Methodology Application:
Logistic Regression Using the CODES Data

Developed For:

Department of Transportation
National Highway Traffic Safety Administration (NHTSA)
National Center For Statistics and Analysis (NCSA)

Developed By:

Jonathan Walker, Ph.D.
Hughes Training, Inc.
Link Division
5111 Leesburg Pike, Suite 300
Falls Church, Virginia 22041

Under Contract OPM 91-2973

For:

U.S. Office of Personnel Management
Office of Employee Development Policy and Programs
Washington, DC 20044-7559

Project Title: CODES Data
Purchase Order Number: 95-PO90521
Work Order Number: 773418
Project Code: O2T368

April 30, 1996 (Revised September 6, 1996)

Table of Contents

Preface

This document was written in WordPerfect 6.1 and has two features only available when viewed on the computer. They are included especially to aid the student:

HYPERTEXT (jumping to connections in the document)
Green, bold, underlined text is **hypertext**, which, when you click on it, takes you to another part of the document, usually to define or otherwise reference the green text. To return from a definition, first turn on the *hypertext feature bar* by clicking **Tools/Hypertext** from the menu bar. Then click on the **Back** button to return you from whence you came.

QUATTRO PRO (experimenting with data in the tables)
In addition, IF you have Quattro Pro 6.02 (for Windows) spreadsheet installed, you can double-click on most of the tables to open an abbreviated version of Quattro Pro (this takes time to open). Then you can experiment with the data, at least in cells that are in *Italics*. Other cells are "protected." If you do experiment, do not resave this document or some of the references in the text to the tables will be wrong. "Comments" near the tables explain particulars.

This document was designed to be printed on a Hewlett-Packard LaserJet 4si. If a printer with less resolution is used, even a Hewlett-Packard LaserJet IIISi, the colors of the fonts in the tables may have to be changed before they are legible.

Section 1: General

1.1. Background

Congress directed the Secretary of Transportation, through the Intermodal Surface Transportation Efficiency Act (**ISTEA**) of 1991, to carry out a study or studies to determine the impact of safety belt and motorcycle helmet use. The Act required the report of findings to be submitted to Congress 40 months after funds had been made available by the Secretary for these studies. In order to carry out the studies described in the Act, the National Highway Traffic Safety Administration (**NHTSA**) used the resources provided in the legislation to fund states to develop Crash Outcome Data Evaluation Systems (**CODES**). NHTSA prepared a Report[1] to Congress based on these analyses.

The data sets resulting from CODES have been used to develop this report and its accompanying Technical Report[2], but are of much broader interest to NHTSA. Therefore, it would be useful to train selected analysts at the National Center of Statistical Analysis (**NCSA**) to use the existing CODES data.

In the past, research questions that needed data from disparate data sets required the construction of small-scale, labor intensive, hand-linked data sets. The states in the CODES project used a cost-efficient method (**probabilistic linkage**) of matching crash data to medical and insurance data. NHTSA wants to expand the establishment and use of these linked data sets within non-CODES states and territories. Therefore, it would also like to produce a training program for analysts in the state agencies corresponding to NCSA.

1.2. Scope and Objectives

This report addresses **logistic regression**, a powerful statistical analysis used extensively in the CODES Report to Congress. It allows more advanced analyses than two simpler types of analysis: the Chi-square test (which tests for the independence of two qualitative (**nominal**) variables), and linear regression (which analyzes relationships between continuous (**interval** or **ratio**) scales). It allows researchers using qualitative measures of effectiveness, such as ‘died versus survived,’ to investigate relationships between that measure and many other measures simultaneously, whether those other measures are qualitative or quantitative. In the future, the CODES data sets, described in the CODES Usage Manual (the deliverable for task 2 of this project), will be analyzed to answer many questions other than those addressed in the Report to Congress. In addition, the CODES states and other states will continue to collect and analyze linked data sets similar to those collected for the original project. This document will introduce logistic regression to analysts who have limited experience or no experience with it.

Section 2: Assumptions

This document is based on the following assumptions:

2.1. User's Experience

Although the document presents some material on the Chi-square test and linear regression, that information is a review rather than an introduction. The document assumes that the user has had a first course in applied statistics (not necessarily theoretical statistics), at the undergraduate level or higher, or has had the equivalent experience with statistics. In particular, the user should be familiar with:

- The four basic scales of measurement: **Nominal**, **Ordinal**, **Interval**, and **Ratio**;
- **Proportions** and simple Chi-square tests of association used to test differences in proportions;
- Basic **linear regression** for predicting one measure from other measures.

2.2. Dependent Variable Type

The dependent variable in the analyses using **logistic regression** will always be a dichotomous (binary, yes-no, true-false) variable, such as died versus survived, or injured versus not injured. Logistic regression does allow an ordinal variable, e.g. a rank order of the severity of injury from 0 to 4, as the dependent variable, but only binary severity measures are discussed in this document.

2.3. Software

The examples in this document use PROC LOGISTIC of SAS®[3], although there are other procedures within SAS that also do logistic regression (PROC CATMOD and PROC PROBIT). Of course, many other statistical software packages can compute logistic regression but they will not be discussed here.

2.4. Application, Not Theory

The thrust of the document is application of the logistic regression, not its underlying theory.

2.5. System Generalization

The CODES data at NHTSA reside only on VAX machines in the NHTSA Research and Development Data Center due to security agreements with the states and, in the case of the larger states, due to data storage requirements. As a result, descriptions of analysis examples will not be generalized beyond that system.

Section 3: Review of Pertinent Statistical Concepts

3.1. Causality

The statistical techniques discussed in this document measure **associations** between variables, but they do not guarantee **causality**. For example, just because urban crash locations are associated with lower mortality rates does not mean rural crash locations are the root cause of more deaths. As another example, just because lower posted speed limits are associated with lower mortality rates does not mean that higher speed limits **cause** more deaths. In both examples, the speed of the vehicles is a better reason for the associations. Greater forces are involved when a vehicle is going 55 miles per hour than 25 miles per hour. However, since NHTSA did not have vehicle speed available, it used these two variables as substitutes or **surrogates** for vehicle speed. Therefore, even if posted speed limit was a perfect **predictor** of injury level, it would not prove that higher speed limits **caused** more injuries.

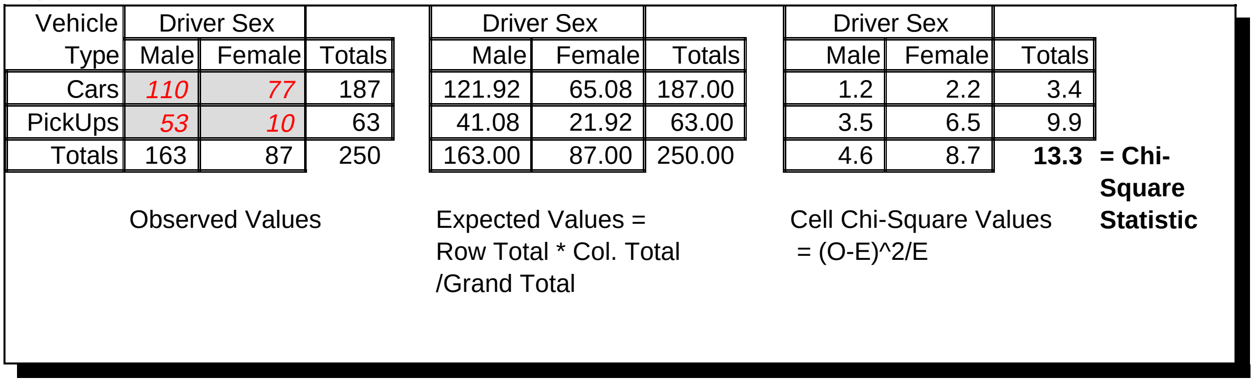
Causality must be established though experimentation, which controls all but one or a few of the variables thought to affect the outcome variable. In the case of human injury as the outcome variable, such experiments would be illegal and unethical. Therefore, studies are made of existing records, such as the Fatal Accident Reporting System (FARS) and the CODES data files, supplemented by experiments using instrument crash dummies, cadavers, or pigs as subjects.

3.2. Chi-square test of association among qualitative variables

This analysis tests for an *association* or *dependence* between two variables. More sophisticated versions can relate **more** than two variables, but in this document we will wait until we discuss **logistic regression** before addressing that situation. Almost any basic statistical textbook will discuss this test. Two books with more detail are Upton[4] and Fleiss[5].

3.2.1 Example: The simplest version of the Chi-square test deals with two binary variables. In this example, type of vehicle (car versus pick-up truck) and gender of driver will be used. One might expect drivers of pick-ups to be male more often than female. Actually, there are more male drivers for almost any vehicle type, so what we really mean is the **probability** of a driver being a male is higher if the vehicle is a pick-up (as opposed to a car). Statistically, this means there is an *association* or *dependence* between driver sex and type of vehicle.

Figure 1. Example of Chi-Square Test.



Vehicle	Driver Sex		
Type	Male	Female	Totals
Cars	*110*	*77*	187
PickUps	*53*	*10*	63
Totals	163	87	250

Observed Values

Driver Sex		
Male	Female	Totals
121.92	65.08	187.00
41.08	21.92	63.00
163.00	87.00	250.00

Expected Values = Row Total * Col. Total /Grand Total

Driver Sex			
Male	Female	Totals	
1.2	2.2	3.4	
3.5	6.5	9.9	
4.6	8.7	**13.3**	**= Chi-Square Statistic**

Cell Chi-Square Values = (O-E)^2/E

The rationale behind the test is shown in Figure 1. If there is **no** association, then the distributions of vehicle type will be the same for males, for females, and for the Row totals. In the Row Totals, about 75% of the vehicles are cars, so if there is no association, 75% of the males should be driving cars and 75% of the females should be driving cars. At the same time, if there is **no** association, then the distributions of sex will be the same for cars, for pickups, and for the Column Totals. In the Column totals about 66% of the drivers are male, so if there is no association, 66% of the cars should be driven by males and 66% of the pickups should be driven by males. Here is how the **Expected Values** are computed: For each cell, the row total for the cell is multiplied by the column total for the cell, and then divided by the grand total (187 * 163 / 250 = 121.92 for the upper left cell). If there is no association, the expected values will be near the observed values.

The Chi-Square statistic measures how much the observed value (O) in each cell is different from its expected value (E). Each difference (O-E) is squared, then divided by the expected value. The total of all the cell Chi-Squares is the Chi-Square statistic for the whole table.

3.2.2 Interpretation: If the Chi-Square statistic is relatively large, then there is an association between the two variables, that is, one can be predicted from the other. In the case of the 2 X 2 table, the statistic must be larger than 3.84 if you are using the 5% **level of significance**, or 6.63 for the 1% level of significance. With a value of 13.3, here we have a statistically significant difference. The statistic can range from 0 (if the expected values are exactly the same as the observed values) to the grand total (250 in this case). The latter means the variables are perfectly associated. However, this can occur in two directions: (1) All males drove pick-ups and all females drove cars, or (2) All females drove pick-ups and males drove cars. Note that the size of the Chi-Square statistic does not tell the direction of the association. It does not tell whether males are associated with pick-ups or cars. It is fairly obvious here that males are associated with pick-up trucks, but the direction is not so obvious in more complex situations. There are many other measures of association, some of which do show the direction of the association. We will deal with two of them, the **Odds Ratio** and **Relative Risk**, later.

