The Tweedie Index Parameter and Its Estimator

An Introduction with Applications to Actuarial Ratemaking

Seth David Temple

A thesis presented for the degree of
Bachelor of Science

Mathematics Department and Robert D. Clark Honors College
University of Oregon
June 2018

An Abstract of the Thesis of

Seth David Temple for the degree of Bachelor of Science
in the Department of Mathematics to be taken June 2018

Title: The Tweedie Index Parameter and Its Estimator: An Introduction with Applications to Actuarial Ratemaking

Approved: _____
Chris Sinclair

Tweedie random variables are exponential dispersion models that have power unit variance functions, are infinitely divisible, and are closed under translations and scale transformations. Notably, a Tweedie random variable has an indexing/power parameter that is key in describing its distribution. Actuaries typically set this parameter to a default value, whereas R's tweedie package provides tools to estimate the Tweedie power via maximum likelihood estimation. This estimation is tested on simulations and applied to an auto severity dataset and a home loss cost dataset. Models built with an estimated Tweedie power observe lower Akaike Information Criterion relative to models built with default Tweedie powers. However, this parameter tuning only marginally changes regression coefficients and model predictions. Given time constraints, we recommend actuaries use default Tweedie powers and consider alternative feature engineering.

Acknowledgments

First and foremost, I want to acknowledge the time and effort my thesis defense committee members put in to support my education. I appreciate Peter Ralph and Samantha Hopkins for reading this thesis and giving thoughtful feedback. Most importantly, Chris Sinclair was an indispensible advisor to me. He met with me weekly for a year, worked through proofs with me, exposed me to new mathematics, and assisted in the editing process. I always left our meetings more inspired to pursue mathematics.

Next, I want to thank my former manager Thomas Wright. He supervised my summer project where I applied geographic information systems (GIS) to territorial ratemaking, encouraged me to study generalized linear modeling, provided literature recommendations, and helped me receive approval to use the encrypted Liberty Mutual dataset seen in this thesis. As great bosses do, Tommy supported my drive to do research and challenge the way we do things.

Lastly, I want to thank my parents and my brother for their unconditional love. I have not been the easiest child at times. Show me a math student who has been such. My mom sometimes texts Matt and me: "You two are my best sons." I respond: "Mom, that's a vacuous statement. Matt and I are your only two sons. We are the best, the worst, and everything in between."

Table of Contents

List of Figures

Foreword

Mathematics is a cumulative study. I have studied math continuously for close to 16 years. The material I present in this thesis draws from advanced probability courses I took during my third year of undergraduate schooling, lessons I learned in statistical modeling as an actuarial intern with Liberty Mutual, and a year's worth of independent research. I do not expect the reader to have any prior exposure to the Tweedie family of distributions or to other statistical concepts discussed in this paper. One aim of this work is to introduce the mathematically literate person to a new class of probability distributions and to explain why actuaries incorporate these distributions into insurance models. On the other hand, I expect the reader to have some basic knowledge of probability and statistics. This text is written for an audience of STEM (Science, Technology, Engineering, and Mathematics) students and professionals who have taken one or two undergraduate courses in statistics. More precisely, this thesis speaks to a niche audience of actuaries.

The main topic of this thesis is the Tweedie family of probability distributions. Bent Jorgensen named this family in honor of the twentieth century hospital physician and statistician Maurice Charles Kenneth Tweedie [15]. Tweedie first discussed these distributions in his 1947 paper "Functions of a Statistical Variate with Given Means, with Special Reference to Laplacian Distributions." I focus on the Tweedie family because it includes many famous distributions that are used as response distributions in generalized linear models (GLMs). Actuaries use GLMs to predict the frequency of claims, the severity of claims, and the combined loss costs. Much of my research follows in the footsteps of prominent mathematicians who have written on the Tweedie family and GLMs: MCK Tweedie, Bent Jorgensen, Peter Dunn, Gordon Smyth, John Nelder, and Robert Wedderburn.

The idea for this project came by accident. During the spring of 2017, Professor Chris Sinclair and I had been reading about and discussing renewal processes. Then,

while working at Liberty Mutual in Seattle during the summer of 2017, I stumbled upon some SAS code that specified a model's response distribution as Tweedie. Out of curiosity, I browsed the Internet to learn more about this new distribution. I quickly noticed how some Tweedie distributions relate to Wald's equation–something Chris and I had previously talked about. Eureka! I had found a connection between my undergraduate research and my summer internship. When I pressed some of my colleagues at Liberty Mutual to learn more about Tweedie distributions and why we set the parameter to a particular value, I came out of the conversations dissatisfied. As a result, I redirected my research and set out to investigate the Tweedie power parameter in more detail.

