

In Sloboda BW (ED), 2009, Transportation Statistics,
Ft. Lauderdale, FL: Ross.

10

QUASI-LIKELIHOOD GENERALIZED LINEAR REGRESSION ANALYSIS OF FATALITY RISK DATA

C. Craig Morris

*Bureau of Transportation Statistics
Research and Innovative Technology Administration
U.S. Department of Transportation*

10.1 INTRODUCTION

Transportation-related fatality risk is a function of many interacting human, vehicle, and environmental factors. Statistically valid analysis of such data is challenged both by the complexity of plausible structural models relating fatality rates to explanatory variables and by the uncertainty regarding the probability distribution of the data. But, fortunately, generalized linear modeling and maximum quasi-likelihood estimation together provide an extraordinarily effective set of statistical tools for the analysis of such data. The goal of this chapter is to illustrate and promote the application of these tools to fatality risk analysis.

First, we review the basic principles of generalized linear modeling and quasi-likelihood estimation. Next we illustrate the use of these tools to analyze

the association of motorcyclist fatality rates with U.S.-state helmet laws while controlling for climate measures (annual heating degree days and precipitation inches) as proxies for motorcycling activity. The focus of the illustration is on Poisson log-linear regression analysis of over-dispersed fatality rate data via quasi-likelihood generalized linear modeling methods.

One popular reference on generalized linear modeling and maximum quasi-likelihood is McCullagh and Nelder's (1989) *Generalized Linear Models*. Another is Agresti's (2002) *Categorical Data Analysis*. Together, these venerable resources provide the essential theoretical and practical background for a variety of multivariate analyses on either continuous or categorical data, including analysis of variance and multiple linear regression, logistic, logit, probit, or log-linear regression, among others. What remains for application of these methods is the choice of statistical software (although one could implement the methods from scratch), and there are many choices. The example we review later in this chapter uses the SAS (2005) V8.0 procedure GENMOD (for generalized linear modeling). Methods for longitudinal, repeated-measures, spatially-correlated, or other correlated data are also available, but these are beyond the scope of this introductory tutorial; two fine reference texts for correlated data problems are Diggle et al. (2002) *Analysis of Longitudinal Data* and McCulloch and Searle (2001) *Generalized, Linear, and Mixed Models*.

10.1.1 Generalized Linear Models

Generalized linear models are an extension of *classical linear models*, so we begin with the latter. In classical linear modeling, a sample of n independent observations ($y_1, y_2, \dots, y_p, \dots, y_n$) is regarded as the realization of n independently distributed components ($Y_1, Y_2, \dots, Y_p, \dots, Y_n$) of a random variable Y with means ($\mu_1, \mu_2, \dots, \mu_p, \dots, \mu_n$). In the *systematic* part of the classical linear model, each mean, μ_p is regarded as a linear function of p explanatory variables ($x_{i0}, x_{i1}, \dots, x_{ij}, \dots, x_{ip-1}$), usually with $x_{i0} = 1$ for the intercept, that is,

$$E(Y_i) = \mu_i = x_{i0}\beta_0 + x_{i1}\beta_1 + \dots + x_{ij}\beta_j + \dots + x_{ip-1}\beta_{p-1} = \sum_j x_{ij}\beta_j \quad (10.1)$$

where, for $i = 1, \dots, n$, and $j = 0, 1, \dots, p - 1$, x_{ij} is the value of explanatory variable j for observation i , and β_j is the parameter determining the direction and degree of association of μ_i with explanatory variable x_{ij} . Equation (10.1) is a *linear prediction function* or *linear predictor* whereby the expected values of the Y_i components [$E(Y_i) = \mu_i$] are predicted by $\sum_j x_{ij}\beta_j$. The systematic part of the classical linear model assumes that all explanatory variables that influence the mean are included in the model and measured without error. In the random part of the classical linear model, each component Y_i is assumed to be independently distributed with a normal (Gaussian) probability distribution and constant variance $\sigma_i^2 = \sigma^2$ for all Y_i , $i = 1, 2, \dots, n$.

In generalized linear modeling, the linear predictor is allowed to predict a chosen function of the mean, $g(\mu_i)$, the variance σ_i^2 is allowed to vary as a function of the mean μ_i , and the random variable Y_i is allowed to have any distribution in the *exponential dispersion family*, a large family including the normal, Poisson, binomial, negative binomial, and multinomial distributions. The generalized linear model thus subsumes and generalizes the classical linear model. Analogous to the classical linear model, the systematic component of the generalized linear model has the form:

$$\eta_i = g(\mu_i) = \sum_j x_{ij}\beta_j \tag{10.2}$$

where:

$$\eta_i = g(\mu_i) \tag{10.3}$$

is called the *link function*, because it links the mean μ_i to the explanatory variables, and:

$$\eta_i = \sum_j x_{ij}\beta_j \tag{10.4}$$

is the linear predictor in the special case of the classical linear model, the link function is the identity:

$$\eta_i = \mu_i \tag{10.5}$$

and, as shown in Equation (10.1), the linear predictor is:

$$\mu_i = \sum_j x_{ij}\beta_j \tag{10.6}$$

By contrast, in a *log-linear model* (for count data), the link function is:

