Computing adjusted risk ratios and risk differences in Stata

Edward C. Norton  
University of Michigan  
Ann Arbor, Michigan  
and NBER  
ecnorton@umich.edu

Morgen M. Miller  
University of Michigan  
Ann Arbor, Michigan  
mmmill@umich.edu

Lawrence C. Kleinman  
Mount Sinai School of Medicine  
New York, New York  
lawrence.kleinman@mssn.edu

Abstract.
This paper explains 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 post-estimation command adjrr that calculates adjusted risk ratios (ARR) and adjusted risk differences (ARD) after running logit or probit models with either binary, multinomial, or ordered outcomes. It 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 estimated model. Data from the Medical Expenditure Panel Survey (MEPS) and the National Health and Nutrition Examination Survey (NHANES) are used to illustrate multiple applications of the command.

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

1 Introduction

Researchers often estimate 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 reasons and other reasons, many people have called for researchers to report risk ratios instead of odds ratios (e.g., Greenland 1987; Spiegelman and Hertzmark 2005; Cummings 2009).

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 (e.g., 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 reproduces Mantel-Haenszel results; Kleinman and Norton (2009) also demonstrate that their method is correct given the distribution of covariates, including complex specifications with con-
Adjusted risk ratios
tinous variables, and is robust in many cases.

This paper makes several contributions. First, we show how to compute adjusted risk ratios and adjusted risk differences 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 limited dependent variables. In addition, it is easy and fast to calculate these statistics in Stata because our command builds on the \texttt{margins} command. However, our command makes it much easier than using \texttt{margins}, especially for multinomial and ordered models. Second, we also compute the adjusted risk difference, in addition to the adjusted risk ratio. 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 \texttt{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 zero and one (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 data sets often have sampling weights, clustered observations, and stratification. These can be taken into account when computing adjusted risk ratios (ARR) and adjusted risk differences (ARD).

2 Methods

2.1 Estimation of adjusted risk ratios and adjusted risk differences

The ARR and ARD are two ways to express the relationship between two predicted probabilities based on the estimated model and a set of observations. One is the predicted probability when the variable of interest equals one; the other is the predicted probability when the variable of interest equals zero (more generally, pick any two values of the variable). These predicted probabilities are then averaged over the entire data set (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 \textit{average treatment effect} because it compares the effect of a change in the variable of interest (the treatment) for all observations. All 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 estimating the model, compute two predicted probabilities of mortality for each observation, one assuming the patient did get the drug and the other assuming she did not. The key thing 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 one, computed over the whole sample, with the variable of interest $x$ set equal to one 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 zero.

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

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

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 zero and one. For a continuous variable, zero and one may not be appropriate comparison values. It might be of policy interest, for example, to compare predicted outcomes for persons age 85 compared to persons age 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, equations (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 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 equations (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 equations (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 \( \omega \) 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 one (as opposed to zero). 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 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 equations (1) and (2) depend 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 one is the logistic cumulative distribution function:

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

(5)

### 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 one is the normal cumulative distribution function:

\[
Pr(y = 1|X, x) = \Phi(\beta x + X \beta)
\]  

(6)

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 \( J \)th category has its coefficient normalized to zero) is as follows:

\[
Pr(y = j, j \neq J|X, x) = \frac{e^{(\beta_j x + X \beta_j)}}{\sum e^{(\beta_j x + X \beta_j)} + 1}
\]

(7)
and for the $J$th category the predicted probability is

$$\Pr(y = J|X, x) = \frac{1}{\sum e^{(\beta J_x x + X \beta J)}} + 1$$  \hspace{1cm} (8)

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. For example, self-reported health status (excellent, good, fair, or poor), or 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 ($j \in 2, \ldots, J - 1$) is as follows:

$$\Pr(y = j, j \neq 1 \text{ or } J|X, x) = \Phi(\beta_j^0 + \beta x x + X \beta) - \Phi(\beta_j^0 - 1 + \beta x x + X \beta)$$  \hspace{1cm} (9)

