaction of the variable of interest with other variables. In our example, `adjrr` reveals that women are 1.1529 times more likely (15.29% more likely) to have diabetes than men, on average. In terms of absolute differences, women, on average, have diabetes 0.68 percentage points more often than men.

Given this model specification, we may be interested in calculating ARRs and ARDs for the variable `female` at a particular age. One approach is using the `at()` option to set the sample to a specific age, such as the mean. The syntax for the `at()` specification follows the **`margins`** command. For example,

```
. adjrr female, at((mean) age)
R1  = 0.0382 (0.0030)    95% CI  (0.0323, 0.0441)
R0  = 0.0257 (0.0028)    95% CI  (0.0201, 0.0312)
ARR = 1.4868 (0.2019)    95% CI  (1.1395, 1.9401)
ARD = 0.0125 (0.0041)    95% CI  (0.0044, 0.0206)
p-value (R0 = R1):  0.0025
p-value (ln(R1/R0) = 0):  0.0035
```

When we set all observations to the mean age, `adjrr` estimates that women are 1.4868 times more likely (48.68% more likely) to have diabetes than men, on average. This large relative difference corresponds to the small absolute difference of 1.25 percentage points.

4.7 Survey commands

Extending the estimation of ARRs and ARDs with survey data is simple. After users identify the survey design of the dataset and run the regression model with the survey prefix command, they can run the `adjrr` command as previously described.

Below we run the equivalent logit model as in section 4.6, but we now incorporate the appropriate sampling units, weights, and strata from `nhanes2.dta`. Notice how including the survey design parameters generates different estimates.

```
. svyset psu [pweight=finalwgt], strata(strata)

      pweight: finalwgt
          VCE: linearized
  Single unit: missing
     Strata 1: strata
         SU 1: psu
        FPC 1: <zero>

. svy: logit diabetes i.female c.age i.female#c.age i.black i.orace, nolog
(running logit on estimation sample)

Survey: Logistic regression

Number of strata   =        31                Number of obs      =      10349
Number of PSUs     =        62                Population size    =  117131111
                                              Design df          =         31
                                              F(   5,     27)    =      61.30
                                              Prob > F           =     0.0000

------------------------------------------------------------------------------
             |             Linearized
    diabetes |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    1.female |    1.76606   .5556064     3.18   0.003     .6328936    2.899227
         age |   .0760729    .005669    13.42   0.000      .064511    .0876348
             |
female#c.age |
          1  |    -.02733   .0086152    -3.17   0.003    -.0449008   -.0097592
             |
     1.black |   .7938007    .128747     6.17   0.000     .5312196    1.056382
     1.orace |  -.3278488    .301033    -1.09   0.285    -.9418097     .286112
       _cons |  -7.408693   .3781967   -19.59   0.000    -8.180031   -6.637356
------------------------------------------------------------------------------

. adjrr female

R1  = 0.0382 (0.0026)    95% CI  (0.0330, 0.0433)
R0  = 0.0301 (0.0027)    95% CI  (0.0248, 0.0354)
ARR = 1.2660 (0.1469)    95% CI  (1.0085, 1.5893)
ARD = 0.0080 (0.0039)    95% CI  (0.0004, 0.0156)
p-value (R0 = R1):  0.0470
p-value (ln(R1/R0) = 0):  0.0507
```

Specifying the survey design also changes the ARR and ARD estimates. Women, on average, are calculated as being 1.2660 times more likely (26.6% more likely) to have diabetes than men. Alternatively, women have diabetes 0.8 percentage points more often than men.

We remind users that Stata does not allow the `if` qualifier of the `adjrr` command with complex survey design, because the weights would be wrong for any subset of the data. The `at()` option is still allowed.

5 Conclusion

Our new Stata command, `adjrr`, easily computes ARRs and ARDs by building on the `margins` command. Calculating these estimates and delta-method standard errors is simple and user friendly with the `adjrr` command. We further extend the basic results from Kleinman and Norton (2009) to models where the variable of interest is not dichotomous, to subsets of the data, to complex survey design, and to models with multinomial and ordered outcomes. Our new Stata command allows for all of these extensions.

6 Acknowledgments

The authors gratefully acknowledge funding from the Agency for Healthcare Research and Quality (grant 1R18HS018032) and the National Institutes of Health: National Center for Research Resources (3UL1RR029887-03S1, UL1RR029887). The authors would also like to thank the reviewer, who provided a thorough and thoughtful review.

7 References

Altman, D. G., J. J. Deeks, and D. L. Sackett. 1998. Odds ratios should be avoided when events are common. *British Medical Journal* 317: 1318.

Bier, V. M. 2001. On the state of the art: Risk communication to the public. *Reliability Engineering & System Safety* 71: 139–150.

Cummings, P. 2009a. The relative merits of risk ratios and odds ratios. *Archives of Pediatrics and Adolescent Medicine* 163: 438–445.

———. 2009b. Methods for estimating adjusted risk ratios. *Stata Journal* 9: 175–196.

———. 2011. Estimating adjusted risk ratios for matched and unmatched data: An update. *Stata Journal* 11: 290–298.

Flanders, W. D., and P. H. Rhodes. 1987. Large sample confidence intervals for regression standardized risks, risk ratios, and risk differences. *Journal of Chronic Diseases* 40: 697–704.

Greenland, S. 1987. Interpretation and choice of effect measures in epidemiologic analyses. *American Journal of Epidemiology* 125: 761–768.

———. 2004. Model-based estimation of relative risks and other epidemiologic measures in studies of common outcomes and in case–control studies. *American Journal of Epidemiology* 160: 301–305.

Greenland, S., and P. W. Holland. 1991. Estimating standardized risk differences from odds ratios. *Biometrics* 47: 319–322.

Klaidman, S. 1990. How well the media report health risk. *Daedalus* 119: 119–132.

Kleinman, L. C., and E. C. Norton. 2009. What's the risk? A simple approach for estimating adjusted risk measures from nonlinear models including logistic regression. *Health Services Research* 44: 288–302.

Localio, A. R., D. J. Margolis, and J. A. Berlin. 2007. Relative risks and confidence intervals were easily computed indirectly from multivariable logistic regression. *Journal of Clinical Epidemiology* 60: 874–882.

Spiegelman, D., and E. Hertzmark. 2005. Easy SAS calculations for risk or prevalence ratios and differences. *American Journal of Epidemiology* 162: 199–200.

StataCorp. 2011. *Stata 12 Base Reference Manual*. College Station, TX: Stata Press.

Teuber, A. 1990. Justifying risk. *Daedalus* 119: 235–254.

About the authors

Edward C. Norton is a professor of health management and policy and a professor of economics at the University of Michigan and the National Bureau of Economic Research.

Morgen M. Miller is a doctoral candidate in health management and policy and in economics at the University of Michigan.

Lawrence C. Kleinman is vice chair and associate professor of health evidence and policy and is associate professor of pediatrics at Icahn School of Medicine at Mount Sinai.