6

Frequency and Severity Models

Edward W. Frees (University of Wisconsin - Madison)

Chapter Preview. Many insurance data sets feature information about how often claims arise, the frequency, in addition to the claim size, the severity. This chapter introduces tools for handling the joint distribution of frequency and severity. Frequency-severity modeling is important in insurance applications because of features of contracts, policyholder behavior, databases that insurers maintain, and regulatory requirements. Model selection depends on the data form. For some data, we observe the claim amount and think about a zero claim as meaning no claim during that period. For other data, we observe individual claim amounts. Model selection also depends upon the purpose of the inference; this chapter highlights the Tweedie generalized linear model as a desirable option. To emphasize practical applications, this chapter features a case study of Massachusetts automobile claims, using out-of-sample validation for model comparisons.

6.1 How Frequency Augments Severity Information

At a fundamental level, insurance companies accept premiums in exchange for promises to indemnify a policyholder upon the uncertain occurrence of an insured event. This indemnification is known as a *claim*. A positive amount, also known as the *severity*, of the claim, is a key financial expenditure for an insurer. One can also think about a zero claim as equivalent to the insured event not occurring. So, knowing only the claim amount summarizes the reimbursement to the policyholder. Ignoring expenses, an insurer that examines only amounts paid would be indifferent to two claims of 100 when compared to one claim of 200, even though the number of claims differ.

Nonetheless, it is common for insurers to study how often claims arise, known as the *frequency* of claims. Let us think about reasons why an insurance analyst should be concerned with models of frequency as well as severity.

Contractual. It is common for insurance contracts to impose deductibles and policy limits on a per occurrence basis. For example, if the policy has a deductible of 100 per occurrence, then two losses of 100 would result in a payout (or claim) of zero from the insurer whereas a single loss of 200 would result in a payout of 100. Models of total insured losses need to account for deductibles and policy limits for each insured event.

Behaviorial. Models of insurance losses implicitly or explicitly account for deci-

sions and behavior of people and firms that can affect losses; these decision-makers can include not only the policyholder but also the insurance adjuster, repair specialist, medical provider, and so forth. Behaviorial explanatory (rating) variables can have different effects on models of how often an event occurs in contrast to the size of the event.

For example, in homeowners insurance, consider a very careful policyholder who lives in an expensive neighborhood. We might look to characteristics of the homeowner as an indication of introduction of loss prevention measures, e.g., sprinklers, as determinants that suggest low frequency. In contrast, we might look to the overall income level of the geographic area where the house is located as a proxy for the level of repair costs in the event of an accident, suggesting high severity.

In healthcare, the decision to utilize healthcare by individuals is related primarily to personal characteristics whereas the cost per user may be more related to characteristics of the healthcare provider (such as the physician).

In automobile insurance, we might think of population density as positively correlated with the frequency of accidents and negatively associated with severity. For example, in a densely populated urban area, the traffic congestion is high, meaning that drivers are likely to have frequent, but relatively low-cost, accidents. This is in contrast to a more sparsely populated rural area where there is an opportunity to drive speedily. Less congestion may mean less frequent accidents but greater speeds mean greater severity.

Prior claims history is also used as a variable that provides information about a policyholder's risk appetite. Especially in personal lines, it is common to use an indicator of whether or not a claim has occurred in, for example, the last three years, rather than the claim amount. (Claim amounts are commonly used in commercial lines through credibility calculations). In many countries, automobile premiums are adjusted by a so-called "bonus-malus" system where prior claim frequency is used to dynamically adjust premiums.

Databases. Many insurers keep separate data files that suggest developing separate frequency and severity models. For example, insurers maintain a "policyholder" file that is established when a policy is written. This file records much underwriting information about the insured(s), such as age, gender and prior claims experience, policy information such as coverage, deductibles and limitations, as well as the insurance claims event. A separate file, often known as the "claims" file, records details of the claim against the insurer, including the amount. (There may also be a "payments" file that records the timing of the payments although we shall not deal with that here.) This recording process makes it natural for insurers to model the frequency and severity as separate processes.

Regulatory and Administrative. Insurance is a closely monitored industry sector. Regulators routinely require the reporting of claims numbers as well as amounts. This may be due to the fact that there can be alternative definitions of an "amount," e.g., paid versus incurred, and there is less potential error when reporting claim numbers.

At a broad level, it is clear that insurers need very different administrative systems for handling small, frequently occurring, reimbursable losses, e.g., prescription drugs, versus rare occurrence, high impact events, e.g., inland marine. Every insurance

claim means that the insurer incurs additional expenses suggesting the that claims frequency is an important determinant of expenses.

There are considerable differences of opinion concerning the importance of frequency models for allocated loss adjustment expenses (ALAE), costs that can be associated with a specific claim, e.g., legal defense fees and claims adjuster costs. According to Werner and Modlin (2010), it is common to assume that ALAE to vary by the amount of the claim rather than frequency.

6.2 Sampling and the Generalized Linear Model

6.2.1 Sampling

For a sampling basis, begin by thinking about the policyholder and claims databases that an insurer maintains. An insurer enters new contracts with insureds and administers claims continuously over time. For some purposes, it is helpful to consider a continuous-time stochastic process as a sampling model; this is the perspective of the loss reserving Chapter 18 and the survival modeling Chapter 19 (where we examine processes that influence policy retention). In contrast, this chapter focuses on collections of policies without an emphasis on the exact calendar start date of the policy. In particular, we leave issues of claim development to those chapters and only consider closed claims. Ratemaking and reinsurance are the primary purposes of this chapter rather than reserving and policy retention.

To establish notation, for each policy $\{i\}$, the potentially observable responses are:

- N_i − the number of claims (events),
- $y_{ij},\ j = 1, ..., N_i$ − the amount of each claim (loss), and
- $S_i = y_{i1} + \cdots + y_{iN_i}$, the aggregate claim amount.

By convention, the set $\{y_{ij}\}$ is empty when $N_i = 0$.

For a specific accounting period (such as a year), the sample of observable responses may consist of:

(i) S_i, so that only aggregate losses are available. For example, when examining losses for commercial insurance, it is common that only aggregate losses are available.
(ii) (N_i, S_i), so that the number and amount of aggregate losses are available.
(iii) $(N_i, y_{i1}, ..., y_{i,N_i})$, so that detailed information about each claim is available. For example, when examining personal automobile claims, losses for each claim are available. Let $\mathbf{y}_i = (y_{i1}, ..., y_{i,N_i})'$ be the vector of individual losses.

We can use ideas from conditional probability to decompose the distribution into frequency and severity components. To be specific, consider the third data form. Suppressing the $\{i\}$ subscript, we decompose the distribution of the dependent variables as:

$$\begin{aligned} \mathrm{f}(N, \mathbf{y}) &= \mathrm{f}(N) \times \mathrm{f}(\mathbf{y}|N) \\ \text{joint} &= \text{frequency} \times \text{conditional severity}, \end{aligned}$$

where $\mathrm{f}(N, \mathbf{y})$ denotes the joint distribution of $(N, \mathbf{y})$. This joint distribution equals the product of the two components:

- claims frequency: $\mathrm{f}(N)$ denotes the probability of having N claims; and

- conditional severity: $\mathrm{f}(\mathbf{y}|N)$ denotes the conditional density of the claim vector $\mathbf{y}$ given N.

The second data form follows similarly, replacing the vector of individual losses $\mathbf{y}$ with the aggregate loss S. We can even decompose the first data form by breaking off the zero event through the indicator notation $r_i = \mathrm{I}(S_i > 0)$ for the frequency component and conditioning on $r_i = 1$ for the severity component. We will examine this data form using "two-part models" in Section 6.3.1.