3.2.3 Warnings: This Chi-Square Test is an approximation of tests which use the more sophisticated multinomial distribution. The Chi-Square formula given is the simplest possible. For a better test of the 2 X 2 table, Yates' correction is generally used, which gives a closer, more conservative approximation. If the expected values are **very** large, the difference is negligible. In the present example the Yates' corrected statistic is 8% lower than the uncorrected. When using SAS®[3], Yates' correction is called "Continuity Adj. Chi-Square." For tables larger than 2 X 2, Yates' correction does not apply and the formulas used in the table are appropriate.

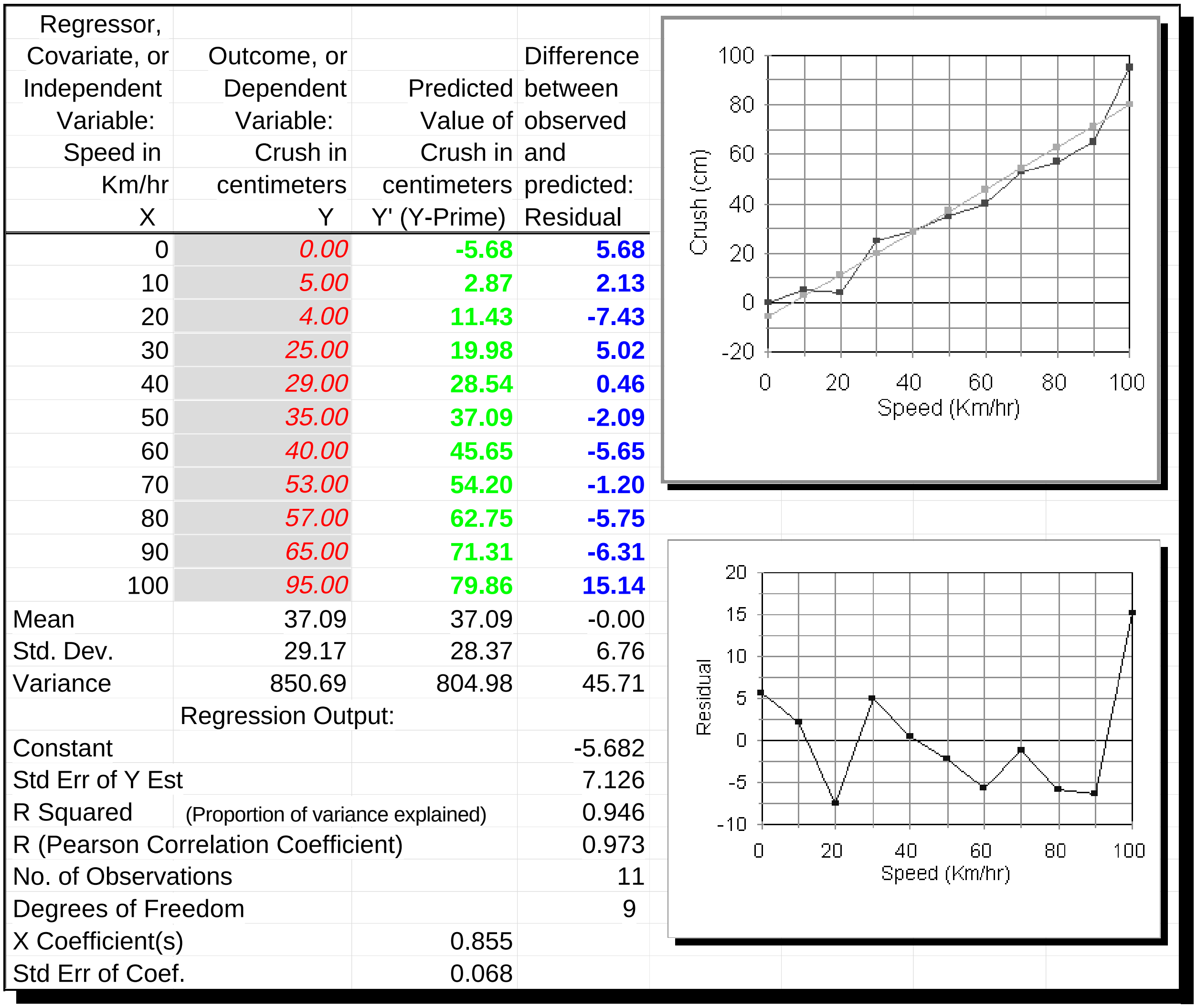
It is important to note that in large computerized data sets, it is easy to run hundreds of tests, and some of these tests will appear significant when actually it is only a random occurrence, and there is really no significant association. Two of the ways to test whether the association is real or not are to (1) check the literature to see if such an association has been significant elsewhere, or (2) run the same test on another independent sample (use data from another year or another location).

3.3. Linear Regression between quantitative variables

Linear regression measures the relationship between quantitative variables and is used to predict the outcome (dependent) variable in future situations. The simplest case is between two variables, such as age and height, or blood pressure and pounds overweight, or speed in a crash and inches of crush on the front of a vehicle. The last will be used as an example here. One variable (crush) is the outcome (dependent) variable, and the other (speed) is the independent variable, also called the covariate or regressor.

There are more sophisticated types of regression. In multiple regression, there is one outcome (dependent) variable and many covariates. In the CODES project, this type was used to investigate the effect of many covariates on the cost of hospitalization. Even more sophisticated is multivariate analysis, which has multiple outcome (dependent) variables and one or more covariates. This approach is required when there are repeated measurements over time on the same case (which, in different projects, could be a person, an intersection, or a state). More detail on linear regression analysis is given in Kleinbaum and Kupper[6].

Figure 2. Example of Linear Regression (Fictitious Data)



Regressor, Covariate, or Independent Variable: Speed in Km/hr X	Outcome, or Dependent Variable: Crush in centimeters Y	Predicted Value of Crush in centimeters Y' (Y-Prime)	Difference between observed and predicted: Residual
0	0.00	-5.68	5.68
10	5.00	2.87	2.13
20	4.00	11.43	-7.43
30	25.00	19.98	5.02
40	29.00	28.54	0.46
50	35.00	37.09	-2.09
60	40.00	45.65	-5.65
70	53.00	54.20	-1.20
80	57.00	62.75	-5.75
90	65.00	71.31	-6.31
100	95.00	79.86	15.14
Mean	37.09	37.09	-0.00
Std. Dev.	29.17	28.37	6.76
Variance	850.69	804.98	45.71
	Regression Output:		
Constant			-5.682
Std Err of Y Est			7.126
R Squared	(Proportion of variance explained)		0.946
R (Pearson Correlation Coefficient)			0.973
No. of Observations			11
Degrees of Freedom			9
X Coefficient(s)		0.855	
Std Err of Coef.		0.068	

3.3.1 Example: The table in Figure 2 shows some data from a fictitious research study in which engineers crashed ten identical cars into an immovable object, varying the speed at which the cars hit and measuring the centimeters of crush in the front end of each car. Since all variables except speed are held constant, one can say in this case that the speed *causes* the crush. Two lines are shown in the top graph: The crooked line (red on the screen or black on a print-out) shows the data as measured. The straight line (green on the screen or gray on a print-out) shows the predictions resulting from the linear regression. The **slope** of the straight line is the X coefficient (see the line in Figure 2 near the bottom of the table). This is also referred to as the 'parameter' for the covariate. The **intercept** is the predicted crush when speed is 0 (see the line labeled Constant just below Regression Output). These are the two computed constants in simple regression. In general, the formula for the line is $Y' = \beta_0 + \beta_1 X$, or in this case, $Y' = -5.682+0.855*X$. β_1 (0.855) is the parameter for the covariate 'speed.' *Y'* (y-prime) is the *predicted* value of crush.

3.3.2 Interpretation: If the observed points are very close to the line, as they are in this case, then the relationship between the variables is very strong. The amount of association is quantified by the Pearson **Correlation Coefficient**, which is 0.97 in this case, almost perfect. The line above, R Squared, shows how much of the variability in the crush data is explained by the linear regression. If speed was unknown, the best guess for crush would be 37.1 cm, and one measure of the variability of that estimate would be 850.69. If the speed was known, the best estimate would be given by the regression formula, and one measure of the variability around the regression line would be 45.71. The difference in these measures (850.69 - 45.71 = 804.98) is the variance explained by the regression formula, and the **proportion** of the variance explained by the regression results is 804.98/850.69 = 0.946. Note that this is R Squared.

This example showed a very strong relationship, as is common in a well-controlled engineering study. In crash investigation the associations will seldom, if ever, be this strong. Many of the statistics in the table, which are obviously significant in this case, may need to be tested when experimental controls are less rigorous. Kleinbaum and Kupper[6] show how to test whether the **slope** is significantly different than zero, whether the intercept is significantly different than zero, whether slopes (or intercepts) from two different sets of data are statistically different, and so forth.

As a rough test, however, you can look at the range of the parameter (the coefficient for speed), plus or minus two standard errors for the parameter. For the slope in this case, the range is 0.855-(2*0.068) to 0.855+(2*0.068), which is 0.719 to 0.991. Because this range does not include zero, the slope is significantly greater than zero. This is good, because you can use speed to predict crush. (In crash reconstructions, you could reverse the outcome measure and the covariate, re-run the regression analysis, and predict speed before impact from crush.) For the intercept in this case, the range is -5.682-(2*7.126) to -5.682+(2*7.126), which is -19.934 to 8.570. Because this range does include zero, the intercept is not significantly greater than zero. In this case, this is good because it is obvious that crush would be zero at zero speed.

3.3.3 Warnings: Linear regression is **not** appropriate when the data do not fit a straight line, as in Y = X^2. Graphing the data and the regression line will show large non-linear trends. The Residual column and the lower graph in the table allow closer inspection. The scatter of points above and below the zero residual line should be random. In this fictitious example, the residual at 100 Km/hr is the highest of all, showing more crush than expected. Above 90Km/hr, the relationship may become non-linear. This relationship is definitely nonlinear at and below zero, because the car does not expand if it sits at the barrier or if it backs away from the barrier. On the other hand, within the range examined, the data are extremely linear. Even in data that are obviously non-linear, small sections may be straight enough to be approximated, for engineering purposes, by linear regression. Common sense must be applied in all situations.

In addition, linear regression requires **interval** or **ratio** scales of measurement, since it assumes that for a constant change at any point on the independent variable, one can expect a constant change in the dependent variable. For instance, if the difference between 5 and 10 on the measurement scale is not the same as the difference between 35 and 40, then the linear regression formula will not be meaningful.

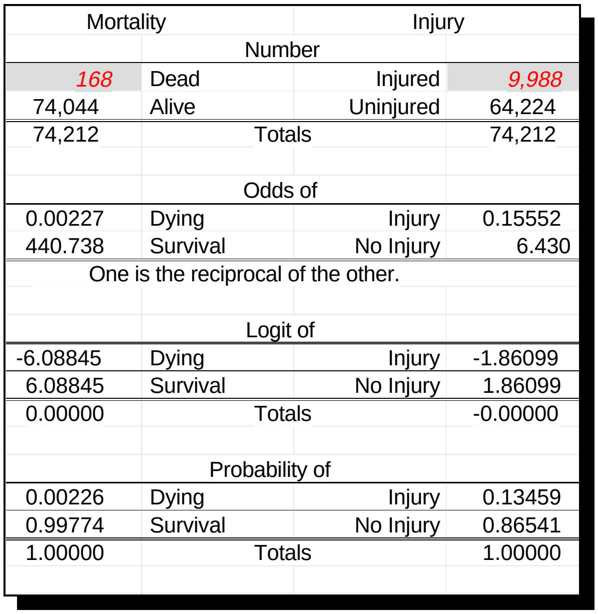
Section 4: Logistic Regression

4.1. Odds, Probabilities, and Logits.

All three of these terms describe how often something happens relative to its opposite happening, such as winning or losing, or dying or surviving. Thus, they all deal with a special case of nominal measurement scales: dichotomous (binary) outcome measures.

