Butting Heads on Matched Cohort Analysis Using SAS® Software

Melvin T. Alexander, Joseph A. Kufera
National Study Center for Trauma and Emergency Medical Systems, University of Maryland, Baltimore, MD

ABSTRACT

Matched-pair cohort studies involve pairs of cases where outcomes of interest (e.g., mortality, morbidity, injury, etc.) among persons exposed to risk factors are compared with persons not exposed. Matched cohort studies have been used in traffic-crash research, for example, to compare the injury risk of drivers and passengers from the same crash or vehicle.

Matching in cohort studies helps reduce bias and confounding effects that may exist in estimating risk or hazard ratios. Only the data where one or both of the paired cases with the outcome of interest are used, which is useful when little or no information is available regarding pairs in which cases did not have the outcome. For example, this technique could be used to analyze a registry of disease or trauma cases, since no information would be available on the number of cases that are disease or trauma free.

Mantel-Haenszel stratification, logistic and conditional Poisson regression, and Cox proportional hazard models are common methods used to analyze data from cohort studies. This paper will show how SAS procedures GENMOD (for conditional Poisson regression) and PHREG (for Cox proportional hazards regression) can be used to analyze matched-pair cohort data in traffic crash research.

INTRODUCTION

Matching has been used in epidemiological studies as a means of reducing confounding and increasing efficiency in cohort studies. In traffic crash research, matching is used to estimate the association between an exposure, such as seat belt use, and an outcome, such as death.

Matched-pair cohort studies differ from matched-pair case-control studies involved in fatal crashes in that the data come from pairs experiencing the same outcome of interest (e.g., death). Case-control studies tend to overstate associations for the same outcome when the outcome is common.

Matching persons in the same vehicle (or vehicles involved in the same crash) gives research investigators the ability to control for variables that may be too costly or impossible to measure. In addition, matched-pair cohort studies generate valid estimates for risk of dying in crashes in the absence of information about crashes in which no deaths occurred (Cummings et al., 2000; Cummings et al., 2003; Greenland, 1998; Olson et al., 2006).

Matching on persons in the same vehicle could also control potential confounding effects of other vehicle-related factors in a crash, such as speed, vehicle make, whether the vehicle rolled over, and so forth. Matching would even control for variables that may have been overlooked, but may be specific to a vehicle or crash and common to individuals in the same vehicle.

We intend to describe how SAS procedures GENMOD and PHREG can be applied to assess the crude risk of death occurring in a motor vehicle crash for drivers and passengers involved in the same crash. Specific options within each procedure will also be used to construct confidence intervals for the risk ratio. Following calculation of the crude risk ratio, adjustment will be made for several occupant and crash characteristics.

METHODS

A simulated matched-pair cohort study was conducted to compare the risk of death for two front-seat occupants of a motor vehicle involved in a crash. Vehicles were included in the study if they were passenger cars, pickup trucks, sport utility vehicles (SUVs), or minivans. Driver age was a dichotomous variable (ages 60 or older versus ages 15-59) to assess the risk of death between younger and older occupants. Use of a seat belt was also included. SAS 9.1 was used for all data analyses (SAS Institute, Inc., 2003).

The methods typically used to analyze matched-paired cohorts in crash research are: the Mantel-Haenszel method; the double-pair method; Cox proportional hazards regression; and conditional Poisson regression. The Mantel–Haenszel method estimates relative risk in data that are stratified on matched-pairs (Mantel-Haenszel, 1959 and Rothman, 1986). The log risk ratio is simple to calculate but has been limited in its ability to control for potential confounders or examine risk-ratio variation.

Evans (1986) introduced the double-pair method as an adaptation of Mantel-Haenszel method to estimate seat belt effects (belted vs. unbelted) when comparing the fatality risks among front-seat drivers and passengers. Like the
Mantel-Haenszel method, the risk ratios and variance estimators are simple to calculate for separate driver and passenger seating positions. Although able to control for potential confounders such as age or sex, the double-pair method cannot produce estimates for discordant pairs on exposure.