Through this decomposition, we do *not* require independence of the frequency and severity components as is traditional in the actuarial science literature. There are many ways to model dependence when considering the joint distribution $\mathrm{f}(N, \mathbf{y})$. For example, one may use a latent variable that affects both frequency N and loss amounts $\mathbf{y}$, thus inducing a positive association. Copulas are another tool used regularly by actuaries to model non-linear associations. The conditional probability framework is a natural method of allowing for potential dependencies and provides a good starting platform for empirical work.

6.2.2 Generalized Linear Model

A natural starting platform for empirical modeling of both frequency and severity components is the generalized linear model (GLM) introduced in Chapter 5. Indeed, one reason for the popularity of this modeling framework is that it has the flexibility to address both frequency and severity models.

Thinking of a generic dependent variable y_i (without regard to whether it represents frequency or severity), we focus on logarithmic links so that the mean function is $\mathrm{E}\ y_i = \exp(\mathbf{x}_i'\boldsymbol{\beta})$.

In some instances, the mean is known to vary proportionally with a variable that we label as E_i, for "exposure;" see the next subsection for more discussion of exposures. To incorporate exposures, one can always specify one of the explanatory variables to be $\ln E_i$ and restrict the corresponding regression coefficient to be 1; this term is known as an *offset*. With this convention, the link function is

$$\ln \mu_i = \ln E_i + \mathbf{x}_i'\boldsymbol{\beta}.$$

Example 6.1 (Relativities). In this example, we consider a small fictitious data set that appears in Werner and Modlin (2010). The data consists of loss and loss adjustment expenses (`LossLAE`), decomposed by three levels of an amount of insurance (`AOI`) and three territories (`Terr`). For each combination of `AOI` and `Terr`, we have available the number of policies issued, given as the exposure.

AOI	Terr	Exposure	LossLAE
Low	1	7	210.93
Medium	1	108	4458.05
High	1	179	10565.98
Low	2	130	6206.12
Medium	2	126	8239.95
High	2	129	12063.68
Low	3	143	8441.25
Medium	3	126	10188.70
High	3	40	4625.34

Source: Werner and Modlin, 2010

Our objective is to fit a generalized linear model (GLM) to the data using `LossLAE` as the dependent variable. We would like to understand the influence of the amount of insurance and territory on `LossLAE`.

We now specify two factors and estimate a generalized linear model using a gamma distribution with a logarithmic link function. In the `R` output that follows, the "`relevel`" command allow us to specify the reference level. For this example, a medium amount of insurance (`AOI = medium`) and the second territory (`Terr = 2`) are chosen as the reference levels. Logarithmic exposure is used as an offset variable so that cells (combinations of the two categorical variables) with larger number of exposures/policies will have larger expected losses.

Selected R Output

```
> Sampdata$AOI = relevel(Sampdata$AOI, ref = "Medium")
> Sampdata$Terr = factor(Sampdata$Terr)
> Sampdata$Terr = relevel(Sampdata$Terr, ref = "2")
> summary(glm(LossLAE ~ AOI + Terr, offset = log(Exposure), data = Sampdata,
+     family = Gamma(link = "log")))

Call:
glm(formula = LossLAE ~ AOI + Terr, family = Gamma(link = "log"),
    data = Sampdata, offset = log(Exposure))

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.180e+00  1.975e-06 2116446   <2e-16 ***
AOIHigh      3.577e-01  2.164e-06  165302   <2e-16 ***
AOILow      -3.147e-01  2.164e-06 -145448   <2e-16 ***
Terr1       -4.601e-01  2.164e-06 -212656   <2e-16 ***
Terr3        2.123e-01  2.164e-06   98109   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for Gamma family taken to be 7.022767e-12)

    Null deviance: 1.3528e+00  on 8  degrees of freedom
Residual deviance: 2.8091e-11  on 4  degrees of freedom
AIC: -47.141
```

Parameter estimates can be readily converted to relativities by exponentiation, as follows:

Variable	Parameter Estimate	Relativity (exponential parameter estimate)
Intercept	4.18	65.366
AOILow	-0.3147	0.730
AOIMedium	0	1
AoIHigh	0.3577	1.430
Terr1	-0.4601	0.631
Terr2	0	1
Terr3	0.2123	1.237

With the relativities and exposures, it straightforward to compute predictions. For example, for a high amount of insurance in territory 1, the exposure is 179, so the fitted value is $179 \times 65.366 \times 1.430 \times 0.631 = 10,558$. This is close to the actual value 10,565.98.

By comparing all actual to fitted values, or the null to the residual deviance, or examining the t-values or p-values, we see that we have done a pretty amazing job of fitting this data. In fact, these data are artificially constructed by Werner and Modlin to prove that various univariate methods of identifying relativities can do poorly. A multivariate method such as GLM is usually preferred in practice. Recall that the purpose of linear, as well as generalized linear, modeling is to simultaneously fit several factors to a set of data, not each in isolation of the others. As will be discussed in the following subsection, we should pay attention to the variability when introducing exposures. However, weighting for changing variability is not needed for this artificial example.

6.2.3 Exposures

As illustrated in the prior example, actuaries commonly use the idea of an "exposure" to calibrate the size of a potential loss. This subsection discusses exposure from a statistical perspective. To begin, an *exposure* is a variable that can be used to explain the distribution of losses; that is, it is a rating variable. It is typically the most important rating variable, so important that both premiums and losses are quoted on a "per exposure" basis. Here are some examples:

Typical Exposure Bases for Several Lines of Business

Line of Business	Exposure Basis
Personal Automobile	Earned Car Year
Homeowners	Earned House Year
Workers Compensation	Payroll
Commercial General Liability	Sales Revenue, Payroll, Square Footage, Number of Units
Commercial Business Property	Amount of Insurance Coverage
Physician's Professional Liability	Number of Physician Years
Professional Liability	Number of Professionals (e.g., Lawyers or Accountants)
Personal Articles Floater	Value of Item

Source: Werner and Modlin, 2010

Naturally, selection of a good exposure base goes beyond statistics. An exposure basis should:

- be an accurate measure of the quantitative exposure to loss,
- be easy for the insurer to determine (at the time the policy is calculated) and not subject to manipulation by the insured,
- be easy to understand by the insured and to calculate by the insurer,
- consider any preexisting exposure base established within the industry, and
- for some lines of business, be proportional to inflation. In this way, rates are not sensitive to the changing value of money over time as these changes are captured in exposure base.

To illustrate, consider personal automobile coverage. Instead of the exposure basis "earned car year," a more accurate measure of the quantitative exposure to loss might be number of miles driven. However, this measure is difficult to determine at the time the policy is issued and subject to potential manipulation by the insured.

For frequency and severity modeling, it is customary to think about the frequency aspect as proportional to exposure and the severity aspect in terms of loss per claim (not dependent upon exposure). However, this does not cover the entire story. For many lines of business, it is convenient for exposures to be proportional to inflation. Inflation is typically viewed as unrelated to frequency but proportional to severity.

Small Exposures

We begin by considering instances where the units of exposure may be fractions. To illustrate, for our automobile data, E_i will represent the fraction of the year that a policyholder had insurance coverage. The logic behind this is that the expected number of accidents is directly proportional to the length of coverage. This can also be motivated by a probabilistic framework based on collections of Poisson distributed random variables known as *Poisson processes,* see, for example, Klugman et al. (2008).

For binary outcomes, this situation is less clear. One way to handle exposures is to let the logit link function depend on exposures. To this end, the basic logit link function is $\pi(z) = \frac{e^z}{1+e^z}$. Define an exposure weighted logit link function to be $\pi_i(z) = E_i \frac{e^z}{1+e^z}$. With this definition, the probability of a claim is

$$\Pr(r_i = 1) = \pi_i = \pi_i(\mathbf{x}_i'\boldsymbol{\beta}) = E_i \frac{\exp(\mathbf{x}_i'\boldsymbol{\beta})}{1 + \exp(\mathbf{x}_i'\boldsymbol{\beta})}. \tag{6.1}$$

For more discussion, see de Jong and Heller (2008, page 102. In particular, page 162 gives illustrative `SAS` code) de Jong and Heller (2008).