$$\eta_i = \log \mu_i \tag{10.7}$$

so the linear predictor is:

$$\log \mu_i = \sum_j x_{ij}\beta_j \tag{10.8}$$

If, as in the example presented later in this chapter, a log-linear model is applied to rate data, where each count y_i is divided by an exposure measure v_i , then the link function is:

$$\eta_i = \log \mu_i / v_i = \log \mu_i - \log v_i \tag{10.9}$$

and the linear predictor is:

$$\log \mu_i / v_i = \sum_j x_{ij}\beta_j \tag{10.10}$$

or, equivalently:

$$\log \mu_i = \sum_j x_{ij}\beta_j + \log v_i \tag{10.11}$$

where the additive term $\log v_i$ is called an *offset*. In generalized linear modeling of rate data, the offset is modeled as an additional explanatory variable (covariate) in the model with parameter $\beta = 1$ forced to unity.

The random component of the generalized linear model specifies the response variable Y with independent observations (y_1, y_2, \dots, y_n) drawn from a probability distribution in the exponential dispersion family, all members of which have the form:

$$f(y_i; \theta_i) = \exp\{[y_i \tilde{\theta}_i - b(\theta_i)] / a(\phi) + c(y_i, \phi)\} \quad (10.12)$$

where θ_i is the *natural parameter*, ϕ is the *dispersion parameter*, and $a(\phi)$, $b(\theta_i)$, and $c(y_i, \phi)$ are functions taking different forms for different members of the exponential family (for example, normal, Poisson, binomial, and so on).

The probability mass or density function for any member of the exponential dispersion family can be written in the form of Equation (10.12). In the case of the Poisson distribution, for example, we have:

$$f(y_i; \mu_i) = e^{-\mu_i} \mu_i^{y_i} / y_i! = \exp[y_i \log \mu_i - \mu_i - \log y_i!] = \exp[y_i \theta_i - \exp(\theta_i) - \log y_i!] \quad (10.13)$$

where $\theta_i = \log \mu_i$, $a(\phi) = 1$, $b(\theta_i) = \exp(\theta_i) = \mu_i$, and $c(y_i, \phi) = -\log y_i!$. Although other link functions might also be considered, the link function for which $g(\mu_i) = \theta_i$ is called the *canonical link*, whereby the natural parameter equals the linear predictor:

$$\theta_i = \sum_j x_{ij} \beta_j \quad (10.14)$$

As can be seen in Equations (10.13) and (10.14), $\log \mu_i$ is the canonical link function for a log-linear model and also usually makes the most sense in practical applications as it precludes negative predictions for count data.

Finally, the variance for any member of the exponential dispersion family is the product of the variance function $V(\mu_i)$ and the dispersion function $a(\phi)$, that is:

$$\text{Var}(Y_i) = V(\mu_i) a(\phi) \quad (10.15)$$

where the variance function:

$$V(\mu_i) = b''(\theta_i) \quad (10.16)$$

is the second derivative of the function $b(\theta_i)$ in Equation (10.12), and the dispersion function $a(\phi)$ commonly has the form:

$$a(\phi) = \phi / w_i \quad (10.17)$$

where the dispersion parameter ϕ is divided by a known prior weight w_p , which can vary for each observation Y_p , but is often unity, whereby the variance is $\text{Var}(Y_i) = V(\mu_i)\phi$. In the classical linear model with normal distribution and variance σ^2 , the variance function is $V(\mu_i) = 1$, and the dispersion parameter is $\phi = \sigma^2$. In the Poisson log-linear model, the variance function is $V(\mu_i) = \mu_i$, and the dispersion parameter is $\phi = 1$.

10.1.2 Parameter Estimation and Statistical Inference

In *maximum likelihood estimation* of the generalized linear model parameters $(\beta_0, \beta_1, \dots, \beta_{p-1})$, the likelihood of the sampled observations (y_1, y_2, \dots, y_n) is expressed as a function of those parameters, and estimates of $(\beta_0, \beta_1, \dots, \beta_{p-1})$ are found that maximize the likelihood, or rather the log likelihood, which yields the same estimates but is more mathematically tractable. The likelihood of an observation from the exponential dispersion family is:

$$f_i = \exp\{[y_i\theta_i - b(\theta_i)]/a(\phi) + c(y_i, \phi)\} \tag{10.18}$$

and the log likelihood of an observation from the exponential dispersion family is thus:

$$L_i = \log f_i = [y_i\theta_i - b(\theta_i)]/a(\phi) + c(y_i, \phi) \tag{10.19}$$

The likelihood of a sample of n independent observations (y_1, y_2, \dots, y_n) is the product of the n individual observation likelihoods f_p that is:

$$f_1 f_2 \dots f_n = \prod_n f_i = \prod_n \exp\{[y_i\theta_i - b(\theta_i)]/a(\phi) + c(y_i, \phi)\} \tag{10.20}$$

and the log likelihood of a joint sample of n independent observations (y_1, y_2, \dots, y_n) is thus the sum of the n individual observation log likelihoods L_p that is:

$$\begin{aligned} \log[f_1 f_2 \dots f_n] &= \log f_1 + \log f_2 + \dots + \log f_n \\ &= \sum_i L_i = \sum_i \{[y_i\theta_i - b(\theta_i)]/a(\phi) + c(y_i, \phi)\} \end{aligned} \tag{10.21}$$