The matched cohorts in our simulation is representative of actual fatality counts of two individuals involved in the same vehicle and crash (e.g., driver and right-front passenger). Allison (2006) described how this approach worked. Let \( y_d \) and \( y_p \) denote the count of driver and passenger deaths involved in the same \( i \)-th vehicle (or crash), respectively. Each variable \( y_i \) is assumed to follow a Poisson distribution with a Poisson parameter \( \mu_i \) that has a probability represented by:

\[
\Pr(y_i = k) = \left( \frac{\mu_i^k e^{-\mu_i}}{k!} \right) \quad \text{where } r=0, 1, 2 \text{ (count of deaths); } j=1(\text{drivers}), 2(\text{passengers}).
\]

The distribution of total deaths may be expressed as: \( n_i = y_d + y_p \). The driver fatality count given the total deaths for drivers and passengers can be approximated using a binomial distribution, i.e.,

\[
y_d | n_i \sim \text{B}(n_i, p_i) \quad \text{with probability parameter} \quad p_i = \frac{\mu_{di}}{\mu_{di} + \mu_{pi}} \quad \text{and } \text{index } n_i.
\]

Poission distribution are the same, i.e., \( \text{E}(y_j) = \text{var}(y_j) = \mu_j \), the loglinear transformation of \( \mu_{di} \) can be written as a function of predictor variables or covariate terms\( (x_i) \), and unobserved fixed effects (\( \epsilon \)). In other words,

\[
\log \mu_{di} = \beta x_i + \epsilon.
\]

A fixed-effects conditional Poisson model [Hardin and Hilbe (2001)] may be estimated with logistic regression programs like SAS LOGISTIC or GENMOD procedure. The GENMOD procedure was chosen because it provided greater flexibility in converting two Poisson regression parameters (i.e., the total count of driver and/or passenger deaths involved in the same crash) into conditional maximum likelihood estimates of binomial parameters expressed in a logistic regression model form. Also, PROC GENMOD provided more generality in estimating standard errors and adjusting for overdispersion than the LOGISTIC procedure.

PROC GENMOD is appropriate when the data satisfy the following conditions:

- Data contain time-dependent covariates;
- Data have missing or incomplete information about crashes in which drivers and passengers survived;
- Data are not-normally distributed but have Poisson distributed, negative-binomial, or binary-binomial outcomes. The negative-binomial would apply if more than two total deaths occurred per vehicle (i.e., back-seat passengers).

Alternatively, the Cox proportional hazards regression stratified on pairs [Hosmer and Lemeshow, (1999)] is also an appropriate tool for analyzing matched-pair cohorts. Cox proportional hazards models are appropriate for analyzing the event history of death occurrence in traffic injuries and have the ability to control for potential confounders and examine risk-ratio variation. The requirements for Cox proportional hazards model assume that:

(i) the follow-up time is the same for all subjects; (ii) stratification is done on pairs; (iii) the Breslow method is used to account for tied times to event (death); and (iv) the likelihood for injury occurrence among all cases is the same. The PHREG procedure in SAS was used to apply the Cox proportional hazards regression. PROC PHREG was chosen because the output from the procedure matches the analytical results pioneered by leading researchers in the traffic-injury research field (Greenland, 1998, Cummings et al., 2004a, 2004b, 2004c, 2004d). One important caveat is that PROC PHREG requires the use of a time variable that has a value of 2 if the passenger died before the driver and 1 if the driver died before the passenger from a 1:1 matched design. This requirement was necessary in order to ensure that the probability of being a driver was properly modeled. Otherwise, the signs of estimated parameter coefficients would be reversed. See Allison (1999) and Stokes et al. (2000) for more information.

To illustrate how matched-pair cohort rates are computed, Table 1 lists hypothetical counts of drivers and passengers from the same vehicle involved in a frontal collision. From the data, the proportion of drivers deaths and the proportion of passenger deaths, along with the crude risk ratio, can be calculated by arranging the data in a 2x2 contingency table of counts for pairs in cells a-d. Using cell-label notation, Greenwood (1998) and Cummings et al. (2003) showed that the crude risk ratio formula comparing driver fatalities to passenger fatalities was:

\[
\frac{(a+b)(a+b+c+d)}{(a+c)(a+b+c+d)}.
\]

The total number of pairs, \( a+b+c+d \), appears in both numerator and denominator. Therefore, the risk ratio simplifies to \( \frac{a+b}{a+c} \). Note that the correct crude risk ratio may be calculated without having any information about the crashes in which both the driver and passenger survived (cell d). Use of this risk ratio would prove helpful if the database used for analysis only includes deaths and therefore
lacks information about crashes in which driver and passenger survived. Formulas for the computing standard errors and the 95% confidence interval are also provided in Table 1.

Table 1: Illustration of Matched-Pair Cohort Analyses of Drivers and Passengers in Fatal Two-Vehicle Head-On Collisions Driver.