Variations. There are alternative ways of incorporating partial year exposures, none clearly superior to the others. Equation (6.1) is based on a uniform distribution of failures within a year. Some others include:

- A constant hazard rate within the year assumption, resulting in: $\pi_{i,H}(z) = 1 - (1 - \frac{e^z}{1+e^z})^{E_i}$.
- A hyperbolic assumption (known as the "Balducci" assumption for an Italian actuary), resulting in: $\pi_{i,B}(z) = \frac{E_i \frac{e^z}{1+e^z}}{1-(1-E_i)\frac{e^z}{1+e^z}}$.

See Bowers et al. (1997) for a discussion of these variations.

For some applications, the event of a claim is a relatively infrequent event and the analyst would like to use all the information available in a claims database. One may wish to "over-sample" policyholders with claims; the idea is to draw a larger proportion of a subset of the population that is of interest in the study. Appendix Section 6.6.2 provides details of this type of sampling scheme.

6.2.4 Grouped versus Individual Data

A discussion of large exposures leads naturally into a discussion of grouped versus individual data.

Using Offsets to Handle Large Exposures

To begin, recall that sums of independent Poisson random variables also have a Poisson distribution. So, when summing random variables from independent policies, it is sensible to think of exposures as large positive numbers. Thus, it is common to model the number of accidents per thousand vehicles or the number of homicides per million population.

For a Poisson distribution, we can use the (logarithmic) number of policies in a group as an offset variable. Mathematically, if we are thinking about E_i independent Poisson variables in group i, each with mean μ_i, then the sum will also be Poisson distributed with mean $E_i\mu_i$. For the Poisson distribution, the variance equals the mean, so both the mean and the variance grow proportionally to the exposure E_i. When using a Poisson distribution with a logarithmic link function, one only needs to specify an offset variable $\ln E_i$ to automatically account for the growing variability.

However, for other distributions, this need not be the case. In the GLM linear exponential family, we saw that the variance can be expressed as a function of the mean, $v(\mu)$. To be specific, consider the gamma distribution where $v(\mu) = \mu^2/\phi$ and ϕ is a dispersion parameter. If we are thinking about E_i independent gamma random variables in group i, each with mean μ and variance θ, then the sum will also be gamma distributed with mean $E_i\mu$ and variance $E_i\theta$. When using a gamma distribution with a logarithmic link function and offset variable $\ln E_i$, the mean will grow proportionally to the exposure E_i but the variability will grow proportionally to E_i^2, not E_i. So, an offset by itself can not handle large exposures.

Using Variable Scale Parameters to Handle Exposures

For a general distribution in the linear exponential family, suppose that we have m independent variables from the same distribution with location parameter θ and scale parameter ϕ. Then, basic arguments given in Section 6.6.1 show that the sample average comes from the same distributional family with location parameter θ and scale parameter ϕ/m. To apply this result, let us consider a problem that analysts regularly face, the use of grouped versus individual data.

To be specific, think about a sample $i = 1, \ldots, n$ categories, or *groups*. For example, each group could be formed by the intersection of amount of insurance, territory, and so forth. For the ith group, we have E_i independent observations with the same distribution from the linear exponential family. This has, for example, location parameter θ_i, mean μ_i, and scale parameter ϕ (that may or may not depend

on i). One could run an analysis with a data set based on individual observations $j = 1, \ldots, E_i$, $i = 1, \ldots, n$.

However, with these assumptions, then the average outcome from the ith group comes from the same exponential family with the same mean μ_i (or location parameter θ_i) but with a scale parameter ϕ/E_i. An alternative method of analysis would be to use the smaller grouped data sample consisting of only n observations, using the reciprocal of exposure as the weight. The Section 6.6.1 result guarantees:

- Estimates of location/mean parameters (e.g., the regression coefficients) will be the same. For ratemaking purposes, analysts typically focus on location parameter estimates.
- Only in the case when the scale parameter is known (e.g., binomial, Poisson) would other inferential aspects (standard errors, t-statistics, p-values, and so forth) be the same. Individual data analysis provides more accurate estimates of scale parameters than the corresponding grouped data analysis.

The book website provides a demonstration of this comparison using the statistical package `R`.

Large Exposures for Frequency and Severity Models

As noted earlier, for frequency and severity modeling, it is customary to think about the frequency aspect as proportional to exposure and the severity aspect in terms of loss per claim. Let us make this advice a bit more specific in the context of an individual versus grouped analyses.

Suppose that individual data consists of a sample $i = 1, \ldots, n$ groups, with $j = 1, \ldots, E_i$ independent observations within each group i. For observation $\{ij\}$, the dependent variables consist of (N_{ij}, S_{ij}), the frequency and total amount of claims.

If explanatory/rating variables are available at the individual observation level, then aggregating information up to the group level is problematic because one loses the information in individual level variables.

Instead, assume that explanatory/rating variables are available only at the group level and we wish to model aggregate frequency and severity variables $\{N_i = \sum_{j=1}^{n} N_{ij}$, $S_i = \sum_{j=1}^{n} S_{ij}\}$.

- For claims frequency, one alternative is to use a Poisson model with the response N_i and offset $\ln E_i$.
- For claims frequency, another alternative is to use a count member from the exponential distribution family, e.g., binomial, with the response N_i/E_i and scale parameter ϕ/E_i.
- For claims severity, use a severity member from the exponential distribution family, e.g., gamma, with the response S_i/N_i and scale parameter ϕ/N_i.

As noted earlier, these modeling strategies provide reliable estimates of location (mean) parameters but not scale parameters. This is a comparative advantage of analysis with individual level analysis; other advantages include:

- Group level analysis was important before modern day computing and databases became available. However, in modern times, rarely do the computing requirements for an individual level analysis present a substantial barrier.

- Group level analysis precludes the examination of individual observations. Often, a highly unusual observation ("outlier") can provide important information to the analyst.
- The equivalence between the two procedures relies on a number of unverifiable assumptions, including the independence of observations within a group. In some instances, we can think of reasons for positive associations among observations from the same category. In this case, the variance of the sum grows faster than linearly and so specifying the scale parameter as inversely proportional to the exposure may give too large a weight to categories with large exposure.

6.3 Frequency-Severity Models

6.3.1 Two-Part Models

In Section 6.2.1, we introduced three forms of dependent variables. We now focus on the first type where the only dependent variable of interest is the total claims from a policy. However, let us think about data sets where we have a large proportion of zeros, corresponding to no claims. For example, in homeowners, it is not uncommon to consider data where 93% of policies do not have a claim.

To address this large proportion of zeros, we consider a two-part model that is a special type of frequency-severity model. In a two-part model, one part indicates whether an event (claim) occurs and the second part indicates the size of the event. Specifically, let r_i be a binary variable indicating whether or not the ith subject has an insurance claim and y_i describe the amount of the claim.

To estimate a two-part model, the analyst first considers the frequency and then the severity, conditional on the frequency.

(i) Use a binary regression model with r_i as the dependent variable and $\mathbf{x}_{1i}$ as the set of explanatory variables. Denote the corresponding set of regression coefficients as $\boldsymbol{\beta}_1$. The logit is a typical binary regression model.

(ii) Conditional on $r_i = 1$, specify a regression model with y_i as the dependent variable and $\mathbf{x}_{2i}$ as the set of explanatory variables. Denote the corresponding set of regression coefficients as $\boldsymbol{\beta}_2$. The gamma with a logarithmic link is a typical severity model.

There is usually overlap in the sets of explanatory variables, where variables are members of both $\mathbf{x}_1$ and $\mathbf{x}_2$. Typically, one assumes that $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are not related so that the joint likelihood of the data can be separated into two components and run separately.