In the case of the log-linear model, the focus of this chapter, the log likelihood for a sample of n independent observations is thus:

$$\sum_i L_i = \sum_i [y_i \log \mu_i - \mu_i - \log y_i!] \tag{10.22}$$

The latter term in Equation (10.21) expresses the sample log likelihood as a function of the known sample observations (y_1, y_2, \dots, y_n) and unknown parameters $(\beta_0, \beta_1, \dots, \beta_{p-1})$ that are estimated by maximizing $\sum_i L_i$ in maximum likelihood estimation. Estimates of $\beta_0, \beta_1, \dots, \beta_{p-1}$ are obtained as solutions of the likelihood equations:

$$\sum_i [(y_i - \mu_i)x_y / v(\mu_i)] (\partial \mu_i / \partial \eta_i) = 0 \tag{10.23}$$

for $i = 1, \dots, n$, and $j = 0, 1, \dots, p - 1$, where η_i is the link function, and $v(\mu_i) = \text{var}(Y_i)$ is the variance (expressed as a function of the mean μ_i). The parameters β_j are implicit in Equation (10.23), since the mean is the inverse of the link function, that is, $\mu_i = g^{-1}(\sum_j x_{ij}\beta_j)$. For example, the variance $v(\mu_i)$ is σ^2 for a classical linear model, μ_i for a Poisson log-linear model, and $\phi \mu_i$ for a log-linear model which assumes that the variance is proportional to the mean to handle overdispersion. For the Poisson log-linear model, $\partial \mu_i / \partial \eta_i = v(\mu_i) = \mu_i$, so the likelihood equations simplify to

$$\sum_i [(y_{.i} - \mu_i) \cdot x_{.j}] = 0 \quad (10.24)$$

for $i = 1, \dots, n$, and $j = 0, 1, \dots, p - 1$. Because the likelihood equations are nonlinear functions of the parameters $(\beta_0, \beta_1, \dots, \beta_{p-1})$ for most generalized linear models, maximum likelihood estimation requires iterative numerical methods. These methods, lucidly covered in McCullagh and Nelder (1989) and Agresti (2002) are beyond the scope of this chapter.

A generalized linear model with as many parameters as observations (that is, $p = n$) is called a *saturated model*. A saturated model perfectly fits the data, explaining 100 percent of the variance and yielding the highest possible maximum likelihood for the sample, but lacks scientific parsimony and other desirable properties such as a smooth curve fit. But the likelihood of the saturated model is a useful baseline for checking model fit. Let L_s denote the maximum log likelihood of a saturated model, and let L_a denote the maximum log likelihood of an alternative model with fewer parameters. For a Poisson or binomial model, the *scaled deviance* is twice the difference of the maximum log likelihoods of the saturated and alternative models, that is:

$$D^* = 2[L_s - L_a] = D / \phi \quad (10.25)$$

which expresses the deviance D as a multiple of the scale parameter. For Poisson or binomial models, $\phi = 1$, so $D^* = D$. Furthermore, for a Poisson or binomial modeling situation in which the number of sample observations n remains fixed regardless of counts, the deviance D divided by its degrees of freedom, $df = n - p$, has a chi-squared asymptotic distribution under the null hypothesis that the two models (saturated and alternative) fit the data equally well. Rejecting the null hypothesis indicates poor alternative model fit. On the other hand, failing to reject the null can indicate either good model fit or insufficient statistical power to detect a poor fit.

Analysis of deviance, a generalization of analysis of variance, is a powerful tool used to compare models and identify explanatory variables associated with variation in the criterion variable Y . Consider the comparison of two generalized linear models, M_0 and M_1 , where M_0 is a special case of M_1 . Then M_0 is said to be *nested* in M_1 . For example, M_1 might include one parameter for a covariate not included

in M_0 , so M_0 would be the special case of M_1 with that parameter forced to equal 0; or M_1 might include an interaction term excluded from M_0 ; or M_1 might include a set of terms, such as interactions, excluded from M_0 ; and so on. Let D_0 denote the deviance of M_0 , and let D_1 denote the deviance of M_1 . Also let p_0 be the number of parameters in M_0 , and let p_1 be the number of parameters in M_1 (with $p_1 > p_0$). The difference of deviances:

$$D_0 - D_1 = 2 \log[\exp(L_1) / \exp(L_0)] \quad (10.26)$$

divided by its degrees of freedom, $df = p_1 - p_0$, is a *likelihood-ratio* statistic with a chi-squared asymptotic distribution under the null hypothesis that models M_0 and M_1 fit the data $(y_1, y_2, \dots, y_p, \dots, y_n)$ equally well. Because a nested model M_0 , with fewer parameters than M_1 , can never fit better than M_1 , $D_0 \geq D_1$, so a likelihood-ratio statistic is nonnegative. The larger the likelihood-ratio statistic, the worse the fit of M_0 compared to M_1 , so rejecting the null hypothesis of no difference in fit indicates better fit of M_1 compared to M_0 .