Figure 3 shows the differences among the three. The data are for drivers in police-reported crashes, derived from one of the CODES states. The identical totals are not a coincidence. It is the same data set, but with two different cut-points to make the dichotomies.

Figure 3. Comparison of Odds, Probabilities, and Logits.



Mortality		Injury	
	Number		
168	Dead	Injured	9,988
74,044	Alive	Uninjured	64,224
74,212	Totals		74,212
	Odds of		
0.00227	Dying	Injury	0.15552
440.738	Survival	No Injury	6.430
One is the reciprocal of the other.			
	Logit of		
-6.08845	Dying	Injury	-1.86099
6.08845	Survival	No Injury	1.86099
0.00000	Totals		-0.00000
	Probability of		
0.00226	Dying	Injury	0.13459
0.99774	Survival	No Injury	0.86541
1.00000	Totals		1.00000

Odds are one category divided by the **other**, (168/74,044) so odds for dying are the reciprocal of the odds for surviving. Thus the odds of dying are approximately 1 **to** 440, and the odds of injury are 1 **to** 6.5 (or 2 **to** 13 to use whole numbers). Odds can range from zero to plus infinity, with the odds of 1 indicating neutrality, or no difference.

The *logits* are simply the natural log of the odds [Ln(odds) or Log$_e$(odds)]. Note that the two logits are always symmetrical (they sum to zero). They range from minus infinity to plus infinity, but because they are logarithms, the numbers usually range from +5 to -5, even when dealing with very rare occurrences.

The *probabilities* are one category divided by the **total** (168/74,212). Note that they always sum to 1.000. The probability of dying is .00226, or approximately 2 **in** a thousand. The probability of injury is .13459, or roughly more than 1 **in** ten. The range of probabilities is zero to one. Note that when a very small number is in the numerator and a very large number is in the denominator, odds almost equal probabilities, but this is not true for the majority of cases.

4.2. Odds Ratios, Relative Risk, and Effectiveness.

Briefly, an odds ratio (O.R.) is the ratio of two odds, and relative risk (R.R.) is the ratio of two probabilities. Another dichotomous factor (called the regressor, covariate, or independent variable), such as safety-belt use, splits the data into two parts, and odds (or probabilities) are computed for each part. Generally, the top of each ratio is associated with the dangerous condition even though, for the CODES report, NHTSA decided to put the odds (or probabilities) for belt **users** (rather than non-users) on top. This was so effectiveness measures would be similar to those reported previously. **Effectiveness** is merely 100*(1- **Relative Risk**). Note that unlike the simple Chi-Square test of association, a direction is implied with **odds ratios** and relative risk. No one expects deaths to increase when safety belts are worn, or incidence of measles to be higher when students are not exposed to a child with the disease.

4.2.1 Example: The table in **Figure 4** continues the example from **Figure 3**, but now the odds and probabilities are separated for each belt condition. The dead and the alive are separated into belted and unbelted, and likewise for the injured and uninjured. The top section presents the raw numbers, the middle section shows the odds and odds ratio, and the bottom section shows probabilities, the relative risk, and effectiveness.

4.2.2 Interpretation: Look at the Mortality side. Your odds of dying while wearing a safety belt are 64 **to** 62,424, or roughly 1 **to** 1000. Your odds of dying without a belt buckled are 104 **to** 11,620, or roughly 1 **to** 100. The odds ratio, using belted divided by unbelted, is 0.1146. This is approximately 1/9, so you can say, "Your odds of dying with your safety belt *fastened* are roughly one ninth your odds of dying with your safety belt *unfastened*." Or you can compute the inverse of 0.1146, which is 8.73, and say, "Your odds of dying if your safety belt is *unfastened* are roughly nine times higher than your **odds** of dying if your safety belt is *fastened*."

The interpretations of the **odds ratio** and **relative risk** are similar, and the actual numbers are very close when dealing with fatalities. However, the differences are greater when dealing with any injury (which is not so rare). Here the odds of injury, given the driver was belted, are 19 **to** 100, but the **probability** in the same situation is 16 **in** 100. The odds ratio is 0.371 but the relative risk is 0.472. The important point is that the two numbers are based on the same data, and have similar meanings. Trouble arises when people compare two studies (where one study uses odds and the another uses probabilities) and do not realize the statistics are different.

For **Effectiveness**, one could say, "If unbelted drivers had worn their belts, 88% of those who died would have survived the crash." For the injury figures, one could say, "If all unbelted drivers had been wearing their belts, 53% of those who were injured would not have been injured."

4.2.3 Warnings: These figures have not been adjusted for over-reporting (people who were unbelted often say they *were* belted). More realistic estimates of odds ratios and effectiveness are given in the Report to Congress[1].

With these measures, one must be very careful when assigning the values for the dependent and independent variables. Reversing which is a 0 and which is a 1 inverts the interpretation. For the outcome measures in this example, 1 meant dead (for mortality) or injured (for morbidity). For the covariate, 1 meant belted.

If 'dead' (the dependent variable) had been coded as 0, and 'survived' had been coded as 1, the odds ratio would be inversed (1 / 0.114552 = 8.73), but the relative risk would not, because it would be using two different probabilities. The probability of surviving, given belt was worn, = 0.998976, and given belt was not worn, = 0.991129. RR(surviving) = 1.007917, which is not the reciprocal of 0.115458. If 'unbelted' (the independent variable) had been coded as 1, the odds ratio would be inversed (1 / 0.114552 = 8.73 for mortality, and 1 / 0.371158 = 2.69 for injury), and the relative risk would be inversed also (1 / 0.115458 = 8.66 for mortality, and 1 / 0.471671 = 2.12 for injury).

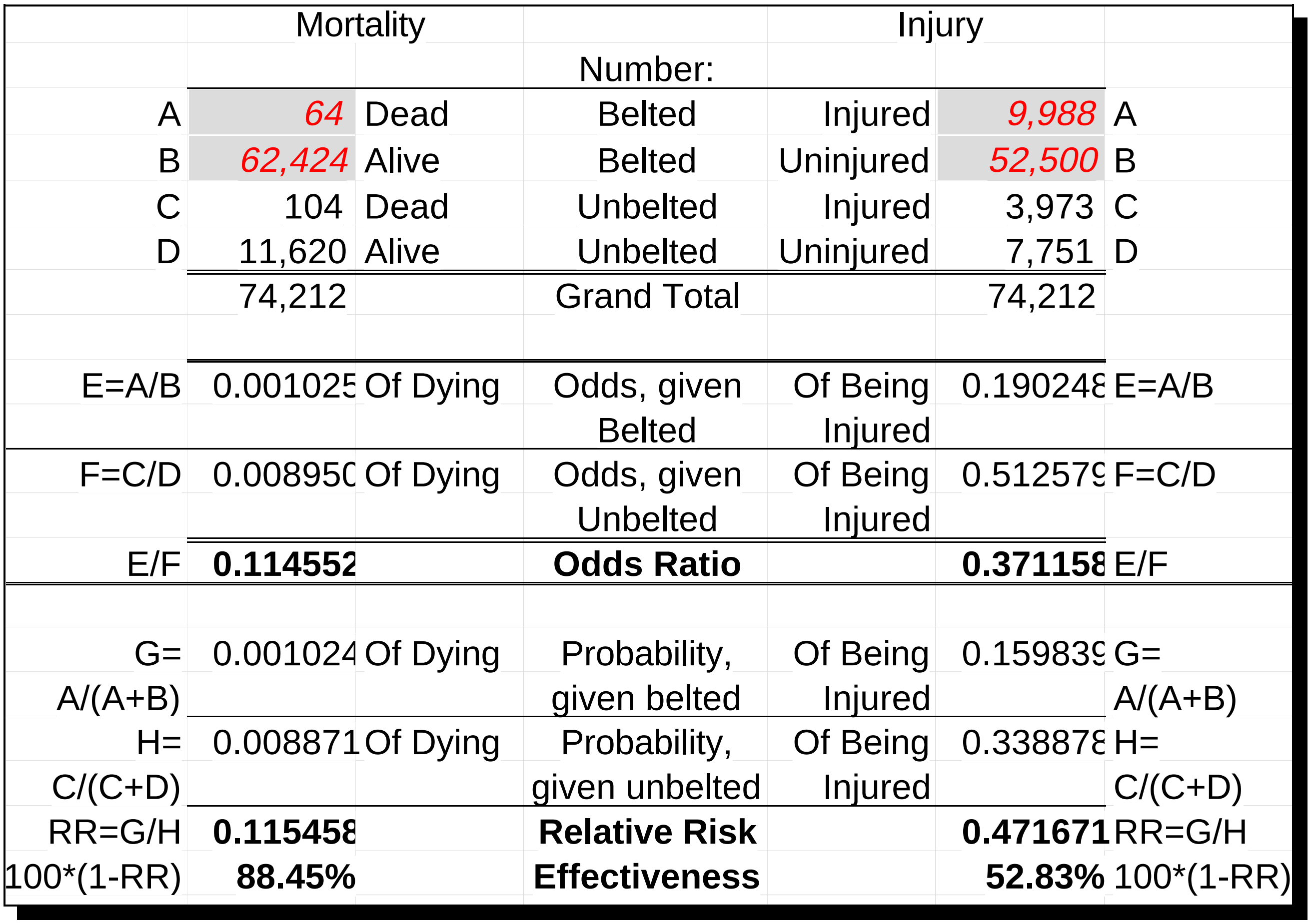
4.3. Regressions Using **Logits**

The logits have an advantage over odds or probabilities. Their neutral point is zero and they are symmetrical about zero. See **Figure 3**. The logit for dying (-6.08845) is the opposite of the logit for surviving (+6.08845). This is not true for odds or probability.

When we transform odds into logits and use logistic regression, we **assume** the resulting function is linear, and the problem can be treated, in a rough sense, like linear regression. As a minimum, we avoid the embarrassing problem of computing odds or probabilities that are less than zero, which might happen if we used odds or probabilities as raw data in a regression analysis. Thus, **logistic regression** allows **nominal** outcome variables to be compared with multiple covariates, which can be any measurement scale, nominal through ratio. For more background and detail in its use and interpretation, see Hosmer and Lemeshow[7].

4.3.1 Example: The table in **Figure 5** displays probabilities, odds, and logits for some hypothetical data which were manipulated to emphasize how logits **could** be more linear than the others. As before, vehicle drivers are separated into injured and uninjured, belted and unbelted, plus a measure of crash intensity called Delta-V.