Tobit Models

Another way of modeling a large proportion of zeros is to assume that the dependent variable is (left) censored at zero. Chapter 19 on survival models provides a more complete introduction to censored regression. This section emphasizes the application to two-part data.

With censored regression models, we use an unobserved, or latent, variable y^* that is assumed to follow a linear regression model of the form

$$y_i^* = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i. \tag{6.2}$$

The responses are censored or "limited" in the sense that we observe $y_i = \max(y_i^*, 0)$. Estimation of this model is typically done by assuming normally distributed disturbances ε_i and using maximum likelihood estimation.

It is straightforward to extend this model to allow for limiting values that vary by policy. In actuarial applications, we think about d_i as representing a (known) deductible that varies by policyholder.

One drawback of the tobit model is its reliance on the normality assumption of the latent response. A second, and more important, drawback is that a single latent variable dictates both the magnitude of the response as well as the censoring. There are many instances where the limiting amount represents a choice or activity that is separate from the magnitude. For example, in a population of smokers, zero cigarettes consumed during a week may simply represent a lower bound (or limit) and may be influenced by available time and money. However, in a general population, zero cigarettes consumed during a week can indicate that a person is a non-smoker, a choice that could be influenced by other lifestyle decisions (where time and money may or may not be relevant).

6.3.2 Other Frequency-Severity Models

We now focus on the second and third types of dependent variables introduced in Section 6.2.1.

For the second form, we have aggregate counts and severities (N_i, S_i) (or use the notation y_i instead of S_i). Then, the two-step frequency-severity model procedure is:

- **1.** Use a count regression model with N_i as the dependent variable and $\mathbf{x}_{1i}$ as the set of explanatory variables. Denote the corresponding set of regression coefficients as $\boldsymbol{\beta}_1$. Typical models include the Poisson and negative binomial models.
- **2.** Conditional on $N_i > 0$, use a GLM with S_i/N_i as the dependent variable and $\mathbf{x}_{2i}$ as the set of explanatory variables. Denote the corresponding set of regression coefficients as $\boldsymbol{\beta}_2$. Typical models include the gamma regression with a logarithmic link and a dispersion parameter proportional to $1/N_i$.

For the third form of dependent variables, we have individual claims $\mathbf{y}_i = (y_{i1}, ..., y_{i,N_i})'$ available. In this case, the first step for the count model is the same. The second step for severity modeling becomes:

- **2*.** Conditional on $N_i > 0$, use a regression model with y_{ij} as the dependent variable and $\mathbf{x}_{2i}$ as the set of explanatory variables. Denote the corresponding set of regression coefficients as $\boldsymbol{\beta}_2$. Typical models include the linear regression (with logarithmic claims as the dependent variable), gamma regression and mixed linear models. For the mixed linear models, one uses a subject-specific intercept to account for the heterogeneity among policyholders.

6.3.3 Tweedie GLMs

The natural exponential family includes continuous distributions, such as the normal and gamma, as well as discrete distributions, such as the binomial and Poisson.

It also includes distributions that are mixtures of discrete and continuous components. In insurance claims modeling, a widely used mixture is the Tweedie (1984) distribution. It has a positive mass at zero representing no claims and a continuous component for positive values representing the total amount for one or more claims.

The Tweedie distribution is defined as a Poisson sum of gamma random variables. Specifically, suppose that N has a Poisson distribution with mean λ, representing the number of claims. Let y_j be an i.i.d. sequence, independent of N, with each y_j having a gamma distribution with parameters α and γ, representing the amount of a claim. Then, $S_N = y_1 + \ldots + y_N$ is a Poisson sum of gammas.

To understand the mixture aspect of the Tweedie distribution, first note that it is straightforward to compute the probability of zero claims as

$$\Pr(S_N = 0) = \Pr(N = 0) = e^{-\lambda}.$$

The distribution function can be computed using conditional expectations,

$$\Pr(S_N \le y) = e^{-\lambda} + \sum_{n=1}^{\infty} \Pr(N = n) \Pr(S_n \le y), \qquad y \ge 0.$$

Because the sum of i.i.d. gammas is a gamma, S_n (not S_N) has a gamma distribution with parameters $n\alpha$ and γ. Thus, for $y > 0$, the density of the Tweedie distribution is

$$\mathrm{f}_S(y) = \sum_{n=1}^{\infty} e^{-\lambda} \frac{\lambda^n}{n!} \frac{\gamma^{n\alpha}}{\Gamma(n\alpha)} y^{n\alpha-1} e^{-y\gamma}. \tag{6.3}$$

At first glance, this density does not appear to be a member of the linear exponential family. To see the relationship, we first calculate the moments using iterated expectations as

$$\mathrm{E}\ S_N = \lambda \frac{\alpha}{\gamma} \qquad \text{and} \qquad \mathrm{Var}\ S_N = \frac{\lambda\alpha}{\gamma^2}(1+\alpha).$$

Now, define three parameters μ, ϕ, p through the relations

$$\lambda = \frac{\mu^{2-p}}{\phi(2-p)}, \qquad \alpha = \frac{2-p}{p-1} \qquad \text{and} \qquad \frac{1}{\gamma} = \phi(p-1)\mu^{p-1}. \tag{6.4}$$

Inserting these new parameters in equation (6.3) yields

$$\mathrm{f}_S(y) = \exp\left[\frac{-1}{\phi}\left(\frac{\mu^{2-p}}{2-p} + \frac{y}{(p-1)\mu^{p-1}}\right) + S(y,p,\phi)\right],$$

where

$$\exp S(y,p,\phi) = \frac{1}{y} \sum_{n=1}^{\infty} \frac{\left(\frac{y^\alpha}{\phi^{1/(p-1)}(2-p)(p-1)^\alpha}\right)^n}{n!\Gamma(n\alpha)}.$$

Thus, the Tweedie distribution is a member of the linear exponential family. Easy calculations show that

$$\mathrm{E}\ S_N = \mu \qquad \text{and} \qquad \mathrm{Var}\ S_N = \phi\mu^p, \tag{6.5}$$

where $1 < p < 2$. The Tweedie distribution can also be viewed as a choice that is intermediate between the Poisson and the gamma distributions.

For the Tweedie glm, we might use $\mathbf{x}_{i,T}$ as a set of covariates and $\boldsymbol{\beta}_T$ as the corresponding set of regression coefficients. With a logarithmic link, $\mu_i = \exp\left(\mathbf{x}_{i,T}'\boldsymbol{\beta}_T\right)$. For the distribution function, there is no closed form expression but we could compute this directly, for example, using the `R` function `ptweedie`.

6.3.4 Comparing the Tweedie to a Frequency-Severity Model

As an alternative, consider a model composed of frequency and severity components. Then, we might use a Poisson regression model for the frequency, thinking of the number of claims for the ith person as:

$$N_i \sim Poisson(\lambda_i), \qquad \lambda_i = \exp(\mathbf{x}_{i,F}'\boldsymbol{\beta}_F),$$

using a logarithmic link function. Here, $\mathbf{x}_{i,F}$ is a set of covariates to be used in the frequency modeling and $\boldsymbol{\beta}_F$ is the corresponding set of regression coefficients.

For the severity, we might use a gamma regression also with a logarithmic link function. Thus, we would model loss amounts as

$$y_{ij} \sim gamma(\alpha, \gamma_i), \qquad \text{where } \frac{\alpha}{\gamma_i} = \mathrm{E}\ y_{ij} = \exp(\mathbf{x}_{i,S}'\boldsymbol{\beta}_S),$$

for $j = 1, \ldots, N_i$. Similar to frequency, $\mathbf{x}_{i,S}$ is a set of covariates to be used in the severity modeling and $\boldsymbol{\beta}_S$ is the corresponding set of regression coefficients. Thus, the frequency and severity models need not employ the same set of covariates.