Maximum likelihood estimation also provides asymptotic (large sample) parameter estimate variances. The standard errors (variance square roots) of estimates of the model parameters β_j , $j = 0, 1, \dots, p - 1$, may be of interest, for example, to construct confidence intervals. While the derivation of these estimates is beyond the scope of our discussion, McCullagh and Nelder (1989) and Agresti (2002) lucidly explain them.

In *maximum quasi-likelihood estimation*, an extension of generalized linear models, one need not assume a particular distribution for Y_j ; instead, one assumes a mean-variance relationship:

$$\sigma_i^2 = v(\mu_i) \quad (10.27)$$

and substitutes the appropriate term for $v(\mu_i)$ in Equation (10.23). For the Poisson distribution, for example, where $\sigma_i^2 = \mu_i$, the quasi-likelihood equations substitute μ_i for $v(\mu_i)$ in Equation (10.23). The equations solved to obtain maximum quasi-likelihood estimates are exactly the same as the likelihood equations used in maximum likelihood estimation, but the equations are not true likelihood equations unless the Y_i distribution is a member of the *natural exponential family*, which is the subset of the exponential dispersion family where the dispersion parameter ϕ is known, for example, the Poisson distribution where $\sigma_i^2 = \mu_i$. Nevertheless, Wedderburn (1974) proposed quasi-likelihood estimation as a further generalization of generalized linear models to handle even more diverse situations and suggested using the estimating equations in Equation (10.23) with any variance function regardless of whether the underlying probability distribution belongs to the natural exponential family.

One important application of quasi-likelihood to analysis of log-linear model rate data is to handle overdispersion, that is, variances that exceed the means, as

often occurs in practice with non-negative integer (count) data. Failure to correct for overdispersion increases the type I error rate (that is, true probability of erroneously rejecting the null hypothesis) for significance tests and erroneously reduces the width of confidence intervals. To handle this situation, the variance is assumed to be proportional to the mean, that is:

$$\sigma_i^2 = \phi \mu_i \quad (10.28)$$

where the dispersion or scale parameter ϕ is estimated and multiplied by the estimated mean to obtain the estimated variance corrected for overdispersion. The scale parameter ϕ can be estimated several ways, but a common method is based on the fact that the scaled deviance D/ϕ has a chi-square asymptotic distribution with $n - p$ degrees of freedom (and thus expectation $n - p$), so the deviance divided by the degrees of freedom is an estimate of the scale parameter, that is:

$$D/(n - p) \approx \phi \quad (10.29)$$

for large samples. Estimation for an overdispersed Poisson log-linear model proceeds by fitting the model by standard maximum likelihood methods, estimating the scale parameter ϕ using the full model deviance divided by its degrees of freedom, dividing log likelihoods used in likelihood ratio tests by the estimated ϕ , adjusting estimated parameter standard errors using the variance estimates multiplied by the estimated ϕ , and proceeding as usual with hypothesis tests and/or confidence intervals.

Finally, there are many well-known diagnostic procedures for testing the adequacy of a model, that is, plotting observed scores against predicted scores to identify potential outlier problems and plotting variances against means to assess the mean-variance assumption. See McCullagh and Nelder (1989) and Agresti (2002) for application of diagnostic methods in generalized linear modeling.

10.1.3 Illustrative Application

To evaluate the effectiveness of universal helmet laws, one approach is to compare motorcyclist fatalities in states with a universal helmet law to those in states without it, adjusting for differences in motorcycling activity between the states. Unfortunately, whereas the number of annual motorcycle registrations is available for individual states, the number of motorcycle miles traveled is not. Although the number of motorcycle registrations is related to exposure, it neglects variation in the activity of the registered motorcycles—a key quantitative measure needed to assess the association of fatality rates with helmet laws. Nevertheless, since motorcycling activity is highly seasonal, with more activity on warm or dry days than on cold or rainy days, and climates vary markedly across states in the United States, fatalities per registered motorcycle in the United States can be compared between

states with and without universal helmet laws while controlling for climate measures correlated with motorcyclist activity.

This study employed maximum quasi-likelihood generalized linear modeling to explore the association of motorcyclist fatality rates with universal helmet laws using climate measures to control for motorcyclist activity (Morris, 2006). The analytic objective was to maintain scientific parsimony and statistical power, with minimal reliance on stringent statistical assumptions, by modeling fatality rates as a function of one explanatory variable (universal helmet law) and two climate-related activity measures (heating degree days, precipitation) along with pertinent quadratic and interaction terms. Quasi-likelihood generalized linear modeling provided crucial flexibility in modeling the relationship between a function of the mean and the covariates, the relationship between the mean and variance, and the error distribution.

Motorcyclist fatality data were from the National Highway Traffic Safety Administration's (NHTSA) Fatality Analysis and Reporting System (FARS) (NHTSA, 2005). FARS is a database of information about the scenarios, vehicles, drivers, and passengers involved in all fatal motor vehicle crashes on public highways and roads in the United States. Data on hospital emergency room-treated injuries were from the U.S. Consumer Product Safety Commission's (CPSC) National Electronic Injury Surveillance System All-Injury Program (NEISS-AIP) (CPSC, 2001). Data on the number of registered motorcycles by states were from the Federal Highway Administration (FHWA, 2005).