<table>
<thead>
<tr>
<th>Died</th>
<th>Passenger Died</th>
</tr>
</thead>
<tbody>
<tr>
<td>a (both driver and passenger died)</td>
<td>Yes</td>
</tr>
<tr>
<td>b (only driver died)</td>
<td>No (Survived)</td>
</tr>
<tr>
<td>Survived</td>
<td>c (only passenger died)</td>
</tr>
<tr>
<td>d (neither died)</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Driver Died</th>
<th>Yes</th>
<th>No (Survived)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>a</td>
<td>b</td>
</tr>
<tr>
<td>No (Survived)</td>
<td>c</td>
<td>d</td>
</tr>
</tbody>
</table>

Hazard or Risk Ratio (RR) = \( \frac{(a + b)(a + b + c + d)}{(a + c)(a + b + c + d)} \)
(simplifies to) = \( \frac{(a + b)}{(a + c)} \), so we do not need to know cell d.

The formula for computing the Mantel-Haenszel Standard Error of Risk Ratio (MHStdErrRR) is:

\[
MHStdErrRR = \sqrt{\frac{(b + c)}{(a + b) \times (a + c)}}
\]

derived from the

Mantel-Haenszel Variance for the Risk Ratio = \( (b + c) ÷ [(a + b) \times (a + c)] \). The 95% Confidence Interval formulas for the Risk Ratios are as follows:

Lower 95% confidence limit (CL) of ln RR = ln RR – 1.96 \times MHStdErrRR
Upper 95% CL of ln RR = ln RR + 1.96 \times MHStdErrRR

Lower 95% CL of RR = exp(lower 95% CL of ln RR)
Upper 95% CL of RR = exp(upper 95% CL of ln RR).

A list of selected records from the simulated database is displayed below. Included in the file are the crash identifier (crashid), which served as the stratification variable to create a separate stratum for each driver-passenger case; indicator variables that distinguished drivers (driver=1) and passengers (driver=0) who died (driverdead=1, passngrdead=1); total count in which at least one of the occupants died (total equals the sum of driverdead and passngrdead), occupants under the age of 60 (age60=0) or over the age of 60 (age60=1); time; gender (sex), vehicle type(car, pickup, utility, minivan); and the matched-cohort cell label.

<table>
<thead>
<tr>
<th>Crashid</th>
<th>driverdead</th>
<th>passngrdead</th>
<th>totaldead</th>
<th>time</th>
<th>driver</th>
<th>age</th>
<th>age60</th>
<th>sex</th>
<th>Cell_label</th>
</tr>
</thead>
<tbody>
<tr>
<td>01</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>40</td>
<td>0</td>
<td>Female</td>
<td>b</td>
</tr>
<tr>
<td>01</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>42</td>
<td>0</td>
<td>Male</td>
<td>b</td>
</tr>
<tr>
<td>02</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>24</td>
<td>0</td>
<td>Male</td>
<td>c</td>
</tr>
<tr>
<td>03</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>51</td>
<td>0</td>
<td>Male</td>
<td>c</td>
</tr>
<tr>
<td>03</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>64</td>
<td>1</td>
<td>Male</td>
<td>c</td>
</tr>
<tr>
<td>04</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>49</td>
<td>0</td>
<td>Female</td>
<td>c</td>
</tr>
<tr>
<td>04</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>43</td>
<td>0</td>
<td>Male</td>
<td>c</td>
</tr>
<tr>
<td>05</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>19</td>
<td>0</td>
<td>Male</td>
<td>c</td>
</tr>
<tr>
<td>05</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>44</td>
<td>0</td>
<td>Female</td>
<td>c</td>
</tr>
<tr>
<td>06</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>72</td>
<td>1</td>
<td>Male</td>
<td>c</td>
</tr>
<tr>
<td>06</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>18</td>
<td>0</td>
<td>Male</td>
<td>c</td>
</tr>
<tr>
<td>07</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>75</td>
<td>1</td>
<td>Male</td>
<td>c</td>
</tr>
<tr>
<td>07</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>32</td>
<td>0</td>
<td>Male</td>
<td>c</td>
</tr>
<tr>
<td>08</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>99</td>
<td>1</td>
<td>Male</td>
<td>b</td>
</tr>
<tr>
<td>08</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>86</td>
<td>1</td>
<td>Male</td>
<td>b</td>
</tr>
<tr>
<td>09</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>25</td>
<td>0</td>
<td>Female</td>
<td>b</td>
</tr>
<tr>
<td>09</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>52</td>
<td>0</td>
<td>Male</td>
<td>b</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>18</td>
<td>0</td>
<td>Female</td>
<td>b</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>70</td>
<td>1</td>
<td>Male</td>
<td>b</td>
</tr>
</tbody>
</table>
RESULTS
A total of 1000 simulated driver-passenger pairs in vehicles involved in a frontal collision were identified as eligible for matched-pair cohort analyses in which one or both of the front-seat occupants of the crashed vehicles died. Table 2 displays the contingency table of the matched cohort data.