Putting the frequency and severity components together yields the aggregate loss

$$S_{N,i} = y_{ij} + \cdots + y_{i,N_i}.$$

This has mean

$$\mathrm{E}\ S_{N,i} = \mathrm{E}\ N_i \times \mathrm{E}\ y_{ij} = \exp\left(\mathbf{x}_{i,F}'\boldsymbol{\beta}_F + \mathbf{x}_{i,S}'\boldsymbol{\beta}_S\right) \tag{6.6}$$

and variance

$$\mathrm{Var}\ S_{N,i} = \lambda_i \frac{\alpha}{\gamma_i^2}(1+\alpha) = \exp\left(\mathbf{x}_{i,F}'\boldsymbol{\beta}_F + 2\mathbf{x}_{i,S}'\boldsymbol{\beta}_S + \ln(1+1/\alpha)\right). \tag{6.7}$$

Note that for frequency-severity modeling, two parameters, λ_i and γ_i, vary with i. To compute the distribution function, one could use the Tweedie for S_N with the `R` function `ptweedie`. This would be done by reversing the relations in (6.4) to get

$$p = \frac{\alpha+2}{\alpha+1}, \qquad \mu_i = \lambda_i \frac{\alpha}{\gamma_i}, \qquad \text{and} \quad \phi_i \mu_i^p = \lambda_i \frac{\alpha}{\gamma_i^2}(1+\alpha). \tag{6.8}$$

Note that if one begins with the frequency-severity model formulation, the scale parameter ϕ depends on i.

6.4 Application: Massachusetts Automobile Claims

We investigate frequency-severity modeling using an insurance automobile claims dataset studied in Ferreira and Minikel (2010), Ferreira and Minikel (2012). These data, made public by the Massachusetts Executive Office of Energy and Environmental Affairs (EOEEA), summarizes automobile insurance experience from the state

of Massachusetts in year 2006. The dataset consists of approximately 3.25 million policies representing over half a billion dollars of claims.

Because the dataset represents experience from several insurance carriers, it is not surprising that the amount of policyholder information is less than typically used by large carriers that employ advanced analytic techniques. Nonetheless, we do have basic ratemaking information that is common to all carriers, including primary driver characteristics and territory groupings. At the vehicle level, we also have mileage driven in a year, the focus of the Ferreira and Minikel study.

6.4.1 Data and Summary Statistics

From the Ferreira and Minikel (2010) data, we drew a random sample of 100,000 policyholders for our analysis. Table 6.1 shows the distribution of number of policies by rating group and territory. The distribution of policies is reasonably level across territories. In contrast, the distribution by rating group is more uneven; for example, over three quarters of the policies are from the "Adult" group. The sparsest cell is business drivers in territory 6; the most heavily populated cell is territory 4 adult drivers.

Table 6.1. *Number of Policies by Rating Group and Territory*

	Territory						
Rating Group	1	2	3	4	5	6	Total
A – Adult	13,905	14,603	8,600	15,609	14,722	9,177	76,616
B – Business	293	268	153	276	183	96	1,269
I – Youthful with less than 3 years Experience	706	685	415	627	549	471	3,453
M – Youthful with 3-6 years Experience	700	700	433	830	814	713	4,190
S – Senior Citizens	2,806	3,104	1,644	2,958	2,653	1,307	14,472
Totals	18,410	19,360	11,245	20,300	18,921	11,764	100,000

For this study, an insurance claim is from only bodily injury, property damage liability, and personal injury protection coverages. These are the compulsory, and thus fairly uniform, types of insurance coverages in Massachusetts; it is critical to have uniformity in reporting standards in an intercompany study such as in Ferreira and Minikel (2010)Ferreira and Minikel (2012). As a result, in Table 6.2, the averages of the loss might appear to be lower than in other studies. This is because the "total" is over the three compulsory coverages and does not represent, for example, losses from the commonly available (and costly) comprehensive coverage. The average total loss in Table 6.2 is 127.48. We also see important differences by rating group, where average losses for inexperienced youthful drivers are over 3 times greater than adult drivers. We can think of this total loss as a "pure premium."

Table 6.2 shows that the average claim frequency is 4.3%. Specifically, for the 100,000 policies, there were 95,875 that had zero claims, 3,942 that had one claim, 176 that had two claims, and 7 that had three claims. The table also reports important difference by rating group, where the average number of losses for inexperienced youthful drivers are about 2.5 times greater than adult drivers.

Table 6.2 also summarizes information on the earned exposure, defined here as the amount of time that the policy was in force in the study, and annual mileage.

Annual mileage was estimated by Ferreira and Minikel (2010)Ferreira and Minikel (2012) based on Massachusetts' Commonwealth's Registry of Motor Vehicles mandatory annual safety checks, combined with automobile information from the vehicle identification number (commonly known using the acronym `VIN`). Interestingly, Table 6.2 shows that only about 90% of our data possess valid information about the number of miles driven, so that about 10% are missing.

Table 6.2. *Averages by Rating Group*

Rating Group	Total Loss	Claim Number	Earned Exposure	Annual Mileage	Number of Policies Total	Number of Policies with Valid Annual Miles
A	115.95	0.040	0.871	12,527	76,616	69,201
B	159.67	0.055	0.894	14,406	1,269	1,149
I	354.68	0.099	0.764	12,770	3,453	2,786
M	187.27	0.065	0.800	13,478	4,190	3,474
S	114.14	0.038	0.914	7,611	14,472	13,521
Total	127.48	0.043	0.870	11,858	100,000	90,131

Table 6.3 provides similar information but by territory. Here, we see that the average total loss and number of claims for territory 6 is about twice that for territory 1.

Table 6.3. *Averages by Territory*

Territory	Total Loss	Claim Number	Earned Exposure	Annual Mileage	Number of Policies Total	Number of Policies with Valid Annual Miles
1	98.24	0.032	0.882	12,489	18,410	16,903
2	94.02	0.036	0.876	12,324	19,360	17,635
3	112.21	0.037	0.870	12,400	11,245	10,092
4	126.70	0.044	0.875	11,962	20,300	18,331
5	155.62	0.051	0.866	10,956	18,921	16,944
6	198.95	0.066	0.842	10,783	11,764	10,226
Total	127.48	0.043	0.870	11,858	100,000	90,131

There are 4,125 (= 100,000 - 95,875) policies with losses. To get a better handle on claim sizes, Figure 6.1 provides smooth histograms of the loss distribution. The left-hand panel is in the original (dollar) units, indicating a distribution that is right-skewed. The right-hand panel shows the same distribution on a logarithmic scale where we see a more symmetric behavior.

Do our rating factors affect claim size? To get some insights into this question, Figure 6.2 shows the logarithmic loss distribution by each factor. The left-hand panel shows the distribution by rating group, the right-hand panel shows the distribution by territory. Neither figure suggests that the rating factors have a strong influence on the size distribution.

6.4.2 Model Fitting

We report three types of fitted models here: (1) frequency models, (2) a severity model, and (3) a pure premium model.

Table 6.4 summarizes the results from two frequency models, Poisson and negative binomial regression models. For both models, we used a logarithmic link with

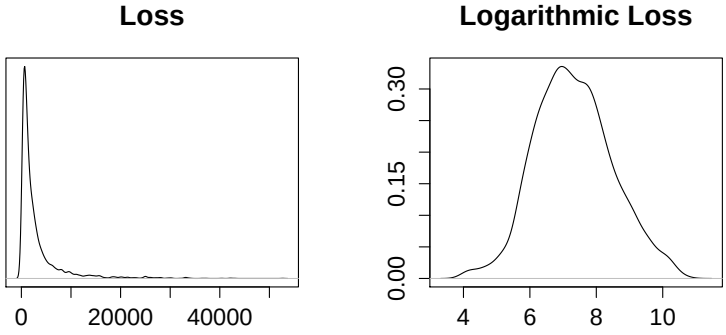

Fig. 6.1. Loss Distribution. The left-hand panel shows the distribution of loss, the right-hand panel shows the same distribution but on a (natural) logarithmic scale.