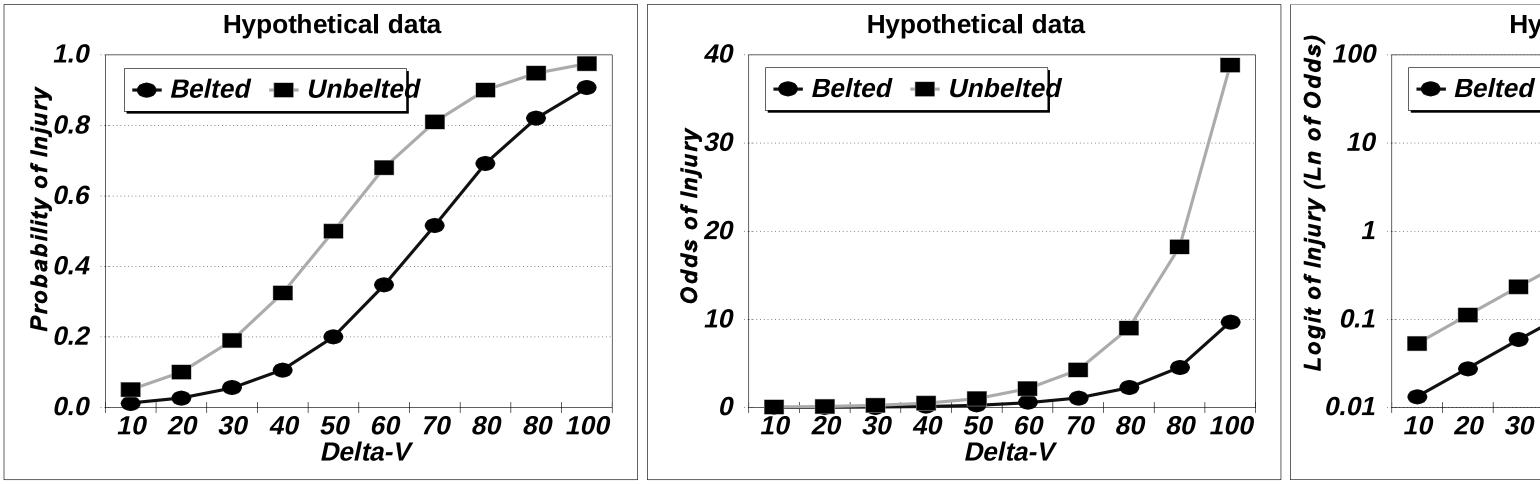
Figure 4. Odds Ratios, Relative Risk, and Effectiveness.



	Mortality			Injury		
			Number:			
A	64	Dead	Belted	Injured	9,988	A
B	62,424	Alive	Belted	Uninjured	52,500	B
C	104	Dead	Unbelted	Injured	3,973	C
D	11,620	Alive	Unbelted	Uninjured	7,751	D
	74,212		Grand Total		74,212	
E=A/B	0.001025	Of Dying	Odds, given	Of Being	0.190248	E=A/B
			Belted	Injured		
F=C/D	0.008950	Of Dying	Odds, given	Of Being	0.512579	F=C/D
			Unbelted	Injured		
E/F	**0.114552**		**Odds Ratio**		**0.371158**	E/F
G=	0.001024	Of Dying	Probability,	Of Being	0.159839	G=
A/(A+B)			given belted	Injured		A/(A+B)
H=	0.008871	Of Dying	Probability,	Of Being	0.338878	H=
C/(C+D)			given unbelted	Injured		C/(C+D)
RR=G/H	**0.115458**		**Relative Risk**		**0.471671**	RR=G/H
100*(1-RR)	**88.45%**		**Effectiveness**		**52.83%**	100*(1-RR)

Figure 5. Relations among Probability, Odds, and Logits. Hypothetical Data on the Relationships among Injury, Belt U
Speed upon Impact).

Col A	Col B	Col C	Col D	Col E	Col F	Col G	Col H	Col I	Col J	Col K	Co
	Belted				Ln(Col D)	**Unbelted**				Ln(Col I)	(D
			Odds of	Probability	Logit			Odds of	Probability	Logit	Od
Delta-V	Injured	Uninjured	Injury	of Injury	Ln(Odds)	Injured	Uninjured	Injury	of Injury	Ln(Odds)	Ra
10	131	9869	0.013	0.013	-4.322	504	9496	0.053	0.050	-2.936	0
20	271	9729	0.028	0.027	-3.581	1003	8997	0.111	0.100	-2.194	0
30	554	9446	0.059	0.055	-2.836	1900	8100	0.235	0.190	-1.450	0
40	1074	8926	0.120	0.107	-2.118	3250	6750	0.481	0.325	-0.731	0
50	2000	8000	0.250	0.200	-1.386	5000	5000	1.000	0.500	0.000	0
60	3469	6531	0.531	0.347	-0.633	6800	3200	2.125	0.680	0.754	0
70	5159	4841	1.066	0.516	0.064	8100	1900	4.263	0.810	1.450	0
80	6923	3077	2.250	0.692	0.811	9000	1000	9.000	0.900	2.197	0
80	8200	1800	4.556	0.820	1.516	9480	520	18.231	0.948	2.903	0
100	9066	934	9.707	0.907	2.273	9749	251	38.841	0.975	3.659	0



4.3.2 Interpretation: As Delta-V goes up, so does the severity of the crash, as can be seen from the steadily increasing **odds**, **probabilities**, and **logits** of injury as Delta-V increases. There is also a safety effect between the belted and unbelted groups. On every row, the odds, probabilities, and logits of injury are lower for belted drivers than for unbelted drivers.

The graphs at the bottom of the figure reveal the patterns among the three measures. In the probability graph, on the left, are two curves showing the familiar ogive shape: little change near zero probability, the sharpest increases between probabilities of 0.4 and 0.6, and decreasing change as probabilities approach 1.0. However, note that the ‘belted’ curve is always lower than the ‘unbelted’ curve, showing the safety effect.

The odds graph, in the middle, appears to show overlap at the low end when there actually is none. As can be seen from column L, there is a constant ratio between the belted odds and the unbelted odds, also showing that belted drivers have a safety advantage. Remember that odds are not symmetric: the neutral point for odds is one, and the space between 0 and 1 holds as much information as the space between 1 and infinity.

The logit graph, on the right, takes care of this problem. The constant **odds ratio** becomes a constant difference (-1.38) between the logits for belted and unbelted. (Remember that subtraction of logarithms is the same as division with regular numbers.) The last column shows that the exponent of the difference between the logits is the same as the odds ratio.
$e^{(\text{Delta Logits})}$ = Odds Ratio (e = approximately 2.718).

In short, at a very low Delta-V, few drivers are hurt; at a very high Delta-V, many drivers are hurt; but the effect of being belted is constant across all Delta-V’s. (Remember that these are hypothetical data: Such well-behaved data are very rare if they exist at all. Also remember that making a logit transformation does not force the data to be linear, but it often makes it more linear.) Finally, although the logit transformation is the one used in PROC LOGISTIC, there are other transformations which are also used and might prove more useful to you. See PROC CATMOD for other possibilities.

4.4. Another Example Using One Covariate and Real Data
This example is the simplest type, and is analogous to a 2 X 2 table. The outcome variable (M_OutC_A) is injury versus no injury, and the covariate (BeltUse) is safety-belt use. The SAS®[3] output on the next page is from Wisconsin. In the program log, lines in **Boldface** are the original program lines, regular text was added by SAS. Lines in *Italics* are additional notes. Following the SAS output are explanatory notes.

```
/*Jon Walker. CODES.   PROC LOGISTIC for Wisconsin data
* From file WiLog2.SAS;1
*/

OPTIONS LS=80 PS=64 NOCENTER NOOVP ; Note 1


PROC FORMAT;
  VALUE Yes2No
```

```
      0 = "No"
      1 = " Yes " /*Space before ‘ Yes’ makes odds ratios go in right
direction*/  Note 2
      ;
RUN ;

LIBNAME CRASH '############## '; Names have been deleted for security
purposes.
/* DataSet within ############# is ######### */
/* Includes proper vehicles (and cycles) for mandated model. */

DATA ;
   SET CRASH.######## ;
Note 3  IF CaseType = 1 /* Car/van/pickup Drivers only*/ ;
ATTRIB M_OutC_A FORMAT=Yes2No. LABEL="Any Injury " ;  Note 4
ATTRIB M_OutC_B FORMAT=Yes2No. LABEL="Transported/worse " ;
ATTRIB M_OutC_C FORMAT=Yes2No. LABEL="Hospitalized " ;
ATTRIB M_OutC_D FORMAT=Yes2No. LABEL="Fatal Injury " ;
/* Reformat Injury Levels: Put Odds Ratios in right direction. */
RUN ;

PROC LOGISTIC SIMPLE ;  Note 5
  MODEL M_OutC_A = BeltUse ;  Note 6
TITLE1 ' CODES: WI: ############# dataset. ' ;
TITLE2 "PROC LOGISTIC (BeltUse only) for the odds of any injury, ";
TITLE3 "for Car/van/pickups drivers. " ;
RUN ;

<<<SAS LOG>>>      <<<SAS LOG>>>      <<<SAS LOG>>>
1 The SAS System                           16:59 Wednesday, August 17, 1994

NOTE: Copyright (c) 1989-1992 by SAS Institute Inc., Cary, NC, USA.
NOTE: SAS (r) Proprietary Software Release 6.08  TS410
      Licensed to U.S. DEPARTMENT OF TRANSPORTATION/NHTSA, Site 0003589006.

NOTE: Running on VAXSTATION Model 4000-90 Serial Number 13000202.

1          /*Jon Walker. CODES.   PROC LOGISTIC for Wisconsin data
4          * From file WiLog2.SAS;1
5          */
6
7          OPTIONS LS=80 PS=64 NOCENTER NOOVP ;  Note 1
8
9          PROC FORMAT;
10           VALUE Yes2No
11          0 = "No"
12          1 = " Yes " /*Makes odds ratios go in right direction*/
13          ;  Note 2
NOTE: Format YES2NO has been output.
14         RUN ;
NOTE: PROCEDURE FORMAT used the following computer resources:
      Buffered IO:         40   Elapsed time:        0 00:00:02.42
      Direct IO:           40   CPU time:            0 00:00:00.22
      Page Faults:        704
```

```
15
16          LIBNAME CRASH '############# :############# ';
NOTE: Libref CRASH was successfully assigned as follows:
      Engine:        V608
      Physical Name:############# :#############
17          /* DataSet within [############# ] is #############  */
18          /* Includes proper vehicles (and cycles) for mandated model. */
19
20          DATA ;
21             SET CRASH.############# ;
22   Note 3   IF CaseType = 1 /* Car/van/pickup Drivers only*/  ;
23          ATTRIB M_OutC_A FORMAT=Yes2No. LABEL="Any Injury " ;Note 4
24          ATTRIB M_OutC_B FORMAT=Yes2No. LABEL="Transported/worse ";
25          ATTRIB M_OutC_C FORMAT=Yes2No. LABEL="Hospitalized " ;
26          ATTRIB M_OutC_D FORMAT=Yes2No. LABEL="Fatal Injury " ;
27          /* Reformat Injury Levels: Put Odds Ratios in right direction. */
28
29          RUN ;
NOTE: The data set WORK.DATA1 has 167642 observations and 44 variables.
NOTE: DATA statement used the following computer resources:
      Buffered IO:        594   Elapsed time:        0 00:02:20.59
      Direct IO:         2597   CPU time:            0 00:00:26.94
      Page Faults:       1144
30
31
32          PROC LOGISTIC SIMPLE ;  Note 5
33            MODEL M_OutC_A = BeltUse ;  Note 6
34          TITLE1 ' CODES: WI: ################ dataset. ' ;
35          TITLE2 "PROC LOGISTIC (BeltUse only) for the odds of any injury, ";
36          TITLE3 "for Car/van/pickups drivers. " ;
37          RUN ;


NOTE: PROC LOGISTIC is modeling the probability that M_OUTC_A=' Yes'. One way
      to change this to model the probability that M_OUTC_A='No' is to specify
      the DESCENDING option on the PROC statement. Refer to Technical Report
      P-229 or the SAS System Help Files for details.
NOTE: At least one W.D format was too small for the number to be printed. The
      decimal may be shifted by the "BEST" format.
NOTE: The PROCEDURE LOGISTIC printed page 1.
NOTE: PROCEDURE LOGISTIC used the following computer resources:
      Buffered IO:        223   Elapsed time:        0 00:03:25.17
      Direct IO:         4927   CPU time:            0 00:01:39.48   Note 7
      Page Faults:       1037
NOTE: The SAS Session used the following computer resources:
      Buffered IO:       1510   Elapsed time:        0 00:16:44.45   Note 8
      Direct IO:        18227   CPU time:            0 00:05:41.32
      Page Faults:       5739
NOTE: SAS Institute Inc., SAS Campus Drive, Cary, NC USA 27513-2414
```

```
-----------------------Start of SAS Output Listing----------------------------
 CODES: WI: ################ dataset.      16:59 Wednesday, August 17, 1994
1
PROC LOGISTIC (BeltUse only) for the odds of any injury,
for Car/van/pickups drivers.

The LOGISTIC Procedure

Data Set: WORK.DATA1
Response Variable: M_OUTC_A  Any Injury
Response Levels: 2
Number of Observations: 167642  Note 9
Link Function: Logit  Note 10


      Response Profile

Ordered
  Value  M_OUTC_A     Count

      1   Yes         17726  Note 11
      2  No          149916


                 Simple Statistics for Explanatory Variables

                                   Standard                                Variable
Variable M_OUTC_A         Mean    Deviation      Minimum      Maximum   Label

BELTUSE   Yes Note 12 0.668848     0.470641            0     1.000000
         No   Note 13 0.896435     0.304696            0     1.000000
                  ------------ ------------ ------------ ------------
         Total Note 14 0.872371     0.333677            0     1.000000


                    Criteria for Assessing Model Fit

                            Intercept
              Intercept        and
Criterion       Only       Covariates    Chi-Square for Covariates

AIC           113163.40     107445.05          .
SC            113173.43     107465.11          .
-2 LOG L      113161.40     107441.05     5720.353 with 1 DF (p=0.0001)
Score             .             .Note 15  7374.305 with 1 DF (p=0.0001)

                 Analysis of Maximum Likelihood Estimates

            Parameter Standard    Wald       Pr >    Standardized     Odds
Variable DF  Estimate   Error  Chi-Square Chi-Square   Estimate      Ratio
               Note 16
INTERCPT 1    -0.9727   0.0153  4029.8457     0.0001            .    0.378
BELTUSE  1    -1.4553   0.0181  6485.1911     0.0001    -0.267717    0.233
               Note 17                                                Note 18

Association of Predicted Probabilities and Observed Responses

 Concordant = 29.7%          Somers' D = 0.228
 Discordant =  6.9%          Gamma     = 0.622
 Tied       = 63.4%          Tau-a     = 0.043
 (2657411016 pairs)          c         = 0.614
```