<table>
<thead>
<tr>
<th>Passenger Died</th>
<th>Driver Died</th>
<th>Yes</th>
<th>No (Survived)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>108</td>
<td></td>
<td>356</td>
</tr>
<tr>
<td>No (Survived)</td>
<td>536</td>
<td></td>
<td>(?)</td>
</tr>
</tbody>
</table>

Risk ratio = \( \frac{108 + 356}{108 + 536} = 0.721 \)

MH Standard Error = \( \sqrt{\left(\frac{356+536}{(108+356) \times (108+536)}\right)^{1/2}} = 0.058 \) (95% confidence interval 0.647-0.802).

Thus, for all passenger vehicle pairs combined, the crude risk ratio for driver deaths relative to front-seat passenger deaths in vehicles in vehicles involved in a frontal collision was 0.72 (95 percent confidence interval (95% CI) = 0.65-0.80). The risk ratio of 0.72 means that drivers had about a 28% lower risk of dying in a head-on collision than passengers.

The PROC GENMOD statements below can be used to analyze the simulated matched pair cohort data. The dependent variable (driverdead/totaldead) in the MODEL statement is expressed with the `events/trials` syntax that tells SAS that the driver deaths occurred out of the total deaths from the crash. By including the binomial and logit options in the MODEL statement, the conditional Poisson distribution of driver deaths is specified. In addition, the "waldci" and 'lrci" options produce Wald and likelihood ratio estimates, respectively, of the 95% confidence interval of the parameter estimates. Subject = crashid in the REPEATED statement distinguishes individual drivers and passengers from the same crash. Crashid must also be identified as a categorical variable in the CLASS statement. The keywords "type = exch modelse" specifies for SAS to use the exchangeable log odds ratio regression model for all matched-paired cohorts that included model-based as well as empirical-based parameter estimates and standard errors. This will allow for the calculation of confidence intervals by different methods.

```
proc genmod data = simfile;
class crashid;
model driverdead/totaldead = / dist=binomial link=logit waldci lrci type3;
repeated subject = crashid / type = exch modelse;
run;
```

Table 3 provides the SAS output. Note that the output displays a parameter estimate of the risk ratio but not the actual risk ratio. The actual risk ratio must be calculated by the user by exponentiating the parameter estimate, i.e., \( \exp(-0.3278) = 0.721 \). This is the same crude risk ratio as that computed in Table 2 above.

```
Table 3: SAS Output from PROC GENMOD
```

<table>
<thead>
<tr>
<th>Parameter</th>
<th>DF</th>
<th>Estimate</th>
<th>Standard Error</th>
<th>Wald 95% Confidence Limits</th>
<th>Likelihood Ratio 95% Confidence Limits</th>
<th>Chi-Square</th>
<th>Pr &gt; ChiSq</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1</td>
<td>-0.3278</td>
<td>0.0609</td>
<td>-0.4472</td>
<td>-0.2085</td>
<td>28.98</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Scale</td>
<td>0</td>
<td>1.0000</td>
<td>0.0000</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
</tbody>
</table>

To produce the actual risk ratio and its 95% confidence interval, create an Output Delivery System (ODS) output data set of the parameter estimates and a separate SAS dataset (RR1). Use the PRINT procedure to display the risk ratio and confidence limits, as shown in Table 4. (To conserve space, the SAS output from PROC GENMOD will not be displayed).

```
ods output ParameterEstimates = genout1;
proc genmod data = simfile;
class crashid;
```
model driverdead/totaldead = / dist=binomial link=logit waldci lrci type3; 
repeated subject = crashid / type = exch modelse ;
run;

data RR1 ;
set genout1 (obs=1); /* the first record has hazard ratio information */
Riskratio = exp(estimate);
RRLowerCL = exp(LowerWaldCL);
RRUpperCL = exp(UpperWaldCL);
run;

Title 'GENMOD Risk Ratios Using the Default Wald Standard Error Estimates';

proc print data = RR1;
  var Riskratio RRLowerCL RRUpperCL ;
  format Riskratio RRLowerCL RRUpperCL 12.5;
run;

ods output close;