Besides providing educational value to the reader, this thesis proposes methods to improve Tweedie severity and loss cost models. Actuaries often exercise judgment in selecting the power parameter. This judgment could mean choosing the power parameter to be 1.5 for pure premium models and 2 or 3 for severity models. But what if the empirical data suggests a different power parameter to be more likely? Would picking the parameter value that better fits the underlying data enhance the predictive power of the model? How easy is it to find an optimal power parameter? I address these questions in the applied work of this paper.

Chapter 1 introduces the mathematics behind the Tweedie family and GLMs. This section reads like a math textbook. It is my aim to bring this material down to a level that is accessible to the undergraduate statistician. Chapter 2 summarizes some papers by Peter Dunn and Gordon Smyth where they use numerical approximation techniques to evaluate Tweedie densities because, in general, Tweedie random variables don't have a closed form density. Their work enables contemporary statisticians to use Tweedie distributions in R. In Chapter 3, I design an experiment to evaluate if estimating the Tweedie power improves a predictive model. Chapters 4 and 5 report the results of the experiment. In Chapter 6, I give an interdiscplinary

analysis of estimating the Tweedie power in a business setting. All in all, I hope that you find something interesting or useful about the Tweedie family in this thesis.

1 The Tweedie Family and Linear Models

Applied probability assumes empirical data behaves according to a theoretical distribution. For example, I could claim by making statistical arguments that a Poisson random variable describes the annual count of earthquakes near Seattle. A Poisson random variable is characterized by the density function

$$f(y;\lambda) = \begin{cases} \frac{e^{-\lambda}\lambda^y}{y!}, & \forall y \in \mathbb{N} \\ 0, & \text{otherwise} \end{cases}.$$

There are many distributions in probability theory: binomial, hypergeometric, negative binomial, Cauchy, normal, et cetera. Random variables define probability distributions, density functions, and other useful functions such as the moment-generating function and the cumulant-generating function. These random variables come with parameters that determine the shape of such functions. For instance, a Poisson random variable comes with the parameter λ and a normal random variable comes with parameters μ (mean) and σ^2 (variance). In building a statistical model, the modeler must select a distribution to describe the empirical data and set values for some parameters.

Actuaries use statistics, financial mathematics, and business intelligence to price insurance policies and to reserve money for claim payments. These insurance professionals often build statistical models to solve regression problems. Regression analysis examines the relationship between a dependent target variable and independent explanatory variables.

The dependent variable in an insurance model is typically either claim frequency, claim severity, or loss costs. Claim frequency measures how many claims a policyholder files. Actuaries build Poisson models to predict claim frequency. Claim severity measures the monetary cost of a claim. A claim severity dataset contains

only observations of filed claims. Actuaries make gamma or inverse-Gaussian models to predict claim severity. Loss costs measure the amount an insurer pays to indemnify a policyholder. Most of the time, policyholders do not file claims. Actuaries use Poisson-gamma models to predict loss costs.

Maurice Charles Kenneth (MCK) Tweedie put forward a framework that encompasses all these random variables in one class [26, 27]. We call that class of random variables the Tweedie family and the distributions within it Tweedie distributions. The Tweedie family is a robust family of probability distributions. It includes a discrete random variable (Poisson), a mixed random variable (Poisson-gamma), continuous random variables, and stable random variables. Tweedie models are implemented often in regression analysis, but they are seldom understood past a superficial level. By providing more context, I hope to inspire actuarial modelers to be more accurate and creative in their application of Tweedie models.

Besides actuarial science, other discliplines employ Tweedie models. Tweedie models have been used to describe monthly rainfall in Australia [12] and to perform catch per unit effort analyses for fisheries research [24]. MCK Tweedie cites himself and other statisticians who implement Tweedie models in the biological and medical sciences [26]. Studying this family in more detail will be fruitful for many researchers, not just actuaries.