Explanation of notes in preceding listing:

NOTE Explanation

1. Sets SAS®[3] options for run: LS=80: output is 80 columns wide; PS=64: 64 lines per page; NOCENTER: Left justify lines; NOOVP: Do not over-print (over-printing was used on mainframe ‘line printers’ to simulate a **BOLD** typeface).

2. PROC FORMAT sets up an output format that can later be assigned to values of a variable using an **ATTRIB** statement. In this case, the space before the Y in ‘ Yes’ also changes the order of the outcome variable when SAS runs PROC LOGISTIC. This is one of several possible ways to make sure SAS interprets the values of the outcome variable in the direction NHTSA wanted. Another way, which does not need any formatting, is to add the word DESCENDING: Line 32 in your program file would look like: PROC LOGISTIC SIMPLE DESCENDING ; NOTE: If you are using SAS version 5, you will use PROC LOGIST (no ‘ic’) and the direction of the parameters will be reversed, so you will not need DESCENDING or the space before the Y in ‘ Yes’.

3. CaseType was a variable used by Wisconsin to separate motorcycle cases from passenger car cases.

4. The ATTRIBute statement assigns a output FORMAT (either predefined by SAS or, as here, defined in a PROC FORMAT step) and a LABEL to a variable. This improves the appearance and understandability of the output. The statement can also be used to assign an INFORMAT and LENGTH, both of which can be used when creating a SAS data set.

5. Beginning of the Logistic Regression Procedure. SIMPLE generates basic statistics about each covariate (independent factor).

6. The outcome variable (only one) goes on the left of the equals sign, all covariates go on the right. The covariates must be numeric (not character variables), even though they usually stand for a qualitative difference, such as male versus female. There is one way to have two variables on the left (SAS, pages 1073-1075), but it does not apply to the data file structure generally used in this type of research.

7. This procedure took about one and two-thirds minutes in the central processing unit of a VAX computer. This is a long time, but would have been up to ten times longer on a PC.

8. This total time on the VAX includes three other simple PROC LOGISTICs not shown in this listing.

9. The total of ‘full’ observations (drivers), i.e., those with no missing data on ALL of the variables in the model. If your data set includes some observations with missing values, SAS will give you a warning telling you how many observations were deleted due to

missing values. Wisconsin used data files that were restricted to 'full' observations so the warning does not appear here.

10. Logit is the function SAS uses by default and is appropriate for these analyses.

11. The number of injured drivers on this line, the number of uninjured drivers on the next.

12. Basic statistics for BeltUse, but only for injured drivers. Because 0 is unbelted and 1 is belted, the mean is also the **proportion** wearing their belts. **If other values are used, this will not be true.** (SAS does not require 0 and 1). The mean multiplied by the number of injured drivers gives the number of belted, injured drivers. If the covariate is continuous, such as age, then the mean will be the average age for the injured drivers, and similarly for the other statistics.

13. Basic statistics for BeltUse, but only for uninjured drivers. As expected, a higher proportion of the uninjured drivers were wearing their safety belts. The mean multiplied by the number of uninjured drivers gives the number of belted, uninjured drivers.

14. Basic statistics for BeltUse for all drivers.

15. First, ignore the column headed 'Intercept Only.' All of the other 'Criteria' are ways to assess how well the model explains the data. In the simplest model, a 2 X 2 analysis, 'Score' will be identical to the simple Chi-square statistic. The 'Criteria' are most useful in comparing results from different models of the same data. In the right column, a higher Chi-Square is better than a lower Chi-Square, but for the tests in the middle column, a lower value is better. Values in the 'Intercept Only' column will not change as long as you are using the same data and same outcome variable.

16. This is the parameter (coefficient) for the intercept. It is the **logit** for injury if the driver was in the 'zero' group on all the covariates. The **odds** for injury appear in the right column. In this 2 X 2 example it is simply the parameter for the unbelted. For the intercept, the label on the last column is a misnomer, because it gives the odds of injury, not an odds ratio. It is not as useful if continuous covariates are included. For instance, if the covariates were belt use, driver age, and posted speed limit, then the predicted odds of injury would be for those drivers who were unbelted, **0** years old, and driving on a road with **0** speed limit.

17. This is the parameter for belt use. Because it is less than zero, then using the belt is associated with relatively lower odds of injury. Because the 'Pr > Chi-Square is less than **level of significance** chosen for this study, the effect is significant. If a covariate's parameter is significantly greater than zero, then the presence of the covariate (whatever was coded as 1) increases the odds of injury, relative to the absence of the covariate. If the parameter is significantly less than zero, then there is a safety effect because the odds are relatively lower in the covariate's presence.

The standard error is a measure of the precision of the parameter estimate. It must be small relative to the parameter or the effect will not be significant. When plotting the parameters, it is advisable to plot a 95% confidence interval. The top of the interval is the parameter plus (1.96 times the standard error), and the bottom is the parameter minus (1.96 times the standard error). Note that if you decide to plot **odds ratios** rather than parameters, you must compute the confidence intervals before changing to odds ratios. Plotting the odds ratios will then have unsymmetrical looking confidence intervals, but this is the correct way to do it.

In the case of a continuous covariate, such as age of the driver, the parameter may be extremely small, because it is the log of the ratio of **odds** for **one** year. When plotting these figures, it is advisable to multiply the parameter and the standard error by a constant, say 25 years, before calculating the confidence interval. This is especially true if the graph compares the continuous covariate to one or more dichotomous covariates.

18. This is the **odds ratio** of injury by BeltUse. It is derived from the parameter: Odds Ratio = $e^{Parameter}$. If **1** in the outcome variable is defined as something negative (here it is injury) then if the odds ratio is less than one, the covariate is a safety effect. If it is more than one, it is a danger effect. If it is not significantly different from one, it has no effect.

4.5. The Same Data Using Multiple Covariates

This section is very similar to the previous section, except it adds the following covariates: Male (versus female: sex of driver), Rural (versus urban: locality of crash), and Intersection (versus no intersection: location of crash). Only the most important parts of the output are reproduced here. Following the SAS®[3] output are more explanatory notes.

```
<<<SAS LOG>>>      <<<SAS LOG>>>      <<<SAS LOG>>>
16                              Note 2
17          PROC LOGISTIC SIMPLE DESCENDING ;
18            MODEL M_OutC_A = BeltUse Male Rural Inter ;Note 19
19          TITLE1 " CODES: WI: PROC LOGISTIC for the odds of any injury, " ;
20          TITLE2 " for Car/van/pickups drivers, 4 Covariates only. " ;
21          RUN ;

NOTE: PROC LOGISTIC is modeling the probability that M_OUTC_A=1.
NOTE: At least one W.D format was too small for the number to be printed. The
      decimal may be shifted by the "BEST" format.
NOTE: The PROCEDURE LOGISTIC printed pages 1-2.
NOTE: PROCEDURE LOGISTIC used the following computer resources:
      Buffered IO:        348   Elapsed time:        0 00:03:02.00
      Direct IO:         5709   CPU time:            0 00:01:26.59
      Page Faults:        934


(From the SAS output)
```

```
The LOGISTIC Procedure
Data Set: WORK.DATA1
Response Variable: M_OUTC_A  MRM Outcome A
Response Levels: 2
Number of Observations: 167642  Note 9
Link Function: Logit  Note 10

      Response Profile

Ordered
  Value  M_OUTC_A      Count
      1         1      17726 Note 11
      2         0     149916

                Simple Statistics for Explanatory Variables

                                      Standard
Variable  M_OUTC_A          Mean     Deviation       Minimum       Maximum

BELTUSE          1      0.668848      0.470641             0      1.000000
Note 20          0      0.896435      0.304696             0      1.000000
                    ------------  ------------  ------------  ------------
             Total      0.872371      0.333677             0      1.000000

MALE             1      0.536669      0.498668             0      1.000000
Note 21          0      0.597895      0.490325             0      1.000000
                    ------------  ------------  ------------  ------------
             Total      0.591421      0.491573             0      1.000000

RURAL            1      0.477152      0.499492             0      1.000000
Note 22          0      0.374570      0.484013             0      1.000000
                    ------------  ------------  ------------  ------------
             Total      0.385417      0.486695             0      1.000000

INTER            1      0.471962      0.499227             0      1.000000
Note 23          0      0.443115      0.496755             0      1.000000
                    ------------  ------------  ------------  ------------
             Total      0.446165      0.497095             0      1.000000

                    Criteria for Assessing Model Fit

                            Intercept
              Intercept        and
Criterion       Only       Covariates    Chi-Square for Covariates

AIC           113163.40     106092.25           .   Note 15  Note 24
SC            113173.43     106142.40           .
-2 LOG L      113161.40     106082.25      7079.147 with 4 DF (p=0.0001)
Score             .             .          8690.529 with 4 DF (p=0.0001)
```

```
CODES: WI: PROC LOGISTIC for the odds of any injury,                        2
for Car/van/pickups drivers, 4 Covariates only.

The LOGISTIC Procedure

                 Analysis of Maximum Likelihood Estimates

            Parameter Standard    Wald       Pr >    Standardized     Odds
Variable DF  Estimate   Error  Chi-Square Chi-Square   Estimate      Ratio
              Note 25
INTERCPT 1    -1.0681   0.0220  2348.3156     0.0001            .     0.344
BELTUSE  1    -1.4922   0.0183  6643.4968     0.0001    -0.274522     0.225 Note 26
MALE     1    -0.3600   0.0165   474.4372     0.0001    -0.097560     0.698
RURAL    1     0.5060   0.0170   887.2171     0.0001     0.135780     1.659
INTER    1     0.2481   0.0170   214.2238     0.0001     0.068000     1.282
              Note 27

Association of Predicted Probabilities and Observed Responses

 Concordant = 62.5%          Somers' D = 0.336
 Discordant = 28.9%          Gamma     = 0.368
 Tied       =  8.6%          Tau-a     = 0.064
 (2657411016 pairs)          c         = 0.668
```