Table 4: GENMOD Risk Ratios Using the Default Wald Standard Error Estimates

<table>
<thead>
<tr>
<th>Obs</th>
<th>Riskratio</th>
<th>RRLowerCL</th>
<th>RRUpperCL</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.72050</td>
<td>0.63944</td>
<td>0.81183</td>
</tr>
</tbody>
</table>

Use a different ODS statement to reproduce the same risk ratio but a different 95% confidence interval that is based on the standard errors from the model-based estimation method after using the exponential transformation (Table 5).

ods output GEEModPESt = genout2;

proc genmod data = simfile;
class crashid;
model driverdead/totaldead = / dist=binomial link=logit waldci lrci type3;
repeated subject = crashid / type = exch modelse ;
run;

data RR2 ;
set genout2 (obs=1); /* first record from GEEModPESt ODS table */
Riskratio = exp(estimate);
RRLowerCL = exp(LowerCL);
RRUpperCL = exp(UpperCL);
run;

Title 'GENMOD Risk Ratios Using Model Standard Error Estimates';
proc print data = RR2;
  var Riskratio RRLowerCL RRUpperCL ;
  format Riskratio RRLowerCL RRUpperCL 12.5;
run;

ods output close;

5
Table 5: SAS Output and GENMOD Risk Ratios Using Model Standard Error Estimates

| Parameter   | Estimate | Standard Error | 95% Confidence Limits | Z      | Pr > |Z|               |
|-------------|----------|----------------|-----------------------|--------|------|---------|
| Intercept   | -0.3278  | 0.0811         | -0.4868               | -0.1689 | -4.04| <.0001  |
| Scale       | 0.9418   | .              | .                     | .      | .    | .       |
| Obs 1       | 0.72050  | 0.61924        | 0.83831               | .      | .    | .       |

Finally, PROC GENMOD will yield the results of the empirical-based method for calculating standard errors by use of a third ODS statement (Table 6). The result is yet a third different 95% confidence interval for the risk ratio.

```sas
ods output GEEEmpPEst = geestd1 ;

proc genmod data = simfile;
   class crashid;
   model driverdead/totaldead = / dist=binomial link=logit waldci lrci type3;
   repeated subject = crashid / type = exch modelse ;
run ;

data RR3 ;
   set geestd1 (obs=1); /* first record from GEEEmpPEst ODS table */
   Riskratio = exp(estimate);
   RRLowerCL = exp(LowerCL);
   RRUpperCL = exp(UpperCL);
run ;

Title 'GENMOD Risk Ratios Using Empirical-Based Standard Error Estimates';

proc print data = RR3;
   var Riskratio RRLowerCL RRUpperCL ;
   format Riskratio RRLowerCL RRUpperCL 12.5 ;
run ;

ods output close ;
```

Table 6: SAS Output and GENMOD Risk Ratios Using Empirical-Based Standard Error Estimates

| Parameter   | Estimate | Standard Error | 95% Confidence Limits | Z      | Pr > |Z|               |
|-------------|----------|----------------|-----------------------|--------|------|---------|
| Intercept   | -0.3278  | 0.0773         | -0.4793               | -0.1764 | -4.24| <.0001  |
| Obs 1       | 0.72050  | 0.61461        | 0.84463               | .      | .    | .       |
To reproduce the analysis with PROC PHREG, we used the SAS code below to produce Table 7. The expression `time*dead(0)` in the MODEL statement of the PHREG procedure indicates a censored case (i.e., whether the driver or passenger survived, coded with values of 1, because fatalities were the characteristics of interest). The variable time is referred to as the censored-event history variable where driver deaths were followed by passenger deaths as a way of modeling event occurrence according to proportional hazard theory and the partial likelihood method. The Breslow method (ties = breslow rl) was used to account for breaking tied times for matched pairs that had identical conditional poisson likelihoods stratified by crashid. Using other methods to break the ties will not produce the correct risk ratio, according to Cummings et al. (2003, 2004a-2004d).

```sas
ods output ParameterEstimates=phreg1;
proc phreg data = simfile nosummary ;
model time*dead(0) = driver / ties = breslow rl ;
strata crashid;
run;
proc print data=phreg1 noobs;
format Hazardratio HRLowerCL HRUpperCL 12.5;
run;
ods output close ;
```

Table 7: Selected Output from PROC PHREG for the crude hazard ratio and 95% Confidence Interval