For the first category, it is

$$\Pr(y = 1|X, x) = \Phi(\beta_1^0 + \beta x x + X \beta)$$  \hspace{1cm} (10)

and for the last (highest, $J$th) category it is

$$\Pr(y = J|X, x) = 1 - \Phi(\beta_J^0 - 1 + \beta x x + X \beta)$$  \hspace{1cm} (11)

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

### 2.6 Survey commands

When a model is estimated with `svy` commands, to adjust for weights, clustering, or stratification, this information is automatically passed along to our new Stata command and is used in `margins` to compute the adjusted risk ratio and risk difference, adjusted for complex survey design. Stata computes linearized standard errors, the default for survey data, which replaces 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 adjrr command

3.1 Mechanics of the adjrr command

The adjrr command uses the margins command to calculate ARRs and ARDs after running logit or probit models with binary, multinomial, or ordered outcomes. 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 estimated 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 zero and the resulting value (x1) equal to one. Users can specify other values at which to evaluate a particular covariate by inputting specific values for x0 and x1. Therefore, when evaluating a continuous covariate such as age, simply typing

```
. adjrr age
```

will calculate the ARR and ARD comparing observations at age 1 to observations at age 0, all else equal. Instead, adjrr can calculate the ARR and ARD for any two ages. 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 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
```
The user may also specify values for other covariates in the model using the at option. If the relevant comparison is between observations at age 20 and age 30, treating all observations as women, the user will input

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

### 3.3 Output

Upon running the command 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. The command 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 adjusted risk ratios in Stata. For alternatives, see Cummings (2009, 2011) and Localio and colleagues (2007). However, we feel that our new Stata command adjrr is both easiest to use and has more features.

### 4 Calculating ARRs and ARDs after running nonlinear models

#### 4.1 Medical Expenditure Panel Survey (MEPS) data

We illustrate the application of the adjrr command using data from the 2004 Medical Expenditure Panel Survey (MEPS). The data used 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 prior year’s National Health Interview Survey. We used a subset of the MEPS 2004 annual file that included all adults age 18 and older who had no missing data on the main variables of interest.
The resulting data set has six 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, multinomial, and ordered outcome. In this MEPS data set, health insurance is divided into three mutually-exclusive categories. About 29 percent are covered by public insurance, 53 by private insurance, and 18 percent are uninsured.

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

The models used in this paper 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 none based on a few demographics.
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 are 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 than the adjusted risk ratio 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 estimated 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 from the logit model.

```
. probit insured female age race_bl race_oth, nolog
Probit regression Number of obs = 19386
```
Adjusted risk ratios

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 | 0.1617922 0.0218733 7.40 0.000 0.1189214 0.2046629
age | 0.0223197 0.0007202 30.99 0.000 0.0209082 0.0237313
race_bl | -0.0052502 0.0313762 -0.17 0.867 -0.0667466 0.0562461
race_oth | 0.2019072 0.0462274 4.37 0.000 0.1113032 0.2925112
_cons | -0.1157799 0.0335045 -3.46 0.001 -0.1814475 -0.0501123
---------------------------------------------

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.

.jr 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

.jr 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 (0.0069) 95% CI (1.0360, 1.1951)
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.

.jr 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 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 running either a logit or probit model with a multinomial outcome variable, ARRs and ARDs can be calculated for each outcome using the `adjrr` command. The multinomial health insurance variable, `ins_group`, captures whether or not an individual has private (`ins_group = 1`), public (`ins_group = 2`), or no insurance (`ins_group = 3`). The following simple multinomial logistic model is estimated.