Explanation of notes in preceding listing:

NOTE Explanation

19. To add more variables, simply list them on the right side of the = sign (no commas or operators). It should not matter which order they are in unless you use SELECTION= ... as an option (SAS®3 page 1080). All models in this document use a simple additive model where all covariates are accounted for simultaneously.

20. These are the statistics for belt use, and are identical to those given previously. Notes 12, 13, and 14) The only difference is that the values of M_OutC_A are not formatted.

21. These are the statistics for gender. Overall, the drivers are 59.1% male, but the injured are only 53.7% male, and the uninjured are 59.8% male. This means that comparing all levels of injury to no injury, males are slightly less likely to be injured. However, this does not hold when you compare fatalities to survivals.

22. These are the statistics for locality. Overall, 38.5% of the crashes were in rural localities, but 47.7% of the crashes with injury were rural. Remember that you can interpret these means as proportions or percentages only if the values for the covariate are 0 and 1.

23. These are the statistics for location. Overall, 44.6% of the crashes were at intersections, but 47.2% of the crashes with injury were at intersections.

24. The scores under the 'Intercept Only' column are identical to the previous model and to the next model. The scores in the 'Intercept and Covariates' column are slightly lower than the simple 2X2 model (Injury and BeltUse only) which means this model gives us a bit more information. The slightly higher Chi-Square scores lead to the same conclusion.

25. The intercept is interpreted similarly to **Note 16**, except that the 'zero' group refers to unbelted women drivers who were in urban crashes not at intersections. Their odds of injury are 0.344.

26. The odds ratio for BeltUse is slightly better that the previous model (0.233) which means adding the other three covariates made using belts look a little safer than without them. BeltUse is by far the most powerful covariate, because its standardized estimate (-0.2745) is farthest from zero.

27. The parameters and odds ratios for the other three covariates are smaller, but all are statistically significant at the 0.0001 level. Being male is associated with lower odds of injury, but crashing in a rural area or at an intersection are both associated with higher odds of injury.

4.6. Coding a Multiple-category Variable into Several Binary Variables

Binary covariates, like those used in the previous section, and continuous covariates (**interval** or **ratio** scales) like age or posted speed limit, can be entered into the model without any further coding as long as they are all numeric variables. However, variables such as crash type or seating position must be reformulated into matrices of binary variables, sometimes called 'Indicator Variables.' In the mandated model for CODES, there were five crash types (Rollover, Single-vehicle hit fixed object, Single-vehicle hit non-fixed object, Multiple-vehicle head-on, and Multiple-vehicle other). In cross-tabulations such as PROC FREQ, a single variable with five categories would be used, but here we use four binary variables (not five) which together give the same information as the multiple-category variable. This is the same process that is used in regular multiple regression (PROC REG).

The crash types that have a corresponding binary variable are defined by a '1' in that variable. A crash cannot be represented by a '1' in more than one variable. Put another way, the crash types are mutually exclusive. The last crash type, Multiple-vehicle-other in this case, is defined by '0's in all four binary variables. This category is called the 'reference category' because all other categories are compared to it. See the table on the next page.

Any category may be chosen as the reference category, but the results might be easier to interpret if you choose the one that has the lowest proportion on the outcome variable. In this instance, that translates to the crash type that has the lowest proportion of injury. This way, the PROC LOGISTIC parameters for the four binary crash variables would be positive, and all the odds ratios would be greater than one. According to this advice, 'Single-vehicle hits non-fixed object' should have been the reference category in the CODES project rather than the one that was chosen.

	BINARY MODEL VARIABLES			
CRASH TYPE	Roll	SVFO	SVO	MVH
Roll over (whether single or multiple vehicle)	1	0	0	0
Single vehicle hits fixed object (pole, tree, etc.)	0	1	0	0
Single vehicle hits non-fixed object (parked car, pedestrian, railway train)	0	0	1	0
Multiple-vehicle, head-on crash	0	0	0	1
Multiple-vehicle crash, other than head-on	0	0	0	0

The table below show the analogous table for seating position, which was collapsed into three categories: Driver, front-seat passenger, and rear-seat passenger.

	BINARY MODEL VARIABLES	
SEATING POSITION	Driver	FrntPas
Driver	1	0
Front Seat Passenger	0	1
Rear Seat Passenger	0	0

4.6.1 Interpretation of Odds Ratios from Multiple-category Variables.

For example, take the second table (Seating Position), and use ‘any injury versus no injury’ as the outcome variable. In New York, the odds ratio for Driver was 1.93 and the odds ratio for FrntPas was 1.35 (both highly significant). This means that the odds of injury to a driver were 93% higher than the odds of injury to a back seat passenger, and likewise, the odds of injury to a front seat passenger were 35% higher than the odds of injury to a back seat passenger. It does not tell you what the odds of injury were in any situation, only the ratios between them. You could conclude that the back seat was the safest place to be in a crash, and you might be tempted to conclude that the driver’s seat was the most dangerous. However, these two odds ratios do not test the difference between the driver and front seat passengers. The driver would have higher odds of injury, but the difference might not be significant.

4.7. An Example Using the Full Set of Mandated Covariates

The example below show the same analysis for Wisconsin with all the covariates. Driver and FrntPas are omitted because Wisconsin only used drivers. Because they did not have sufficient information on uninjured passengers, all passengers were omitted. Only the most important parts of the output are reproduced here. Following the SAS output are more explanatory notes.

From the SAS Log:

```
25
26          PROC LOGISTIC SIMPLE ;
27            MODEL M_OutC_A = BeltUse Roll SVFO SVO MVH Rural Age Note 6
28                 Male Slim Wet Time Inter PC ;
29          TITLE1 " CODES: WI: PROC LOGISTIC for the odds " ;
30          TITLE2 " of any injury, for Car/van/pickups drivers. " ;
31          RUN ;


NOTE: PROC LOGISTIC is modeling the probability that M_OUTC_A=' Yes'. One way
      to change this to model the probability that M_OUTC_A='No' is to specify
      the DESCENDING option on the PROC statement. Refer to Technical Report
      P-229 or the SAS System Help Files for details.
NOTE: At least one W.D format was too small for the number to be printed. The
      decimal may be shifted by the "BEST" format.
NOTE: The PROCEDURE LOGISTIC printed pages 1-3.
NOTE: PROCEDURE LOGISTIC used the following computer resources:
      Buffered IO:        714   Elapsed time:        0 00:05:17.34
      Direct IO:        10947   CPU time:            0 00:02:26.11
      Page Faults:       1314

///////////////////////// START of SAS output listing \\\\\\\\\\\\\\\\\\\\\\\\
CODES: WI: PROC LOGISTIC for the odds of any injury, for Car/van/pickups
drivers.

The LOGISTIC Procedure

Data Set: WORK.DATA1
Response Variable: M_OUTC_A  Any Injury
Response Levels: 2
Number of Observations: 167642  Note 9
Link Function: Logit  Note 10


      Response Profile

Ordered
  Value  M_OUTC_A     Count

      1   Yes         17726  Note 11
      2  No          149916
```

```
               Simple Statistics for Explanatory Variables

                                      Standard
Variable  M_OUTC_A          Mean     Deviation       Minimum       Maximum

BELTUSE    Yes Note 12  0.668848      0.470641             0      1.000000
          No   Note 13  0.896435      0.304696             0      1.000000
                    ------------  ------------  ------------  ------------
          Total Note 14 0.872371      0.333677             0      1.000000

ROLL       Yes          0.065835      0.248001             0      1.000000
          No            0.017196      0.130003             0      1.000000
Note 28             ------------  ------------  ------------  ------------
          Total         0.022339      0.147785             0      1.000000

SVFO       Yes          0.157057      0.363865             0      1.000000
          No            0.066937      0.249915             0      1.000000
                    ------------  ------------  ------------  ------------
          Total         0.076467      0.265744             0      1.000000

SVO        Yes          0.106172      0.308066             0      1.000000
          No            0.188752      0.391313             0      1.000000
                    ------------  ------------  ------------  ------------
          Total         0.180021      0.384206             0      1.000000

MVH        Yes          0.059461      0.236492             0      1.000000
          No            0.015402      0.123146             0      1.000000
                    ------------  ------------  ------------  ------------
          Total         0.020061      0.140208             0      1.000000

RURAL      Yes          0.477152      0.499492             0      1.000000
Note 22   No            0.374570      0.484013             0      1.000000
                    ------------  ------------  ------------  ------------
          Total         0.385417      0.486695             0      1.000000

AGE        Yes         34.817048     16.804835      8.000000     93.000000
Note 29   No           36.115531     16.611883      9.000000     99.000000
                    ------------  ------------  ------------  ------------
          Total        35.978233     16.637133      8.000000     99.000000

MALE       Yes          0.536669      0.498668             0      1.000000
Note 21   No            0.597895      0.490325             0      1.000000
                    ------------  ------------  ------------  ------------
          Total         0.591421      0.491573             0      1.000000

SPLIM      Yes         40.730847     12.763017      5.000000     65.000000
Note 30   No           37.166080     12.902392      5.000000     65.000000
                    ------------  ------------  ------------  ------------
          Total        37.543008     12.934223      5.000000     65.000000
```

```
The LOGISTIC Procedure

                Simple Statistics for Explanatory Variables

                                                 Standard
Variable  M_OUTC_A            Mean      Deviation       Minimum       Maximum

WET        Yes            0.350220       0.477052             0      1.000000
Note 31   No             0.384008       0.486362             0      1.000000
                      ------------   ------------  ------------  ------------
          Total           0.380436       0.485495             0      1.000000

TIME       Yes            0.257926       0.437506             0      1.000000
Note 32   No             0.184163       0.387618             0      1.000000
                      ------------   ------------  ------------  ------------
          Total           0.191963       0.393845             0      1.000000

INTER      Yes            0.471962       0.499227             0      1.000000
Note 23   No            0.443115       0.496755             0      1.000000
                      ------------   ------------  ------------  ------------
          Total           0.446165       0.497095             0      1.000000

PC         Yes            0.810674       0.391779             0      1.000000
Note 33   No             0.758545       0.427967             0      1.000000
                      ------------   ------------  ------------  ------------
          Total           0.764057       0.424588             0      1.000000


                    Criteria for Assessing Model Fit

                                Intercept
              Intercept            and
Criterion       Only          Covariates    Chi-Square for Covariates

AIC           113163.40      101099.40 Note 34  .
SC            113173.43      101239.81          .
-2 LOG L      113161.40      101071.40    12090.006 with 13 DF (p=0.0001)
Score             .              .        14481.074 with 13 DF (p=0.0001)


                 Analysis of Maximum Likelihood Estimates

               Parameter Standard    Wald       Pr >    Standardized     Odds
Variable DF   Estimate    Error  Chi-Square Chi-Square   Estimate       Ratio
              Note 35
INTERCPT 1     -2.4227    0.0466  2703.6718     0.0001            .     0.089
BELTUSE  1     -1.4185    0.0193  5388.7167     0.0001    -0.260958     0.242 Note 36
ROLL     1      1.1945    0.0429   774.3640     0.0001     0.097323     3.302 Note 37
SVFO     1      0.7324    0.0280   684.2170     0.0001     0.107309     2.080 Note 38
SVO      1     -0.6293    0.0301   436.6856     0.0001    -0.133311     0.533 Note 39
MVH      1      1.4680    0.0416  1243.6812     0.0001     0.113477     4.340 Note 40
```

```
RURAL     1     0.0477   0.0230     4.3057     0.0380     0.012801    1.049 Note 41
AGE       1    0.00207 0.000514    16.1678     0.0001     0.018950    1.002 Note 42
MALE      1    -0.3677   0.0174   449.0257     0.0001    -0.099652    0.692 Note 43
SPLIM     1     0.0269 0.000882   926.7895     0.0001     0.191530    1.027 Note 44
WET       1    -0.2661   0.0178   222.7600     0.0001    -0.071232    0.766 Note 45
TIME      1     0.2736   0.0210   170.1425     0.0001     0.059417    1.315 Note 46
INTER     1     0.4126   0.0191   464.5763     0.0001     0.113074    1.511 Note 47
PC        1     0.3327   0.0219   230.3157     0.0001     0.077872    1.395 Note 48


Association of Predicted Probabilities and Observed Responses

 Concordant = 72.5%          Somers' D = 0.460  Note 49
 Discordant = 26.5%          Gamma     = 0.464
 Tied       =  0.9%          Tau-a     = 0.087
 (2657411016 pairs)          c         = 0.730
```