Normalized state climate data, including population-weighted annual heating degree days and precipitation inches, were from the National Oceanic and Atmospheric Administration (NOAA). The heating degree days statistic is a measure of cold weather energy consumption and is defined as the annual sum of daily differences in mean daily temperature from a 65° base (with the difference set to 0 if the mean daily temperature exceeds the 65° base temperature), averaged across all stations within the state, with the average weighted by population distribution in the area. At one station in a given year, for example, five days with a mean daily temperature of 64° would result in five degree days, as would one day with a mean daily temperature of 60°. NOAA's normalized heating degree day measure, an annual average derived over the 30-year period 1971–2000, is a climate measure that estimates the annual average heating degree days for each state during the normalization period. The advantage of heating degree days over average temperature as a measure of motorcyclist activity consists both in its theoretical utility for ratio-scale measurement of the change in thermal energy necessary to maintain a comfortable ambient temperature and in its empirical utility in accounting for substantial nuisance variation in fatality rates.

To demonstrate the seasonality of motorcyclist fatalities and injuries, and their strong association with climate measures, Table 10.1 gives monthly motorcyclist

fatalities and injuries in the United States for 2001–2002 along with normalized heating degree days and precipitation inches for the coterminous United States. (The normalized climate measures by month were only available for the coterminous United States). Table 10.1 shows that the largest percentages of fatalities (11.1–13.5 percent) and injuries (10.7–13.1 percent) occurred during warm months (May–September) associated with the smallest percentages of normalized heating degree days (0.2–3.5 percent) and the largest percentages of precipitation inches (8.5–10.0 percent). Conversely, the smallest percentages of fatalities (2.6–3.6 percent) and injuries (3.2–3.7 percent) occurred during cold months (December–January) associated with the largest percentages of normalized heating degree days (16.2–20.3 percent) and the smallest percentages of precipitation inches (6.8–7.5 percent).

Table 10.2 confirms large statistically significant Pearson correlations among the monthly measures in Table 10.1, with a correlation of .98 for motorcyclist

Table 10.1 Monthly motorcyclist fatalities and U.S. hospital emergency room-treated injuries during 2001–2002 and normalized heating degree days and precipitation inches

Month	Motorcyclist fatalities, 2001–2002		U.S. Emergency room-treated motorcyclist injuries, 2001–2002		Normalized heating degree days (coterminous United States)		Normalized precipitation inches (coterminous United States)	
	Number	%	Number	%	Number	%	Number	%
1	170	2.6	8,098	3.6	917	20.3	2.27	7.5
2	222	3.4	8,370	3.7	732	16.2	2.04	6.8
3	332	5.2	11,652	5.2	593	13.1	2.59	8.6
4	549	8.5	21,868	9.7	345	7.6	2.44	8.1
5	713	11.1	23,938	10.7	159	3.5	3.01	10.0
6	850	13.2	28,476	12.7	39	0.9	2.92	9.7
7	842	13.1	25,888	11.5	9	0.2	2.79	9.2
8	870	13.5	29,364	13.1	15	0.3	2.65	8.8
9	796	12.4	25,348	11.3	77	1.7	2.58	8.5
10	504	7.8	19,296	8.6	282	6.2	2.29	7.6
11	359	5.6	14,900	6.6	539	11.9	2.40	7.9
12	234	3.6	7,188	3.2	817	18.1	2.23	7.4
Total	6,441	100.0	224,386	100.0	4,524	100.0	30.21	100.0

Sources: NHTSA, CPSC, NOAA, 2004.

Table 10.2 Correlation matrix for monthly motorcyclist fatalities and U.S. hospital emergency room-treated injuries during 2001–2002 and normalized heating degree days and precipitation inches

	U.S. emergency room-treated motorcyclist injuries	Normalized heating degree days	Normalized precipitation inches
Motorcyclist fatalities	0.983**	-0.983**	0.800*
U.S. emergency room-treated motorcyclist injuries		-0.979**	0.772*
Normalized heating degree days			-0.764*

* $p < .005$, ** $p < .0001$; 2-tail.

Source: Bureau of Transportation Statistics, 2005.

fatalities and injuries, $-.98$ for fatalities and heating degree days, $.80$ for fatalities and precipitation, $-.98$ for injuries and heating degree days, $.77$ for injuries and precipitation, and $-.76$ for heating degree days and precipitation.

Figure 10.1 gives fatalities per 10,000 registered motorcycles per year as a function of universal helmet law and annual heating degree days for all 50 states. Fatality rates are linearly associated with annual heating degree days in both universal helmet law ($R^2 = .284$) and non-universal helmet law ($R^2 = .519$) states, with essentially parallel least-squares regression lines relating fatality rates to heating degree days. Range restriction in the universal helmet law states ($n = 20$) is the most likely explanation of the smaller proportion of variance in fatality rates accounted for by heating degree days in those states, which exclude both Alaska and Hawaii. Although there is dispersion about the regression lines in both groups, reflecting other relevant factors influencing state motorcyclist fatality rates, a substantial portion of that dispersion is attributable to variation in annual precipitation as shown in Figure 10.2.