Exponential Dispersion Models

The Tweedie family is a subset of a class of random variables described by Bent Jorgensen in ***The Theory of Dispersion Models***. As a result, we must first cover exponential dispersion models (EDMs) before we discuss the Tweedie family. Jorgensen presents two descriptions of EDMs in his monograph: one axiomatic and one constructive. The axiomatic version defines EDMs without justifying the origins of the distributions; the constructive version begins with a cumulant function and builds the

theory from there. (Later, Jorgensen proves that the axiomatic definition fits with the constructive definition.) Here we provide the axiomatic definition for exponential dispersion models [15].

Ideas inspire more ideas. This adage applies to the development of exponential dispersion models. EDMs maintain the structure of the normal distribution. In order to talk about this structure in abstract terms, we must establish some definitions.

Definition 1.1. Let f be a real-valued density function for the random variable X. The *support* of X is the set of elements in the domain of f that do not map to zero. The *convex support* of X is the smallest interval containing the support.

Definition 1.2. Let C be a convex support and let Ω be the interior C. Ω and C are intervals satisfying that $\Omega \subset C \subset \mathbb{R}$. A *unit deviance* is a function $d : C \times \Omega \to \mathbb{R}$ that satisfies the following:

(i) $d(y; y) = 0, \ \forall y \in \Omega$;

(ii) $d(y; \mu) > 0, \ \forall y \neq \mu$.

Remark. This definition looks familiar to the definition of a metric, except without the triangle inequality. Essentially, the unit deviance is a tool to measure distance.

Equipped with these two definitions, we are ready to present the definition of an exponential dispersion model. Consider the density function for a normal random variable $N(\mu, \sigma^2)$:

$$f(y; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(y-\mu)^2\right\}.$$

Density functions for exponential dispersion models share this format.

Definition 1.3. An *exponential dispersion model* $EDM(\mu, \sigma^2)$ is a probability distribution whose density function with respect to a suitable measure has the form

$$f(y; \mu, \sigma^2) = a(y; \sigma^2) \exp\left\{-\frac{1}{2\sigma^2} d(y; \mu)\right\}, \quad y \in C$$

where $a \geq 0$ is a suitable function, d is a unit deviance of the form $d(y; \mu) = yg(\mu) + h(\mu) + k(y)$, C is the convex support, $\mu \in \Omega = \text{int } C$, and Ω is an interval.

Remark. This definition mentions measure theory. This thesis doesn't cover probability from a measure-theoretic lens. We consider $a(y; \sigma^2)$ suitable if the axioms of probability hold. It suffices to check that

$$\int f(y; \mu, \sigma^2)\, dy = 1.$$

Note. We call μ the position parameter and σ^2 the dispersion parameter. This language and notation draws from normal theory. Figure 1 shows the normal bell curve [2]. μ is where the center of mass of the distribution is located. σ^2 describes how spread out the mass of the distribution is.

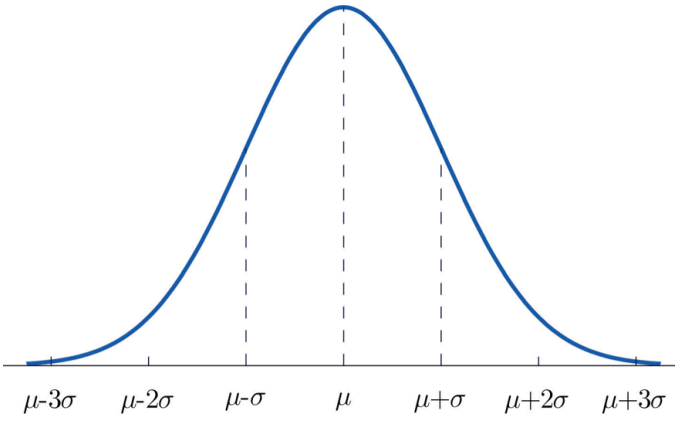

Figure 1: Bell Curve Interpretation of μ and σ.

Proposition 1.1. The following distributions are exponential dispersion models:

(i) normal distribution;

(ii) Poisson distribution;

(iii) binomial distribution;

(iv) gamma distribution.

Proof. To show a distribution is an EDM, we first propose $a(y;\sigma^2)$ and $d(y;\mu)$. Next, we argue that the unit deviance d has the correct form. Lastly, we do the algebra necessary to show that $f(y;\mu,\sigma^2)$ corresponds with the distribution's usual density function.