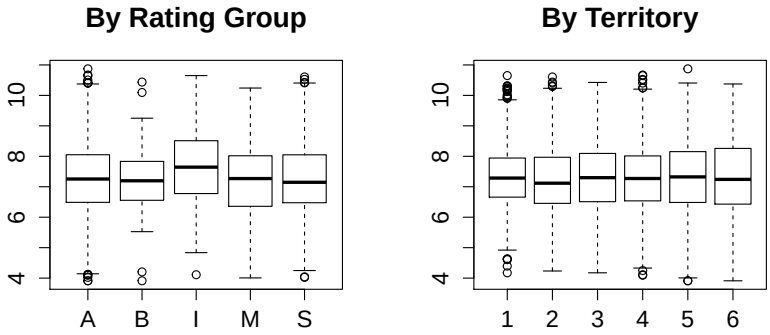

Fig. 6.2. Logarithmic Loss Distribution by Factor. The left-hand panel shows the distribution by rating group, the right-hand panel shows the distribution by territory.

logarithmic exposure as an offset variable. Focussing on the Poisson fit, we see that the t-statistics indicate strong statistical significance for several levels of each factor, rating group and territory. Additional tests confirm that they are statistically significant factors. Although not reported in Table 6.4, we also ran a model that included interactions among terms. The interaction terms were statistically insignificant with a $p - value = 0.303$ level. Hence, we report on the model without interactions, in part because of our desire for simplicity.

We also ran an analysis including annual mileage. This variable turned out to be strongly statistically significant with a t-statistic equal to 12.08. However, by including this variable, we also lost 9,869 observations due to missing values in annual mileage. From the perspective taken in the Ferreira and Minikel (2010, 2012) study, mileage is the key variable of interest and so the analyst would wish to retain this variable. From another perspective, an analyst might be concerned that including the mileage variable results in analyzing a biased sample; that is, the roughly 10% population without mileage differs dramatically from the 90% with mileage. Because of the biased sample concern, we treat the potential inclusion of the mileage variable as an interesting follow-up study.

For some audiences, analysts may wish to present the more flexible negative binomial regression model. Table 6.4 shows that there is little differences in the estimated coefficients for this data set, indicating that the simpler Poisson model is acceptable for some purposes. We will use the Poisson distribution in our out-of-sample analysis in Section 6.4.3, primarily because we can get an analytic expression for the predictive distribution (using the fact that a Poisson sum of independent gammas has a Tweedie distribution).

Table 6.4. *Comparison of Poisson and Negative Binomial Models*

	Poisson		Negative Binomial		Relativity
Effect	Estimate	t-statistic	Estimate	t-statistic	(Poisson)
(Intercept)	-2.636	-70.92	-2.639	-69.67	
Rating Group					
B	0.344	2.85	0.343	2.79	1.41
I	1.043	18.27	1.038	17.64	2.84
M	0.541	8.58	0.539	8.38	1.72
S	-0.069	-1.49	-0.069	-1.48	0.93
Territory					
1	-0.768	-14.02	-0.766	-13.79	0.46
2	-0.641	-12.24	-0.640	-12.04	0.53
3	-0.600	-9.87	-0.598	-9.70	0.55
4	-0.433	-8.81	-0.432	-8.64	0.65
5	-0.265	-5.49	-0.264	-5.37	0.77
Both models use logarithmic exposure as an offset					
Estimated negative binomial dispersion parameter is 2.128					
Reference levels are "A" for Rating Group and "6" for Territory					

Table 6.5 summarizes the fit of a gamma regression severity model. As described in Section 6.3.2, we use total losses divided by number of claims as the dependent variable and the number of claims as the weight. We fit a gamma distribution with a logarithmic link and the two factors, rating group and territory. Table 6.5 shows small t-statistics associated with levels of rating group and territory - only "inexperienced" drivers are statistically significant. Additional tests indicate that the territory factor is not statistically significant and the rating group factor is marginally statistically significant with a $p-value = 0.042$. This is an interesting finding.

Table 6.5 also shows the result of using claim number as an explanatory variable in the severity model. For our data, the variable was not statistically significant and so was not included in subsequent modeling. Had the variable been statistically significant, a proxy would need to be developed for out-of-sample prediction. That is, although we can condition on claim number and it may be a sensible explanatory variable of (average) severity, it is not available apriori and so cannot be used directly for out-of-sample prediction.

As an alternative to the frequency severity approach, we also fit a model using "pure premiums," total losses, as the dependent variable. Similar to the frequency and severity models, we use a logarithmic link function with the factors rating group and territory. The Tweedie distribution was used. We approximated the Tweedie shape parameter p using profile likelihood and found that the value $p = 1.5$ was acceptable. This was the value used in the final estimation.

Table 6.6 reports the fitted Tweedie regression model. The t-statistics associated

Table 6.5. *Gamma Regression Models*

	Without Number		With Number	
Effect	Estimate	t-statistic	Estimate	t-statistic
(Intercept)	7.986	137.33	7.909	76.87
Rating Group				
B	0.014	0.08	0.020	0.11
I	0.222	2.49	0.218	2.43
M	-0.013	-0.13	-0.015	-0.15
S	0.036	0.50	0.038	0.52
Territory				
1	0.026	0.31	0.027	0.31
2	-0.137	-1.67	-0.138	-1.68
3	0.004	0.04	0.005	0.05
4	-0.029	-0.38	-0.029	-0.38
5	0.019	0.26	0.018	0.23
Claim Number	–	–	0.071	0.90
Estimated Dispersion Parameter	2.432		2.445	
Reference levels are "A" for Rating Group and "6" for Territory				

with several levels of rating group and territory are statistically significant. This suggests, and was confirmed through additional testing, that both factors are statistically significant determinants of total loss. The table also reports the relativities (computed as the exponentiated parameter estimates). Interestingly, these relativities turn out to be close to those of the frequency model; this is not surprising given the lack of statistical significance associated with the factors in the severity model.

Table 6.6. *Tweedie Regression Model*

Effect	Estimate	t-statistic	Relativity
(Intercept)	5.356	63.47	
Rating Group			
B	0.340	1.28	1.41
I	1.283	9.39	3.61
M	0.474	3.22	1.61
S	-0.033	-0.36	0.97
Territory			
1	-0.743	-6.53	0.48
2	-0.782	-6.92	0.46
3	-0.552	-4.37	0.58
4	-0.480	-4.44	0.62
5	-0.269	-2.50	0.76
This model uses logarithmic exposure as an offset Estimated dispersion parameter is 2.371 Reference levels are "A" for Rating Group and "6" for Territory			

6.4.3 Out-of-Sample Model Comparisons

To compare the frequency-severity and pure premium approaches, we examined a "held-out" validation sample. Specifically, from our original database, we drew a random sample of 100,000 policies and developed the models reported in Section 6.4.2. Then, we drew an (independent) sample of 50,000 policies. For the frequency-

severity model, our predictions are based on equation (6.6), using Poisson frequency coefficients in Table 6.4 to estimate $\boldsymbol{\beta}_F$, severity coefficients in Table 6.5 to estimate $\boldsymbol{\beta}_S$, and with values of the independent variables from the held-out validation sample. The predictions for the Tweedie model followed similarly using the coefficients reported in Table 6.6.

Figure 6.3 compares the predictions for frequency-severity and the pure premium models. The left-hand panel shows the distribution of our pure premium predictions. The right-hand panel shows the strong relationship between the two predictions; it turns out that the correlation is approximately 0.999. For the purposes of predicting the mean, typically the primary focus of ratemaking, these two models yield virtually indistinguishable predictions. Both models provided some ability to predict total losses; the (Spearman) correlation between held-out losses and (either) predictor turned out to be 8.2%.