<table>
<thead>
<tr>
<th>Variable</th>
<th>DF</th>
<th>Parameter Estimate</th>
<th>Standard Error</th>
<th>Chi-Square</th>
<th>Pr &gt; ChiSq</th>
<th>Hazard Ratio</th>
<th>95% Hazard Ratio Confidence Limits</th>
</tr>
</thead>
<tbody>
<tr>
<td>driver</td>
<td>1</td>
<td>-0.32781</td>
<td>0.08612</td>
<td>14.4906</td>
<td>0.0001</td>
<td><strong>0.72050</strong></td>
<td>0.60860 0.85297</td>
</tr>
</tbody>
</table>

Observe that the hazard ratio computed from PROC PHREG resembles the same crude risk ratios computed manually from the contingency table and the GENMOD procedure.

The 95% Confidence Intervals computed from the GENMOD and PHREG procedures vary because different formulae were used to compute the standard errors. PROC GENMOD provided three standard error computing methods (Wald/likelihood ratio, model-based, and empirical-based), whereas PROC PHREG provided one method of computing the standard error (partial likelihood). Table 8 summarizes the 95% Confidence intervals based on the Mantel Haenszel, three GENMOD, and PHREG procedures below:

Table 8: Standard Errors, Risk Ratios, and 95% Confidence Intervals by the Various Conditional Poisson Estimation Methods

<table>
<thead>
<tr>
<th>Method</th>
<th>Standard Error</th>
<th>Riskratio</th>
<th>RRLowerCL</th>
<th>RRUpperCL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Contingency Table 2: Mantel-Haenszel</td>
<td>0.0578</td>
<td>0.72050</td>
<td>0.64326</td>
<td>0.80700</td>
</tr>
<tr>
<td>GENMOD Table 4: Wald</td>
<td>0.0609</td>
<td>0.72050</td>
<td>0.63944</td>
<td>0.81183</td>
</tr>
<tr>
<td>GENMOD Table 5: Model-Based</td>
<td>0.0811</td>
<td>0.72050</td>
<td>0.61924</td>
<td>0.83831</td>
</tr>
<tr>
<td>GENMOD Table 6: Empirical-Based</td>
<td>0.0773</td>
<td>0.72050</td>
<td>0.61461</td>
<td>0.84463</td>
</tr>
<tr>
<td>PHREG Table 7: Partial-Likelihood</td>
<td>0.0861</td>
<td>0.72050</td>
<td>0.60860</td>
<td>0.85297</td>
</tr>
</tbody>
</table>

Note how the empirical-based 95% confidence interval from the GENMOD procedure approached the wider 95% confidence intervals from the PHREG procedure. The narrowest confidence intervals were computed using the Mantel-Haenszel and Wald (default) standard errors of PROC GENMOD. Cummings et al. (2003, 2004a, 2004d) used bootstrap simulation studies to find that Cox proportional hazards models (i.e., PROC PHREG) produced more accurate, but wider, risk ratio confidence intervals than the conditional Poisson regression methods given by PROC GENMOD. This helps to explain the wide-spread use of this technique in matched-paired cohort analysis.
especially when other covariates are added to the model. Thus, the remaining analysis of the matched cohort data will be conducted using PROC PHREG.

The following datastep creates dummy variables to combine Pickups and SUVs and separate out minivans, using passenger cars as a reference group. These variables, along with other occupant characteristics, will be added as covariates in the PHREG procedure.

```sas
data newsimfile;
set simfile;
if type = 'Car' then do;
pickupSUV = 0; minivan = 0;
end;
else if type in('Pickup', 'Utility') then do;
pickupSUV = 1; minivan = 0;
end;
else if type = 'Van' then do;
pickupSUV = 0; minivan = 1;
end;
run;
ods output ParameterEstimates=phreg2;
proc phreg data = newsimfile nosummary ;
model time*dead(0) = driver age60 unbelted pickupSUV minivan /
ties = breslow rl ;
strata crashed;
Title 'PHREG Hazard Ratios Using Maximum Likelihood Standard Error Estimates';
proc print data=phreg2;
  var hazardratio hrlowercl hruppercl ;
  format Hazardratio HRLowerCL HRUpperCL 12.5;
run;
ods output close;
```

Table 9 below shows how the risk (hazard) ratio changed from 0.72 (95% CI=0.61-0.85) in Table 7 to 0.85 (95% CI=0.68-1.05) after adjusting for driver/passenger, age, belt status, and vehicle type, indicating the reduced risk of fatality for drivers is not statistically significant.