(i) Consider a $N(\mu,\sigma^2)$ random variable with $\mu \in \mathbb{R}$, $\sigma^2 \in \mathbb{R}_+$, and $y \in \mathbb{R}$. Let $a(y;\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}$ and $d(y;\mu) = (y-\mu)^2$. Notice that

$$\begin{aligned} d(y;\mu) &= (y-\mu)^2 \\ &= y^2 - 2y\mu + \mu^2. \end{aligned}$$

Define $g(x) = -2x$, $h(x) = x^2$, and $k(x) = x^2$. The unit deviance is of the right form and $f(y;\mu,\sigma^2)$ matches the usual normal density.

(ii) Consider a Poisson random variable is defined with $\mu \in \mathbb{Z}_+$ and $y \in \mathbb{Z}_{0+}$. Let $a(y;\sigma^2) = \frac{y^y}{y!}\exp\{-y\}$ and $d(y;\mu) = 2(y\log\frac{y}{\mu} - y + \mu)$. Here the $a(y;\sigma^2)$ function is independent of the dispersion parameter σ^2. For a Poisson random variable, $\sigma^2 = 1$. Notice that

$$\begin{aligned} d(y;\mu) &= 2\Big(y\log\frac{y}{\mu} - y + \mu\Big) \\ &= 2y\log y - 2y\log\mu - 2y + 2\mu. \end{aligned}$$

Define $g(x) = -2\log x$, $h(x) = 2x$, and $k(x) = 2x\log x - 2x$. The unit deviance is of the right form. Now,

$$\begin{aligned} f(y;\mu,1) &= a(y)\exp\left\{\frac{-1}{2}\cdot 2\left(y\log\frac{y}{\mu} - y + \mu\right)\right\} \\ &= a(y)\exp\left\{y\log\mu - y\log y + y - \mu\right\} \\ &= a(y)e^{y\log\mu}e^{-y\log y}e^{y}e^{-\mu} \\ &= a(y)y^{\mu}y^{-y}e^{y}e^{-\mu} \\ &= \frac{1}{y!}y^{\mu}e^{-\mu}y^{-y}y^{y}e^{y}e^{-y} \\ &= \frac{1}{y!}y^{\mu}e^{-\mu}. \end{aligned}$$

Thus, $f(y;\mu,\sigma^2)$ matches the usual Poisson probability mass function.

(iii) To prove that a binomial random variable is an EDM, we adjust our notation slighlty. In practice, binomial random variables occur when we have an event with two possible outcomes and we want to know the probability of an outcome occurs m times out of n independent attempts. Let the probability space of random variable X be $\{0,1\}$ and set $Pr(X=1) = p$. Replace y with m, μ with p, and σ^2 with n. Also, we know the parameter n because it is how many times we run the experiment. Fix n to be a positive integer.

Consider a binomial random variables with $p \in (0,1)$, $m \in \mathbb{Z}_{0+}$, and $n \in \mathbb{Z}_{+}$. Let $a(m;n) = \binom{n}{m}$ and $d(m;p) = -2n\Big(m\log(\frac{p}{1-p}) + n\log(1-p)\Big)$. Notice that

$$\begin{aligned} d(m;p) &= -2n\Big(m\log\left(\frac{p}{1-p}\right) + n\log(1-p)\Big) \\ &= -2nm\log\left(\frac{p}{1-p}\right) - 2n^2\log(1-p). \end{aligned}$$

Define $g(x) = -2n\log(\frac{x}{1-x})$, $h(x) = -2n^2\log(1-x)$, and $k(x) = 0$. The unit

deviance is of the right form. Now,

$$\begin{aligned} f(m;p,n) &= a(m;n)\exp\left\{\frac{-1}{2n}\cdot(-2n)\Big(m\log\Big(\frac{p}{1-p}\Big)+n\log(1-p)\Big)\right\} \\ &= \binom{n}{m}\exp\left\{m\log\Big(\frac{p}{1-p}\Big)+n\log(1-p)\right\} \\ &= \binom{n}{m}\exp\Big\{m\log p+(n-m)\log(1-p)\Big\} \\ &= \binom{n}{m}p^m(1-p)^{n-m}. \end{aligned}$$

Thus, $f(y;\mu,\sigma^2)$ matches the usual binomial probability mass function.