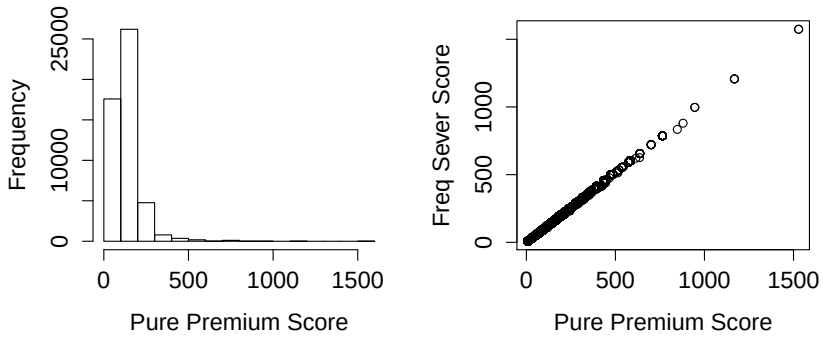

Fig. 6.3. Out-of-Sample Mean Performance. The left-hand panel shows the distribution of the out-of-sample predictions calculated using the pure premium, or Tweedie, model. The right-hand panel shows the strong relationship between the scores from the frequency-severity and the pure premium models.

Because the mean predictor did not provide a way of discriminating between the pure premium and frequency-severity models, we also looked to tail percentiles. Specifically, in the Tweedie regression model, in Section 6.4.2 we cited $p = 1.5$ and $\hat{\phi} = 2.371$ and described how to estimate $\hat{\mu}_i$ for each observation i. Then, in Section 6.3.3 we noted that one could use the `pTweedie` function in `R` to get the distribution function. We did this for *each* held-out observation and evaluated it using the actual realized value. Recall the "probability integral transform," a result in probability theory that says that when a continuous random variable is evaluated using its distribution function, the resulting transformed random variable should have a uniform (on [0,1]) distribution. Thus, if our distribution function calculation is approximately correct, then we can expected the held-out transformed random variables to have an approximate uniform distribution.

The procedure for Poisson frequency and gamma severity models is similar but a bit more complex. In Section 6.3.3, we noted that a Poisson sum of gamma random variables has a Tweedie distribution. So, even though we estimate the frequency and severity parameters separately, they can still be combined when we look at the loss distribution. In display (6.8), we show explicitly how to get Tweedie parameters

from the Poisson frequency and gamma severity models. Then, as with the Tweedie GLM, we can calculate, the transformed (using the distribution function) actual realized value.

Table 6.7 provides the comparisons for selected percentiles. Both models provide disappointing results below the 98th percentile; perhaps this is to be expected for a distribution with approximately 96% zeros. For the 99th percentile and above, the Tweedie does a good job tracking the actual held-out losses. In comparison, the frequency-severity approach is only competitive at the 99.9th percentile. On the one hand, this table suggests that fitting the tails of the distribution is a more complex problem that requires more refined data and sophisticated models. On the other hand, the similarity of results in Figure 6.3 when predicting the mean suggests a robustness of the GLM procedures that gives the analyst confidence when providing recommendations.

Table 6.7. *Out-of-Sample Quantile Performance*

Percentile	Pure Premium	Frequency Severity
0.960	0.50912	0.42648
0.970	0.85888	0.79766
0.980	0.93774	0.86602
0.985	0.97092	0.90700
0.990	0.99294	0.93948
0.995	0.99528	0.97722
0.999	0.99784	0.99870

6.5 Further Reading and References

There is a rich literature on modeling the joint frequency and severity distribution of automobile insurance claims. There has been substantial interest in statistical modeling of claims frequency yet the literature on modeling claims severity, especially in conjunction with claims frequency, is less extensive. One possible explanation, noted by Coutts (1984), is that most of the variation in overall claims experience may be attributed to claim frequency. Coutts (1984) also remarks that the first paper to analyze claim frequency and severity separately seems to be Kahane and Levy (1975), see also Weisberg and Tomberlin (1982).

In the econometrics literature, Cragg (1971) introduced different frequency and severity covariates in two-part models, citing an example from fire insurance. Mullahy (1998) provides and overview of two-part models and discusses healthcare applications.

Brockman and Wright (1992) provide an early overview of how statistical modeling of claims and severity can be helpful for pricing automobile coverage. Renshaw (1994) shows how generalized linear models can be used to analyze both the frequency and severity portions based on individual policyholder level data. At the individual policyholder level, Frangos and Vrontos (2001) examined a claim frequency and severity model, using negative binomial and Pareto distributions, respectively. They used their statistical model to develop experience rated (bonus-malus) premiums.

A trend in recent research has been to explore multivariate frequency-severity

models, examining different lines of business or different perils simultaneously. The first papers in this area seem to be due to Pinquet (1997)Pinquet (1998), fitting not only cross-sectional data but also following policyholders over time. Pinquet was interested in two lines of business, claims at fault and not at fault with respect to a third party. Frees et al. (2012) examine multivariate two-part models for different perils in homeowners insurance. Frees et al. (2013) review multivariate two-part models, examining several types of medical care expenditures jointly.

References

Bowers, N. L., H. U. Gerber, J. C. Hickman, D. A. Jones, and C. J. Nesbitt (1997). *Actuarial Mathematics*. Society of Actuaries.

Brockman, M. J. and T. S. Wright (1992). tatistical motor rating: making effective use of your data. *Journal of the Institute of Actuaries 119*, 457–543.

Coutts, S. M. (1984). Motor insurance rating, an actuarial approach. *Journal of the Institute of Actuaries 111*, 87–148.

Cragg, J. G. (1971). Some statistical models for limited dependent variables with application to the demand for durable goods. *Econometrica 39*(5), 829–844.

de Jong, P. and G. Z. Heller (2008). *Generalized Linear Models for Insurance Data*. Cambridge University Press, Cambridge, UK.

Ferreira, J. and E. Minikel (2010). Pay-as-you-drive auto insurance in Massachusetts: A risk assessment and report on consumer, industry and environmental benefits. In *Conservation Law Foundation, Boston, Mass.* `http://www.clf.org/wp-content/uploads/2010/12/CLF-PAYD-Study_November-2010.pdf`.

Ferreira, J. and E. Minikel (2012). Measuring per mile risk for pay-as-you-drive automobile insurance. *Transportation Research Record: Journal of the Transportation Research Board 2297*, 97–103.

Frangos, N. E. and S. D. Vrontos (2001). Design of optimal bonus-malus systems with a frequency and a severity component on an individual basis in automobile insurance. *ASTIN Bulletin 31*(1), 1–22.

Frees, E., X. Jin, and X. Lin (2013). Actuarial applications of multivariate two-part regression models. *Annals of Actuarial Science*, To appear.

Frees, E., G. Meyers, and A. D. Cummings (2012). Predictive modeling of multi-peril homeowners insurance. *Variance 6*(1), 11–31.

Kahane, Y. and H. Levy (1975). Regulation in the insurance industry: determination of premiums in automobile insurance. *Journal of Risk and Insurance 42*, 117–132.

Klugman, S. A., H. H. Panjer, and G. E. Willmot (2008). *Loss Models: From Data to Decisions*. John Wiley & Sons, Hoboken, New Jersey.

Mullahy, J. (1998). Much ado about two: Reconsidering retransformation and the two-part model in health econometrics. *Journal of Health Economics 17*, 247–281.

Pinquet, J. (1997). Allowance for cost of claims in bonus-malus systems. *ASTIN Bulletin 27*(1), 33–57.

Pinquet, J. (1998). Designing optimal bonus-malus systems from different types of claims. *ASTIN Bulletin 28*(2), 205–229.

Renshaw, A. E. (1994). Modeling the claims process in the presence of covariates. *ASTIN Bulletin 24*(2), 265–285.