Explanation of notes in preceding listing:

NOTE Explanation

28. Roll, SVFO, SVO, and MVH are the indicator variables that allow us to interpret crash-type effects, as discussed in section 4.6. These means can be interpreted similarly to those discussed before, but with the addition of a dependence among these four and the implied reference group, 'Multiple-Vehicle Other.' The proportions for the reference group can be found by subtracting the other four from 1. Thus, the proportion of the injured who were in 'Multiple-Vehicle Other' crashes is 1.0 - .065835 - .157057 -.106172 - .059461 = .611475. So 61% of those injured were in a 'Multiple-Vehicle Other' crash.

29. Age is a ratio scale, so the mean is not a proportion, but the average age of all who were injured, all who were uninjured, and everyone in the file. By comparing the injured and uninjured means, you would think younger people would be more likely to be injured and the parameter would be negative. However, the opposite is true. After adjusting for all the other variables, older people are more likely to be injured. However, more study needs to be done on the age effect.

30. Posted speed limit is another ratio scale, even if it does make jumps of 5 or 10 mph. As you might suspect, those who were injured were traveling on roads with higher speed limits on the average, although the difference seems small.

31. Wet refers to the slipperiness of the road. Any slick road, due to ice, snow, rain, or oil, was considered 'wet.' Surprisingly, more of the uninjured people were on wet roads compared to the injured people. Perhaps slick roads results in slower speeds and lower energy crashes.

32. Time is a binary classification of the time of the crash. Crashes between 8:00 p.m. and 3:59 a.m. were classed as 1, all others were 0. Alcohol-involved crashes are much higher between 8 p.m. and 4 a.m., so this variable served more as a surrogate for alcohol involvement than as a time variable. 26% of the injured were in this category, but only 18% of the uninjured were in this category.

33. PC is a surrogate for vehicle size. 1 means passenger car, 0 means pick-up truck or van. Vehicle weight, coded from the VIN, would have been preferable, but states either did not have it on the crash file or it was too seldom recorded. The breakdown by injury shows that 81% of the injured were in a lighter vehicle, while 76% of the uninjured were in the lighter vehicle. This would imply injury is slightly more likely in a car than in a pick-up or van, as is the case (see **Note 48**).

34. The AIC is now 101,099.40 with all thirteen covariates. With four covariates it was 106,092.25, and with one (BeltUse) it was 107,445.05. Thus, there is more information in the more complex model, although it is only a 6.2 percent decrease compared to the simplest model.

35. The intercept still refers to the ‘zero’ groups for all covariates, but because there are now 13 covariates, it has little practical value in itself, but is still useful for reconstructing probabilities of injury for specific groups.

36. The BeltUse parameter has remained very stable over all three models: -1.4553 with BeltUse alone, -1.4922 when three more covariates were added, and -1.4185 with all thirteen covariates. In every case, use of a safety belt significantly reduces one’s odds of injury.

37. Being in a rollover crash significantly increases one’s odds of injury relative to the reference crash group, “Multiple-Vehicle Crashes, other than Head-On.” This covariate is third in magnitude, after “Multiple-Vehicle Head-on Crashes” and BeltUse. With an odds ratio of 3.302, people in rollover crashes had more than three times the odds of injury than people in the reference crash group.

38. Single-vehicle crashes where a fixed object was struck are also relatively more dangerous than the reference crash group, with a parameter of 0.7324 and an odds ratio of better than two.

39. Single-vehicle crashes where a non-fixed object was struck are relatively safer than the reference group. By taking the reciprocal of the odds ratio: 1/0.533 = 1.876, one can conclude that “Multiple-vehicle, Other” crashes have odds of injury that are 87.6% higher than “Single-vehicle, Other” crashes.

40. “Multiple-vehicle, Head-on” crashes are most dangerous of all, with odds more than four times higher than the reference group. A rough test shows that they are significantly more dangerous than rollover crashes. By using the standard errors of the parameter, we can find the 5th percentile for the head-on crashes and compare it to the 95th percentile for the rollover crashes. If the two do not overlap, they are probably significantly different. For the head-on parameter, 1.4680 - (1.96 * 0.0416) = 1.3865 = the 5th percentile, and for the rollover parameter, 1.1945 + (1.96 * 0.0429) = 1.2786 = the 95th percentile. They do not overlap and we can reasonably conclude that head-on crashes are more likely to cause more injury.

41. A positive rural parameter (0.0477) and a chi-square probability of less than 0.05 (0.0380) mean that rural crashes are associated with higher odds of injury. However, not much can be made of the fact, since there are many variables that could explain the odds ratio of 1.049: Higher speed limits and poorer road design, to name two.

42. The age effect, although very small (0.00207), is very significant (0.0001). Remember that the parameter represents the effect for merely one year, so it may be more meaningful to look at difference for a score of years. Remember, if plotting the result, multiply both the parameter and the standard error by 20.

43. The negative parameter (-0.3677) and odds ratio less than one (0.692) indicate males are less likely to be injured. However, unlike most other covariates, the strength of the effect weakens when more extreme output measures are used. In a few states, when comparing died to survived, the sign of the parameter is reversed, meaning it is better to be female at that level of analysis.

44. The Speed Limit effect, also very small (0.0269), is also very significant (0.0001). Here the parameter represents the effect for one mile per hour, so it may be more meaningful to look at difference for a 25 mph. Remember, if plotting the result, multiply both the parameter and the standard error by 25. Crashes with more speed have more energy, and consequently more injuries.

45. Wet roads are relatively safer, according to this negative parameter. This agrees with the raw percentages in the simple statistics section (See **Note 31**).

46. Time, the alcohol surrogate, is associated with higher odds of injury. Note that in Utah, where per capita consumption of alcohol is lowest in the country, this variable is not significant.

47. Intersections are also associated with higher odds of injury.

48. If PC = 1, the person was in a passenger car; If PC = 0, in a truck or van. This odds ratio indicates cars are more dangerous, perhaps because the average car is lighter than the average truck.

49. Somer's D is an indicator of association that ranges from -1 to +1 with 0 meaning no association. It is used here to tell how well the predictions fit the observations. With only BeltUse, it was .228, with three more variables it was .336, and with nine more variables it is now .460.

Section 5: Summary

5.1. Uses of Logistic Regression

Logistic Regression is a great improvement over previous methods of analyzing dichotomous outcome variables for several reasons. It allows simultaneous investigation of many covariates, allowing separation of covariate effects and makes each effect more easily interpretable. The regression model allows prediction of high-risk groups so prevention dollars can be spent where they will be most effective. Finally, this approach also allows construction of more complex models than the straight additive approach, so interactions among the covariates can also be explored.

It should be noted that logistic regression uses the logit transformation, and there are other transformations that may result in a better fit of the data. In SAS®[3], PROC CATMOD, a more general procedure for categorical modeling, offers other transformations.

5.2. Advanced Topics

Other researchers both in NCSA and the CODES states, have investigated additional topics of interest using logistic regression. Interested parties should contact these people directly. For current phone numbers and addresses contact Dennis Utter (202 366 5351), or Sandra Johnson (202 366 5364).

Karl Kim, of Hawaii, (808 956 7381) has done extensive modeling with Hawaii's CODES data. He has developed a structural model to explain the relationships between certain driver characteristics and behaviors, crash types, and injury severity.

Because belt use estimates as reported by the occupants in crashes are higher than those reported by studies using independent roadside observers, Missouri, Utah, and Wisconsin have experimented with models for adjusting reported belt use. The adjusted data are then used to estimate new belt-use odds ratios.

Douglas Thompson, of Maine, (207 780 4682) has studied alternate models of safety-belt effectiveness using within-vehicle analysis to better control variability, and has compared these results to those derived from the standard model used in the CODES report[1].

Researchers in New York have used logistic regression to investigate differing injury patterns among older drivers.

Ellen Hertz of NCSA (202 366 5360) has explored the use of PROC CATMOD as an alternative to PROC LOGISTIC to compute relative risk.

Section 6: Definitions

Correlation Coefficient: A measure of association of quantitative variables. After plotting the relationship of two variables, such as height and weight, using Cartesian coordinates, if all points line up exactly (and the line is not vertical or horizontal) then the correlation is perfect: It will be +1.0 if it is a direct relationship (as one goes up, the other goes up), or -1.0 if it is an inverse relationship (as one goes up, the other goes down). If the points are scattered randomly, then the correlation is zero, and there is no relation. Its symbol is usually a lower-case r.

It is related to the regression coefficient (in simple regression) by the following formula: $r=\frac{S_X}{S_Y}\beta_1$, where β_1 is the regression coefficient (slope), S_X is the sample standard deviation of the X values, and S_Y is the sample standard deviation of the Y values. Its interpretation is different from the regression coefficient because the correlation coefficient is the same no matter which variable is considered the X and which the Y. For the regression coefficient it will make a difference. The two coefficients are also related in the following way: If you standardized the raw data, and computed the regression coefficient, it would be *identical* to the correlation coefficient. Standardizing means each X has its sample mean subtracted from it, then is divided by the sample standard deviation: $Z_X=\frac{(X_i-\bar{X})}{S_X}$. Likewise for the Y scores, $Z_Y=\frac{(Y_i-\bar{Y})}{S_Y}$.