Figure 10.2 gives fatalities per 10,000 registered vehicles per year as a function of universal helmet law and annual precipitation inches for all 50 states. Figure 10.2 reveals quadratic association of fatality rates with annual precipitation. The linear and quadratic components of precipitation inches together account for about 18 percent of the variance in fatality rates among universal helmet law states and 35 percent in non-universal helmet law states. It is beyond the scope of this analysis to attempt an explanation of why the relation of state fatality rates with state average annual precipitation should take the J form in Figure 10.2; rather, the purpose is to control nuisance variation in state fatality rates (that is, mainly due to variation in activity) to permit a parsimonious and statistically powerful assessment of the association of fatality rates with universal helmet laws. The full linear model relat-

y with normalized
ous United States.
ble for the coter-
percentages of fatali-
rred during warm
ges of normalized
ages of precipita-
tages of fatalities
ring cold months
normalized heat-
es of precipitation
rrelations among
3 for motorcyclist

ergency
g degree days

Normalized precipitation inches (coterminous United States)	
Number	%
2.27	7.5
2.04	6.8
2.59	8.6
2.44	8.1
3.01	10.0
2.92	9.7
2.79	9.2
2.65	8.8
2.58	8.5
2.29	7.6
2.40	7.9
2.23	7.4
30.21	100.0

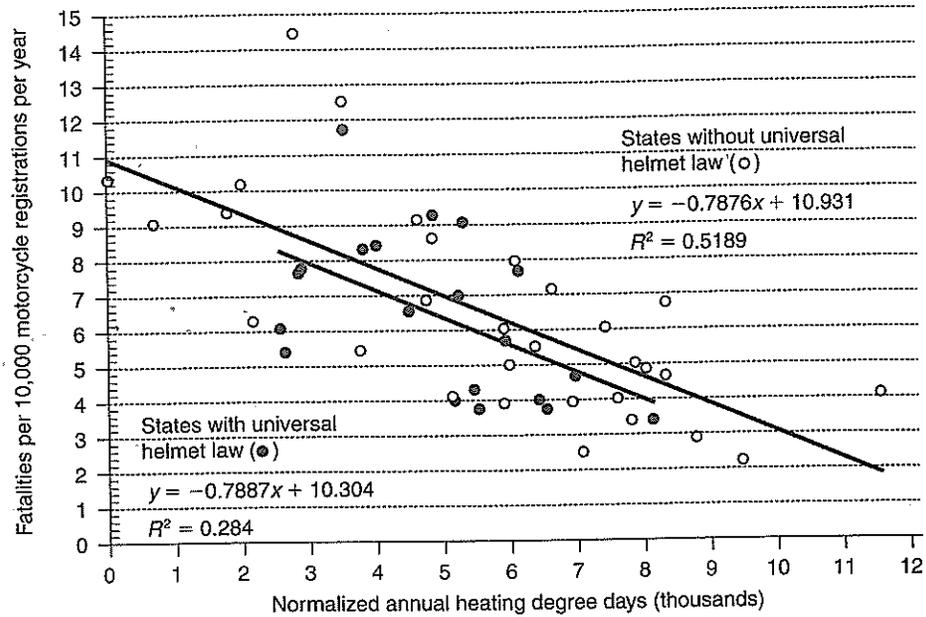


Figure 10.1 Motorcyclist fatalities per 10,000 registered motorcycles per year in states with or without a universal helmet law every year from 1993 through 2002 as a function of annual heating degree days. Source: NHTSA, FHWA, NOAA, 2004.

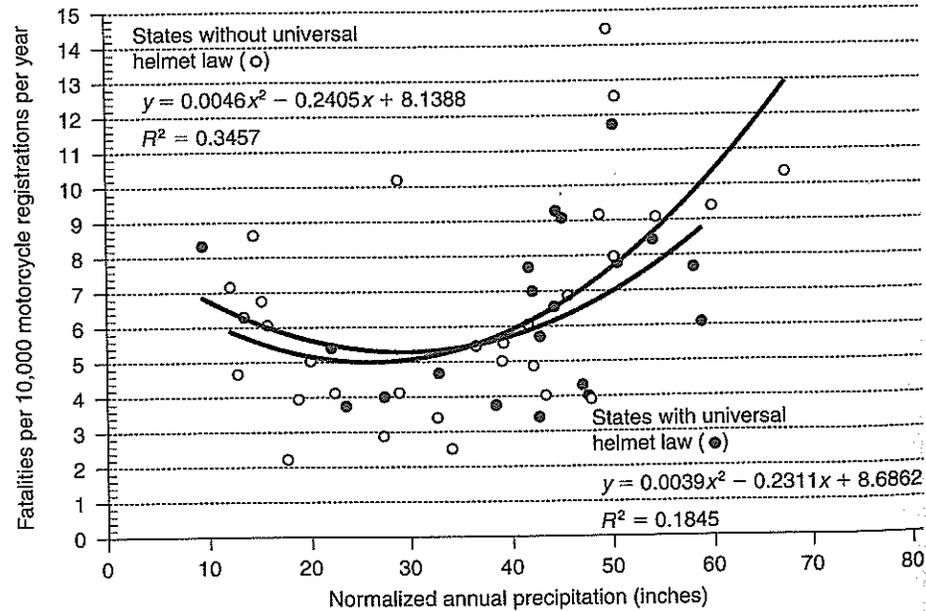


Figure 10.2 Motorcyclist fatalities per 10,000 registered motorcycles per year in states with or without a universal helmet law every year from 1993 through 2002 as a function of annual precipitation (inches). Source: NHTSA, FHWA, NOAA, 2004.