<table>
<thead>
<tr>
<th>Variable</th>
<th>DF</th>
<th>Parameter Estimate</th>
<th>Standard Error</th>
<th>Chi-Square</th>
<th>Pr &gt; ChiSq</th>
<th>Hazard Ratio</th>
<th>95% Hazard Ratio Confidence Limits</th>
</tr>
</thead>
<tbody>
<tr>
<td>driver</td>
<td>1</td>
<td>-0.16786</td>
<td>0.10933</td>
<td>2.3573</td>
<td>0.1247</td>
<td>0.845</td>
<td>0.682 1.048</td>
</tr>
<tr>
<td>age60</td>
<td>1</td>
<td>0.91226</td>
<td>0.20574</td>
<td>19.6604</td>
<td>&lt;.0001</td>
<td>2.490</td>
<td>1.664 3.727</td>
</tr>
<tr>
<td>unbelted</td>
<td>1</td>
<td>1.01031</td>
<td>0.17318</td>
<td>34.0335</td>
<td>&lt;.0001</td>
<td>2.746</td>
<td>1.956 3.856</td>
</tr>
<tr>
<td>pickupSUV</td>
<td>1</td>
<td>-1.40737</td>
<td>0.17794</td>
<td>62.5585</td>
<td>&lt;.0001</td>
<td>0.245</td>
<td>0.173 0.347</td>
</tr>
<tr>
<td>minivan</td>
<td>1</td>
<td>-0.37062</td>
<td>0.36104</td>
<td>1.0538</td>
<td>0.3046</td>
<td>0.690</td>
<td>0.340 1.401</td>
</tr>
</tbody>
</table>

Adding covariates to the model showed increased the crude risk ratio from 0.72 to 0.85. There were significant effects due to age, belt use, pickups and SUVs (p<0.0001). No significant minivan effect occurred relative to passenger cars.
The DATA step statements below were used to combine the ODS Output datasets into a single dataset containing the hazard ratios for all covariates. Table 10 was formed using the PRINT procedure.

```plaintext
data phreg1;
set phreg1;
covariate = 'driver';
run;

data phreg2;
set phreg2;
covariate = 'driver age60 unbelted pickupSUV miniVan';
run;

***Combine tables by rows (Concatenate tables by Append method)***;
data combined_rr;
length variable $9 df 8 estimate 8 stderr 8 chisq 8 probchisq 8
hazardratio 8 hrlowercl 8 hruppercl 8 covariate $100;
set phreg1 phreg2;
keep variable estimate hazardratio hrlowercl hruppercl covariate;
run;

title1 'Multivariate Models, Adjusted Risk Ratios, and 95% CIs for';
title2 ' Various Covariates';
proc print data=combined_rr;
var covariate hazardratio hrlowercl hruppercl;
run;
```

<table>
<thead>
<tr>
<th>Obs</th>
<th>Covariate</th>
<th>HAZARDRATIO</th>
<th>HRLOWERCL</th>
<th>HRUPPERCL</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>driver</td>
<td>0.720</td>
<td>0.609</td>
<td>0.853</td>
</tr>
<tr>
<td>2</td>
<td>driver age60 unbelted pickupSUV minivan</td>
<td>0.845</td>
<td>0.682</td>
<td>1.048</td>
</tr>
<tr>
<td>3</td>
<td>driver age60 unbelted pickupSUV minivan</td>
<td>2.803</td>
<td>1.898</td>
<td>4.139</td>
</tr>
<tr>
<td>4</td>
<td>driver age60 unbelted pickupSUV minivan</td>
<td>2.625</td>
<td>1.895</td>
<td>3.636</td>
</tr>
<tr>
<td>5</td>
<td>driver age60 unbelted pickupSUV minivan</td>
<td>0.314</td>
<td>0.220</td>
<td>0.446</td>
</tr>
<tr>
<td>6</td>
<td>driver age60 unbelted pickupSUV minivan</td>
<td>0.836</td>
<td>0.417</td>
<td>1.676</td>
</tr>
</tbody>
</table>

The matched-pair cohort results support a general conclusion that, in certain circumstances, fatalities among drivers involved in frontal collisions were significantly lower than passengers. Controlling for factors that could confound the association of the risk of driver deaths, such as belt use, driver age, and vehicle type mismatch, increased the estimated risk ratio from 0.72 to 0.85 for all passenger vehicles; however, the confidence interval for the adjusted risk ratio included 1.00.