Effectiveness: A measure of how effective a safety device/program is, expressed as a percentage. It can be interpreted as, "If all people not using the device had used the device, X percent of them would not have suffered the consequence." X is the effectiveness figure. It is derived from Relative Risk: Eff. = (1-R.R.)*100. See **Figure 4** for a comparison of **odds ratios**, **relative risk**, and, effectiveness.

Hypertext: A link that will let you move from one location in a current document to another place within that document, to another document, or to a macro.

Intercept: In graphs of straight lines using Cartesian coordinates, the value of Y when X = 0, or the point at which the line crosses the vertical (Y) axis. In **logistic regression**, the **logit** of a positive response (e.g., injury) when all the covariates are 0.

Interval Scale: A measurement scale in which different intervals along the scale have the same meaning. For example, both Fahrenheit and Centigrade temperature scales fall into this category, because on each scale (taken by itself), the 10 degrees between 20 and 30 degrees means the same as the 10 degrees between 90 and 100 degrees, at least from a physical standpoint. Also, where two interval scales measure the same thing, there is a multiplicative formula *which may include the addition or subtraction of a constant* relating the two scales: C° = (5/9)*(F°- 32°). Interval scales have all the properties of **Ordinal** and **Nominal** scales, but may not have a 'real zero' point, necessary for a **Ratio** scale.

Level of significance: Research is a gamble: There is so much variability in most data that we use statistical tests to decide whether the patterns we observe (differences among groups or associations among variables) are really true, or just chance occurrences. There are two ways we can be correct: The test says there was an association, and there really is, or the test says there is no association, and there really is not. There are also two ways we can be incorrect: The test says there was an association but there really is none (called a Type 1 error, Alpha [α] error, or 'false alarm'), or the test says there was no association but there really is one (called a Type 2 error, Beta [β] error, or 'miss'). (This β is not related to the regression coefficient.)

Type 1 errors are set by the researcher: A **5% level of significance** means a 5% **probability** of a Type 1 error, which in turn means that if there is really no association the researcher will make the wrong decision 5 times (on the average) out of 100 repetitions of the same study. It is important to note that in large computerized data sets, it is easy to run hundreds of tests, and some of these will very likely be Type 1 errors!

Type 2 errors are not further discussed except to note that the two errors must be balanced out. When a researcher selects a very low level of significance to reduce Type 1 error, she is raising the probability of a Type 2 error. See advanced texts on experimental design for further discussion.

Logistic Regression: A statistical technique for examining relationships between an outcome measure that is a **nominal** scale, and one or more other variables, often called covariates, which can be any type of scale. The outcome measure can also be called a dependent variable, and the covariates can be called independent variables or regressors. In the CODES project, one of the outcome variables was died (versus survived) and some of the covariates were safety belt use, sex, age, posted speed limit, roadway condition, and vehicle type. Effects are reported as the natural logarithms of the odds ratios (parameter estimates) and as **odds ratios**.

Logit: The natural logarithm of the **odds**. Where N_0 = the number of cases in one category, and N_1 = the number of cases in the other category, then the logit = $\log_e(N_1/N_0) = \ln(N_1/N_0)$. In terms of probability, where $p = N_1/(N_1+N_0)$, the logit = $\ln(p/[1-p])$. In this document, its main use is to allow linear regression when analyzing relationships with a dichotomous outcome variable (dependent variable). Its main drawback comes when a category has a zero count (an empty cell) which leads to an odds of 0 or undefined (division by 0). There is no logarithm for either situation, which invariably leads to a failed solution when trying to compute a **logistic regression**. In this case the variable with a zero must be removed from the model and the regression re-run. See **Figure 3** for a comparison of **odds**, **probability**, and **logits**.

Nominal Scale: A measurement scale in which the only differences are qualitative, not quantitative. Examples in crash research are safety belt use (yes or no) or crash type (rollover, single-vehicle hitting fixed object, single vehicle hitting movable object, multiple-vehicle with head-on collision, or multiple-vehicle with all other types of collision). This is the simplest type of measurement scale. Two nominal scales measuring the same thing can be related only if the definitions for all categories are exact matches.

Odds: A way to express how often something happens, relative to something else happening. They are often expressed as a ratio of two whole numbers. If a bettor says the odds are 3 **to** 2 (1.5:1) of the home team winning, it means she expects the home team to win 3 of every 5 games and the away team to win 2 of every 5 games. Note this is not the same as probability: The expected probability of the home team winning is 3/5 = 0.60. See **Figure 3** for a comparison of odds, **probability**, and **logits**.

Odds Ratio: A measure of the dependence of a variable, usually a dichotomous variable but sometimes an ordinal variable, on a second variable, which often is also a dichotomous variable. When both variables are dichotomous, it is a ratio of two odds, where one odds is affected by one part of the second dichotomous variable, and the other odds is affected by the other part of the second dichotomous variable. For example: Does a team win more at home than away? Here the first variable (the outcome or dependent variable) is Won versus Lost, and the second variable (the covariate, regressor, or independent variable) is Home versus Away. If their ‘home’ record is 32 won and 17 lost, and their ‘away’ record is 26 won and 24 lost, then the odds ratio would be (32/17) / (26/24), or 1.74. An odds ratio of 1.00 means the covariate (home versus away in this case) had no effect on the outcome variable. An odds ratio significantly greater than one means that the numerator of the covariate (in this case, playing at home) *in*creases the outcome in the numerator of the ratio (in this case, winning) relative to the denominator of the covariate (playing away). An odds ratio significantly less than one means the numerator of the covariate *de*creases the outcome in the numerator of the odds. An odds ratio is never negative, but can approach 0 at the lower limit and infinity at the upper limit. See **Figure 4** for a comparison of odds ratios, **relative risk** , and **effectiveness**.

Ordinal Scale: A measurement scale in which there are differences in quantity, such as the KABC0 scale of injury used by police, or the injury/treatment scale used in the CODES analysis of safety belts. Measurements on an ordinal scale are in order from high to low (or vice versa) but do not necessarily have the properties of **interval** or **ratio** scales, e.g. the difference between K and A is not necessarily the same as the difference between A and B, and the bottom of the scale is not necessarily a true zero even though this is the case in these two examples. Ordinal scales are also called Rank scales. In the social sciences, a large category of measures called Likert scales are also ordinal scales. An example of a Likert scale is: “How comfortable do you feel right now? Assign a number from 1 [very uncomfortable] to 5 [very comfortable].”

Probabilistic Linkage: A technique for linking computerized large data files efficiently. It automatically sorts the data files into manageable blocks of cases, weighs each datum that might be useful in a possible match, and computes a composite weight, or score. This weight measures the possibility that ‘case X from file A’ is the same as ‘case Y from file B.’ High weights indicate matches, negative weights indicate non-matches, and low-positive weights are questionable, and can be clerically verified. The technique allows population files, such as state-wide medical files and state-wide crash files, to be linked. Thus, questions can be researched with less time and money and with more resulting cases, compared to single-purpose, small scale studies that have to be limited due to enormous clerical costs.

Probability: A number, ranging from 0 to 1, expressing the likelihood of something happening. Using the example of a sports team, the probability of winning is the number of wins divided by the total number of games played. In this sense, it is the number of *successes* divided by the total number of *trials*. However, it can also represent other quantities that are not simple ratios, such as an area in a theoretical statistical distribution. See **Figure 3** for a comparison of probability, **odds**, and **logits**.

Proportion: In this document, proportion refers to *a part of the whole,* and thus is tied to probability. It is a ratio of two whole numbers, whereas a probability does not have to be such. Thus if a team has won 16 of their first 25 games, the *proportion* won can be expressed as 16/25, 64/100, 64%, or 0.64.

Ratio Scale: A measurement scale in which different intervals along the scale have the same meaning and there is a meaningful 'zero point.' For temperature, only the Kelvin scale is a ratio scale because it uses 'absolute zero,' meaning at 0° there is no heat at all, which is not true on the Fahrenheit and Centigrade scales. Two ratios scales measuring the same quantity are related by only a constant multiplier, *without* adding or subtracting a constant. For example, in length, 2.54 centimeters = 1 inch, and in angles, 2π/360 radians = 1°. A ratio scale includes all the properties of the lesser scales: **Nominal**, **Ordinal**, and **Interval**.

Relative Risk: A measure of the dependence of one dichotomous variable on another. It is the ratio of two probabilities. As the name implies, it is used to measure the effect of some risk factor (not wearing a safety belt, speeding, being male, etc.) on the probability of some negative outcome (losing, dying, becoming infected, etc.). For example, what is the effect of playing out of town on losing a game? Using the numbers from the problem discussed under the definition of the **odds ratio**, the probability of losing 'away' would be 24/(26+24) = 0.48, and the probability of losing at 'home' would be 17/(32+17) = 0.35. The relative risk would then be 0.48/0.35, or 1.256. In this case, a ratio of greater than 1 implies that playing 'away' games increases the chances of losing. Notice that this setup of the 'away' problem is the converse of that in the definitions of odds ratio. Nevertheless, the overall results are the same. Here, 'playing away' was associated with losing more, there, 'playing at home' was associated with winning more. Both statements are saying the same thing. See **Figure 4** for a comparison of relative risk, **odds ratios**, and **effectiveness**.

Slope: A measure (to be precise, the *tangent*) of the angle of a line on a graph which shows the relationship between two variables. A down slope is a negative slope (an inverse relationship between the variables), an up slope is a positive slope (a direct relationship), a flat line is a zero slope (no relationship: X changes but Y does not), and a vertical line is an "undefined" slope (also no relationship: Y changes but X does not. It is undefined due to division by zero). The short definition is 'rise over run,' which in terms of Cartesian coordinates means the change in the Y-coordinate (vertical) is divided by the change in the X-coordinate (horizontal). If the slope is +2.0, then for every increase of 1 in the independent variable (X), the dependent variable (Y) *in*creases 2. If the slope is -2.0, for every increase of 1 in the independent variable (X), the dependent variable (Y) *de*creases 2.

Section 7: References

1. *Report to Congress*: *Benefits of Safety Belts and Motorcycle Helmets, Based on Data from the Crash Outcome Data Evaluation System (CODES)*. DOT HS 808 347. Washington, DC: Department of Transportation, National Highway Traffic Safety Administration, February 1996.

2. Johnson, Sandra W. and Walker, Jonathan. *NHTSA Technical Report: The Crash Outcome Evaluation System (CODES)*. DOT HS 808 338. Washington, DC: Department of Transportation, National Highway Traffic Safety Administration, January 1996.

3. SAS Institute Inc., *SAS/STAT® User's Guide, Version 6, Fourth Edition, Volume 2,* Cary, NC: SAS Institute Inc., 1989, 846 pp.

4. Upton, Graham J. G. *The Analysis of Cross-tabulated Data.* New York: John Wiley & Sons, 1978.

5. Fleiss, Joseph L. *Statistical Methods for Rates and Proportions, 2nd Ed.* New York: John Wiley & Sons, 1981.

6. Kleinbaum, D. G. and Kupper, L. L. *Applied Regression Analysis and Other Multivariable Methods*. Belmont, CA: Wadsworth Publishing, 1978.

7. Hosmer, D. W., Jr. and Lemeshow, Stanley. *Applied Logistic Regression.* New York: John Wiley & Sons, 1989.