ing fatality rates to heating degree days, precipitation inches, squared precipitation inches (a quadratic term), the products of heating degree days with precipitation and squared precipitation inches (interaction terms), and the dichotomous universal helmet law indicator accounts for 55 percent of the variance in state fatality rates and 58 percent of the (natural) log-transformed fatality rates. Thus, normalized heating degree days and precipitation inches together account for substantial variation in state motorcyclist fatality rates.

Generalized linear regression analysis compared fatality rates in states with and without universal helmet laws adjusting for exposure as indexed by normalized state climate data. For each state, the annual average fatality rate was estimated by dividing the sum of fatalities across the decade from 1993 to 2002 by the sum of motorcycle registrations each year across the same period. The annual average heating degree days and precipitation were obtained likewise. To include all 50 states in the analysis, the five states that repealed a universal helmet law sometime during the 1993–2002 decade (3 in 1997, 1 in 1999, and 1 in 2000) were grouped with states that did not have a universal helmet law during that decade. Also, following NOAA, data for the District of Columbia were combined with those for Maryland, both of which had a universal helmet law throughout the 1993–2002 decade. The comparison is thus between 20 states in the United States that had a universal helmet law every year from 1993 through 2002 and 30 states that did not (that is, 25 states that did not have a universal helmet law and five states that did for some years, but not for the entire decade). Since helmet usage is known to correlate highly with universal helmet law requirements, this distinction is likely to correlate highly with actual helmet usage in the states, with higher total usage in states that had a universal helmet law in effect during the entire decade as compared to states that did not. An additional analysis was run in which the dichotomous universal helmet law grouping variable was replaced with a continuous variable measuring the number of years (rounded to the nearest 1/2 month) that a universal helmet law was actually in effect in each state from 1993 through 2002. The results of both analyses were similar.

To assess the association of fatality rates with helmet laws while controlling for annual heating degree days, precipitation inches, squared precipitation inches, and their interaction, quasi-likelihood generalized linear regression analyses were performed using the SAS (V8.0) GENMOD procedure and log-linear model:

$$\log(\mu_F / V) = \beta_0 + \beta_1 D + \beta_2 P + \beta_3 P^2 + \beta_4 DP + \beta_5 DP^2 + \beta_6 H \quad (10.30)$$

where $\log(\cdot)$ denotes the natural log function, μ_F = expected annual fatalities, V = annual motorcycle registrations (10,000s), D = annual heating degree days, P = annual precipitation (inches), and for comparison, either (a) $H = 0$ or 1 indicating whether the state had a universal helmet law every year from 1993 to 2002 or (b) $0 \leq H = \text{years helmet law in effect from 1993 to 2002} \leq 10$.

Whether the universal helmet law factor was regarded as dichotomous or continuous, each analysis included an intercept parameter β_0 and parameters for linear association of log fatality rate with D and P (β_1, β_2), quadratic association with P (β_3), interaction of D with the linear (β_4) and quadratic (β_5) components of P , and association with a universal helmet law (β_6). Model parameters were estimated via maximum likelihood assuming a Poisson distribution, with parameter estimate variances adjusted for overdispersion via quasi-likelihood generalized linear modeling methods using the square root of the deviance divided by the degrees of freedom to estimate the generalized linear model scale parameter.

Analyses employed the SAS (V8.0) GENMOD procedure with the form:

```
proc genmod; model F = D P P2 DP DP2 H/
      dist=poi link=log offset=LV scale=deviance type1 type3;
```

This SAS code specifies the model depicted in Equation (10.30), in which the variables are defined as previously stated, except that F denotes fatalities, $P2$ denotes P^2 , $DP2$ denotes DP^2 , and LV denotes the natural log of V . The modeling options specified after the slash (/) are defined as follows: `dist = poi` specifies the Poisson distribution; `link = log` specifies a log-linear model; `offset = LV` specifies the offset for this log-linear rate analysis; `scale = deviance` specifies that the scale parameter is to be estimated using the full model deviance divided by its degrees of freedom $n - p$; `type 1` specifies a set of hierarchical analyses using the scale parameter estimated by the deviance obtained for the model including only the parameters in the model at that point (for example, after D and P are in the model); and `type 3` specifies a simultaneous analysis using the scale parameter estimated by the deviance obtained for the full model depicted in Equation (10.30).