**SUMMARY AND CONCLUSIONS**

Comparison of PHREG and GENMOD GEE ODS models for risk ratios indicated that all methods computed the same risk (hazard) ratio as that found in the contingency table. However, the 95% confidence limits differed because of the way standard errors were computed by the procedures. PROC GENMOD has three ways of computing standard errors for a variety of regression models using general estimating equations.

One limitation of conditional Poisson regression is the standard error estimates of risk ratios used for constructing confidence intervals may be larger than they should have been (Cummings et al., 2000). McNutt et al. (2004) pointed out that Poisson regression produced wide confidence limits because the Poisson model overestimated binomial errors with common outcomes of matched pairs. They also went on to say that, although the confidence
intervals from Poisson regression may be wide (more conservative) than using stratified analysis, the confidence intervals “can be thought of as bounding the true confidence interval” (Am J Epidemiol, 2004, 157:943).

PROC PHREG was restricted to computing maximum likelihood estimates and had wider confidence limits than PROC GENMOD. Cummings (Am J Epidemiol, 2004; 159:213-215; Epidemiol Rev. 2003; 25:43-50; The Stata Journal, 2004;4:274-281), though, preferred the Cox proportional hazards (PROC PHREG) approach for matched-cohort analysis when the time to an event (e.g., crash) is of interest rather than just if the event occurred or not (Epidemiol Rev. 2003; 25:47), as indicated by his repeated use of the method in traffic injury research based on simulation studies.

A listing of the unique features of the GENMOD and PHREG procedures, along with the pros and cons of each, is presented in Table 12.

Table 12: Comparative Features of PROCs GENMOD and PHREG

<table>
<thead>
<tr>
<th>GENMOD</th>
<th>PHREG</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Uses REPEATED statement for strata variables (matched pairs); equivalent to PHREG’s STRATA statement</td>
<td>• Uses STRATA variable statement to match pairs; equivalent to GENMOD’s REPEATED statement</td>
</tr>
<tr>
<td>• Uses the “Events/Trials” syntax (individual-level data can be summarized such that observations contain the count of the number of events and trials for the particular patterns of covariates expressed in the “Events/Trials” syntax)</td>
<td>• Breslow method is used to break ties of discordant pairs</td>
</tr>
<tr>
<td>• Data can be arranged in 2x2 contingency table summary for matched pairs with cohort cells A, B, C. Requires dummy variable coding</td>
<td>• Uses binomial link function that conditions on the total count of two cases in which each case is binomially distributed (undefined if total count =0)</td>
</tr>
<tr>
<td>+ More general in fitting a wide range of log-linear models than PHREG</td>
<td>+ To estimate the Cox model, the time variable is specified along with the status variable, denoting that driver-passenger observations are “censored”.</td>
</tr>
<tr>
<td>+ Corrects for overdispersion</td>
<td>- Limited to modeling survival data</td>
</tr>
<tr>
<td>+ Optionally produces likelihood-ratio hypothesis tests</td>
<td>- Requires covariates to have numeric 0,1 dummy variable coding</td>
</tr>
<tr>
<td>+ Provides multiple methods for computing standard errors (and confidence Intervals)</td>
<td>- Does not handle CLASS variables. They have to be defined with dummy-variable coding in a DATA Step (In version 9.1, an experimental procedure, TPHREG, allows categorical variables to be included with a CLASS statement in Cox proportional hazards modeling analysis).</td>
</tr>
<tr>
<td>- Covariates have to be defined for specified effects via CONTRAST and ESITMATE statements using GLM-like specifications (use ±1 effect coding to define reference vs. other levels)</td>
<td>+ Standard error formula considered to be the preferred method for use in traffic injury research</td>
</tr>
<tr>
<td>- Onus is on the statistician/analyst to choose the appropriate model for his/her situation</td>
<td>- Limited to one method of computing standard errors (and confidence intervals)</td>
</tr>
</tbody>
</table>

REFERENCES


**ACKNOWLEDGEMENTS**

We wish to thank Marge Scerbo for her suggestions and support.

**CONTACT INFORMATION**

Your comments and questions are values and encouraged. Please contact the authors at:

Melvin T. Alexander  
National Study Center for Trauma and Emergency Medical Systems  
University of Maryland School of Medicine  
701 West Pratt Street, Room 544  
Baltimore, MD 21201-1032  
Work Phone: (410) 328-6844  
Fax: (410) 328-3699  
E-mail: mtalexander@som.umaryland.edu