Weisberg, H. I. and T. J. Tomberlin (1982). A statistical perspective on actuarial methods for estimating pure premiums from cross-classified data. *Journal of Risk and Insurance 49*, 539–563.

Werner, G. and C. Modlin (2010). *Basic Ratemaking* (4th ed.). Casulty Actuarial Society.

6.6 Appendices

6.6.1 Sample Average Distribution in Linear Exponential Families

The distribution of the linear exponential family with parameters θ and ϕ is

$$\mathrm{f}(y;\theta,\phi) = \exp\left(\frac{y\theta - b(\theta)}{\phi} + S\left(y,\phi\right)\right).$$

With this notation, it can be readily shown (e.g., Frees, 2010, Chapter 13) that the moment generation function can be expressed as

$$M(s;\theta,\phi) \quad = \quad \mathrm{E}\ e^{sy} = \exp\left(\frac{b(\theta+s\phi) - b(\theta)}{\phi}\right).$$

Suppose that $y_1,\ldots,y_m$ are independently distributed with this moment generating function. Then, the moment generating function of the sample average is

$$\begin{aligned}
\mathrm{E}\ \exp\left(s\frac{y_1+\cdots+y_m}{m}\right) &= \prod_{i=1}^{m} \mathrm{E}\ \exp\left(\frac{s}{m}y_i\right) \\
&= \prod_{i=1}^{m} \exp\left(\frac{b(\theta+\frac{s}{m}\phi) - b(\theta)}{\phi}\right) \\
&= \exp\left(\frac{b(\theta+s\frac{\phi}{m}) - b(\theta)}{\phi/m}\right) = M(s;\theta,\phi/m).
\end{aligned}$$

Thus, the sample average is from the same linear exponential family with parameters θ and ϕ/m.

6.6.2 Over-Sampling Claims

If you work with government surveys such as the Medical Expenditure Survey (MEPS) in Chapter 2 or the Survey of Consumer Finances (SCF) in Frees (2010), you will have seen that it is common for such surveys to use unequal probabilities when drawing samples from larger populations. For example, the MEPS data over-samples poor and minority individuals; the SCF over-samples the wealthy. The idea is to draw a larger proportion of a subset of the population that is of interest in the study. In insurance, it is common to "over-sample" policyholders with claims.

Specifically, consider the two-part model introduced in Section 6.2.1 and considered in more detail in Section 6.3.1. Suppose that we have a very large database consisting of $\{r_i, y_i, \mathbf{x}_i\}$, $i = 1,\ldots,N$ observations. We want to make sure to get plenty of $r_i = 1$ (corresponding to claims or "cases") in our sample, plus a sample of $r_i = 0$ (corresponding to non-claims or "controls"). Thus, we split the data set into two pieces. For the first piece consisting of observations with $r_i = 1$, take a random sample with probability τ_1. Similarly, for the second piece consisting of observations with $r_i = 0$, take a random sample with probability τ_0. For example, we might use $\tau_1 = 1$ and $\tau_0 = 0.2$, corresponding to taking all of the claims and a 20% sample of non-claims. Thus, the "sampling weights" τ_0 and τ_1 are considered known to the analyst. This over-sampling procedure is sometimes known as the "case-control" method.

How does this sampling procedure affect the inference in a two-part model?

Think about this question from a likelihood perspective. To develop the likelihood,

let $\{s_i = 1\}$ denote the event that the observation is selected to be included in the sample and $\{s_i = 0\}$ means that it is not included. Suppressing the $\{i\}$ subscript, we decompose the likelihood of the dependent variables that can be observed as:

$$\begin{aligned} \mathrm{f}\,(r, y|s = 1) &= \mathrm{f}\,(r|s = 1) \times \mathrm{f}\,(y|s = 1, r) \\ \text{“observable” likelihood} &= \text{conditional frequency} \times \text{conditional severity.} \end{aligned}$$

For the conditional severity, it is common to assume that $\mathrm{f}\,(y|s = 1, r) = \mathrm{f}\,(y|r)$ - given the absence or presence of a claim, the selection mechanism has no affect on the amount. This is an assumption that may need to be verified but does seem to commonly hold.

For the conditional frequency, here are some basic probability calculations to show how the conditional (on selection) claim frequency relates to the (population) claim frequency. Conditional on $\{r_i = 1\}$, we have that $\Pr(s_i = 1|r_i = 1) = \tau_1$, a Bernoulli distribution. Similarly, $\Pr(s_i = 1|r_i = 0) = \tau_0$. From this, we have

$$\begin{aligned} \Pr(r_i = 1, s_i = 1) &= \Pr(s_i = 1|r_i = 1)\Pr(r_i = 1) = \tau_1\pi_i \\ \Pr(r_i = 0, s_i = 1) &= \Pr(s_i = 1|r_i = 0)\Pr(r_i = 1) = \tau_0(1 - \pi_i) \end{aligned}$$

Thus, the probability of the observation being selected into the sample is

$$\Pr(s_i = 1) = \tau_1\pi_i + \tau_0(1 - \pi_i).$$

Further, the probability of observing a claim in the sample is:

$$\begin{aligned} \Pr(r_i = 1|s_i = 1) &= \frac{\Pr(r_i = 1, s_i = 1)}{\Pr(s_i = 1)} = \frac{\tau_1\pi_i}{\tau_1\pi_i + \tau_0(1 - \pi_i)} \\ &= \frac{\tau_1\pi_i/(1 - \pi_i)}{\tau_1\pi_i/(1 - \pi_i) + \tau_0}. \end{aligned}$$

Now, using the logit form in equation (6.1), we can express the odds ratio as

$$\frac{\pi_i}{1 - \pi_i} = \frac{\frac{E_i}{1+\exp(-\mathbf{x}_i'\boldsymbol{\beta})}}{1 - \frac{E_i}{1+\exp(-\mathbf{x}_i'\boldsymbol{\beta})}} = \frac{E_i}{1 + \exp(-\mathbf{x}_i'\boldsymbol{\beta}) - E_i}. \tag{6.9}$$

Thus,

$$\begin{aligned} \Pr(r_i = 1|s_i = 1) &= \frac{\tau_1 \frac{E_i}{1+\exp(-\mathbf{x}_i'\boldsymbol{\beta})-E_i}}{\tau_1 \frac{E_i}{1+\exp(-\mathbf{x}_i'\boldsymbol{\beta})-E_i} + \tau_0} = \frac{\tau_1 E_i}{\tau_1 E_i + \tau_0(1 + \exp(-\mathbf{x}_i'\boldsymbol{\beta}) - E_i)} \\ &= \frac{\tau_1 E_i}{c_i + \tau_0 \exp(-\mathbf{x}_i'\boldsymbol{\beta})}, \end{aligned}$$

where $c_i = \tau_1 E_i + \tau_0(1 - E_i)$. From this, we can express the probability of observing a claim in the sample as

$$\Pr(r_i = 1|s_i = 1) = \frac{E_i^*}{1 + \gamma_i \exp(-\mathbf{x}_i'\boldsymbol{\beta})} \tag{6.10}$$

where $E_i^* = \tau_1 E_i/c_i = \frac{\tau_1 E_i}{\tau_1 E_i + \tau_0(1-E_i)}$ and $\gamma_i = \tau_0/c_i = \frac{\tau_0}{\tau_1 E_i + \tau_0(1-E_i)}$.

In summary, equation (6.10) has the same form as equation (6.1) with a new definition of exposure and the introduction of an offset term, $-\ln \gamma_i$, assuming that

you use logistic regression (not probit) for your claim frequency modeling. If all of your exposures are identically equal to 1 ($E_i \equiv 1$), then γ_i is a constant and you simply re-interpret the constant in the systematic component $\mathbf{x}_i'\boldsymbol{\beta}$ (which we typically ignore). If exposures are not constant, then equation (6.10) gives a straightforward method of adjusting the exposure and introducing an offset term, running the usual logistic regression software without the need for specialized software routines.