As shown in Table 10.3, all parameter estimates in the quasi-likelihood generalized linear model analysis differ significantly from zero, including the fatality rate reduction associated with the universal helmet law whether the latter was measured dichotomously [$F(1, 43) = 2.72, p = .053$, one-sided] or continuously [$F(1, 43) = 2.63, p = .055$, one-sided]. A one-sided test of the universal helmet law effect is justified by *a priori* expectation of a safety benefit from existing empirical and biophysical evidence. The overall fit of either generalized linear model with estimated scale parameter was excellent whether the universal helmet law was measured dichotomously [scaled $\chi^2(43) / 43 = 1.04$] or continuously [scaled $\chi^2(43) / 43 = 1.03$]. In conclusion, with climate measures statistically controlled, state universal helmet laws were associated with lower motorcyclist fatality rates. This finding is consistent with studies using a variety of methodologies that have also reported motorcycle helmet safety benefits (Norvell and Cummings, 2002; Sass and Zimmerman, 2000).

Table 10.3 Generalized linear regression results

Source	Universal helmet law factor					
	Dichotomous			Continuous		
	Estimate	SE	F	Estimate	SE	F
Intercept	-8.3587	0.6125		-8.4609	0.6056	
Heating degree days (D)	0.2783	0.1170	5.69*	0.2893	0.1163	6.20*
Precipitation (P)	0.0860	0.0364	5.97*	0.0939	0.0365	7.04*
P ²	-0.0011	0.0005	5.73*	-0.0012	0.0005	6.50*
DP	-0.0255	0.0071	13.27*	-0.0263	0.0072	13.97*
DP ²	0.0004	0.0001	14.99*	0.0004	0.0001	15.26*
Universal helmet law	-0.1284	0.0778	2.72**	-0.0145	0.0089	2.63**

* $p < .05$, ** $p = .106$, *** $p = .112$; 2-sided.

Note: Results are for likelihood-ratio tests with a full model log-likelihood of 4569.0251 and a scale parameter of 5.5666 estimated by the square root of the full model deviance divided by the degrees of freedom (i.e., $5.5666 = (\sqrt{1332.4549 / 43})$) and do not depend on the order of entry into the model.

Source: Bureau of Transportation Statistics, 2005.

10.2 SUMMARY

The results show that climate measures have considerable promise as indirect measures (proxies) of motorcycling activity. And, more to the point of this chapter, illustrate the utility of quasi-likelihood generalized linear regression modeling in the analysis of fatality risk data.

The views in this chapter are those of the author and do not necessarily represent the views of the Bureau of Transportation Statistics, the Research and Innovative Technology Administration, the U.S. Department of Transportation, or any other agency or staff.

REFERENCES

- Agresti, A. 2002. *Categorical Data Analysis*. Hoboken, New Jersey: Wiley.
- Diggle, P., P. Heagerty, K.-Y. Liung, and S. Zeger. 2002. *Analysis of Longitudinal Data*. New York: Oxford.
- Federal Highway Administration, U.S. Department of Transportation. 2004. Highway Statistics 1992–2002. Washington, DC. Annual vehicle-miles of travel and related data by highway category and vehicle type for each year.

- from <http://www.fhwa.dot.gov/policy/ohpi/hss/hsspubs.htm> (accessed March 3, 2004).
- McCullagh, P. and J. A. Nelder. 1989. *Generalized Linear Models*. New York: Chapman and Hall.
- McCulloch, C. E. and S. R. Searle. 2001. *Generalized, Linear, and Mixed Models*. New York: Wiley.
- Morris, C. C. 2006. Generalized linear regression analysis of association of universal helmet laws with motorcyclist fatality rates. *Accident Analysis and Prevention*, 38: 142-147.
- National Highway Traffic Safety Administration, U.S. Department of Transportation. 2004. Fatality Analysis Reporting System (FARS). See <http://www-nrd.nhtsa.dot.gov/departments/nrd-30/ncsa/FARS.html>. December 30, 2004.
- National Climatic Data Center, National Environmental Satellite, Data, and Information Service, National Oceanic and Atmospheric Administration. 2002. State, Regional, and National Monthly Heating Degree Days Weighted by Population (2000 Census) 1971-2000 (and previous normal periods), Historical Climatology Series No. 5-1. Asheville, NC. Downloaded from <http://www.ncdc.noaa.gov/oa/climate/normals/usnormals.html#Release>. December 30, 2004.
- National Climatic Data Center, National Environmental Satellite, Data, and Information Service, National Oceanic and Atmospheric Administration. 2002. State, Regional, and National Seasonal Temperature and Precipitation Weighted by Area 1971-2000 (and previous normals periods), Historical Climatology Series No. 4-3. Asheville, NC. Downloaded from <http://www.ncdc.noaa.gov/oa/climate/normals/usnormals.html#Release>.
- Norvell, D. C. and P. Cummings. 2002. Association of helmet use with death in motorcycle crashes: a matched-pair cohort study. *American Journal of Epidemiology*, 156(5): 483-487.
- SAS Institute. 2005. SAS version 8.0, online documentation at <http://v8doc.sas.com/sashtml/>.
- Sass, T. R. and P. R. Zimmerman. 2000. Motorcycle helmet laws and motorcyclist fatalities. *Journal of Regulatory Economics*, 18(3): 195-215.
- U.S. Consumer Product Safety Commission. 2001. NEISS All Injury Program: Sample Design and Implementation. Washington, DC.
- Wedderburn, R. W. M. 1974. Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika*, 61(3): 439-447.

L
V
F
C
F
-
R
D
N
Y
P

T
R
h
st
ch
th
w