```
. 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. Isolating 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)
```
Adjusted risk ratios

<table>
<thead>
<tr>
<th>Outcome 3</th>
<th>R0</th>
<th>R1</th>
<th>ARR</th>
<th>ARD</th>
<th>p-value (R0 = R1)</th>
<th>p-value (ln(R1/R0) = 0)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.2001 (0.0041)</td>
<td>0.7936 (0.0239)</td>
<td>-0.0413 (0.0054)</td>
<td>-0.0413 (0.0054)</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

In modeling health insurance as a categorical variable, we are able to 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. Alternatively, men are more likely than women to have private insurance or be uninsured. Interpreting the ARR estimates for outcome 1 (private insurance), we find that women are 7.23% less likely to have private insurance than men, on average. In terms of absolute differences in insurance coverage, women, on average, have private insurance 4 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 assuming for this illustration only that private insurance represents a higher level of coverage in comparison to public insurance. 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

| 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 |

```

```
. 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) = 0.0058 (0.0067) 95% CI (.0068, 0.0158)
ARD(outcome 1) = -0.0058 (0.0067) 95% CI (-.0189, .0072)
```

```
```
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 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 data set from the Stata 11 Survey Data Reference Manual. This data set is a selected sample from the National Health and Nutrition Examination Survey (NHANES).

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

We ran a logit model using the NHANES data that estimates the outcome of diabetes as a function of sex, age, and race. Our outcome of interest is a binary variable denoting whether or not 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 is a factor variable, age is 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 differentially between men and women.

```
. webuse nhanes2, clear
```
Adjusted risk ratios

.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]
-------------+----------------------------------------------------------------
diabetes\n1.female | 1.352549 .4851081 2.79 0.005 .4017545 2.303343
1.age | .071462 .0063009 11.34 0.000 .0591124 .0838115
1.female#c.age | -1.352549 .4851081 2.79 0.005 .4017545 2.303343
1.age | .071462 .0063009 11.34 0.000 .0591124 .0838115
1.female#c.age | -2.0197972 .0078278 -2.53 0.011 -.0351395 -.0044549
1.black | .7177365 .127091 5.65 0.000 .4686427 .9668304
1.orace | .1989662 .3520485 0.57 0.572 -.4910362 .8889687
_cons | -7.142681 .3961564 -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. The adjrr command automatically takes into account the interaction of the variable of interest with other variables (as long as Stata’s standard # notation is used). In our example, the adjrr command reveals that women are 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 specifications follows the margins command. For example,

adjrr female, at((mean) age)

R1 = 0.0382 (0.0029) 95% CI (0.0323, 0.0441)
R0 = 0.0267 (0.0028) 95% CI (0.0201, 0.0332)
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

Setting all observations to the mean age, the adjrr command estimates that women
are 48.68% more likely to have diabetes than men, on average.

### 4.7 Survey commands

Extending the estimation of ARRs and ARDs using survey data is simple. After identifying the survey design of the data set and running the regression model with the survey prefix command, the `adjrr` command can be run 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 the NHANES data set. Notice how including the survey design parameters generates different estimates.

```
.svysset 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  

| Coef. Std. Err. t P>|t| [95% Conf. Interval] |
|-------------+--------------------------------------------------|
| diabetes | | | | | |
| 1.female | 1.76606 .5556063 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
```
Adjusted risk ratios

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

5 Conclusion
Our new Stata command adjrr easily computes adjusted risk ratios and adjusted risk differences by building on the margins command. Calculating these estimates and delta-method standard errors is simple and user-friendly using our new adjrr command. We further extend the basic results from Kleinman and Norton 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 these extensions.

6 Acknowledgements
The authors gratefully acknowledge funding from AHRQ (Grant 1R18HS018032), NIH/NCRR (3UL1RR029887-03S1), and from NIH/NCRR (UL1RR029887). The authors would also like to thank the reviewer who provided a thorough and thoughtful review.

7 References


About the authors

Edward C. Norton is a professor of health management and policy and professor of economics at the University of Michigan, and NBER.

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 & Policy and Associate Professor of Pediatrics at Mount Sinai School of Medicine.