(iv) We again adjust the notation to prove that a gamma random variable is an EDM. Conventionally, a gamma random variable is parameterized by shape parameter α and rate parameter β. The density is

$$f(y;\alpha,\beta)=\frac{\beta^\alpha y^{\alpha-1}e^{-\beta y}}{\Gamma(\alpha)}$$

where the the gamma function is defined as

$$\Gamma(x)=\int_0^\infty y^{x-1}e^{-y}dy\ .$$

Set $\mu=\alpha/\beta$ and $\sigma^2=1/\alpha$. $\alpha>0$ and $\beta>0$, so $\mu>0$ and $\sigma^2>0$. y takes values in $\mathbb{R}_+$.

Let $a(y;\sigma^2)=a(y;1/\alpha)=\frac{1}{\Gamma(\alpha)}\alpha^\alpha e^{-\alpha}y^{-1}$ and $d(y;\mu)=2\Big(\frac{y}{\mu}-\log(\frac{y}{\mu})-1\Big)$. We omit the argument that unit deviance d is in the correct form. It mimics previous

arguments. Now,

$$\begin{aligned} f(y;\mu,\sigma^2) &= f(y;\alpha/\beta,1/\alpha) \\ &= a(y;1/\alpha)\exp\left\{\frac{-\alpha}{2}\cdot 2\Big(y\beta/\alpha-\log(\frac{y\beta}{\alpha})-1\Big)\right\} \\ &= a(y;1/\alpha)\exp\left\{-y\beta+\alpha\log(y\beta)-\alpha\log(\alpha)+\alpha\right\} \\ &= \frac{1}{\Gamma(\alpha)}\alpha^{\alpha}\alpha^{-\alpha}e^{-\alpha}e^{\alpha}(y\beta)^{\alpha}y^{-1}e^{-y\beta} \\ &= \frac{1}{\Gamma(\alpha)}\beta^{\alpha}y^{\alpha-1}e^{-y\beta}. \end{aligned}$$

Thus, $f(y;\mu,\sigma^2)$ matches the usual gamma density function.

Note. The equations $\mu=\alpha/\beta$ and $\sigma^2=1/\alpha$ are useful if you use both tweedie and gamma functions in the statistical software R. The tweedie functions expect parameters μ and σ^2 whereas the gamma functions expect parameters α and β.

□

Remark. The proof of this proposition highlights the flexibility of the $a(y;\sigma^2)$ function in the EDM framework. This flexibility is needed to relate dissimilar distributions within the same framework.

It is important to underline that the dispersion parameter σ^2 is not, in general, equal to the variance of the random variable. Let $X\sim \text{EDM}(\mu,\sigma^2)$. The following is true:

(i.) $\text{E}[X]=\mu$;

(ii.) There exists a function $V(\mu)$ such that $\text{Var}(X)=\sigma^2\cdot V(\mu)$.

$V(\cdot)$ is called the unit variance function. If $V(\mu)=1$, then the $\text{Var}(X)=\sigma^2$. When $V(\mu)\neq 1$, then $\text{Var}(X)\neq\sigma^2$. This logic makes sense, assuming that we know what a unit variance function is. We provide a formal definition below.

Definition 1.4. The unit deviance d is *regular* if $d(y;\mu)$ is twice continuously differentiable with respect to (y,μ) on $\Omega \times \Omega$ and satisfies $\frac{\partial^2 d}{\partial \mu^2}(\mu;\mu) > 0, \ \forall \mu \in \Omega$.

Definition 1.5. The *unit variance function* $V : \Omega \to \mathbb{R}_+$ of a regular unit deviance is

$$\frac{2}{\frac{\partial^2 d}{\partial \mu^2}(\mu;\mu)}.$$

Math is taught by examples. Let's consider some examples of regular unit deviances and compute their corresponding variance functions.

Example 1.1. $d(y;\mu) = (y-\mu)^2$ is a regular unit deviance. The expansion $y^2 - 2y\mu + \mu^2$ is clearly twice continuously differentiable with respect to (y,μ). Compute the partial derivatives:

$$\frac{\partial d}{\partial \mu} = -2y + 2\mu;$$

$$\frac{\partial^2 d}{\partial \mu^2} = 2.$$

By definition, $V(\mu) = 1$. Recall that the unit deviance $(y-\mu)^2$ belongs to the normal random variable. Therefore, the variance of a normal random variable is equal to the value of its dispersion parameter.

Example 1.2. The Poisson unit deviance is $2y \log y - 2y \log \mu - 2y + 2\mu$. This unit deviance is twice continuously differentiable with respect to (y,μ). $\frac{\partial^2 d}{\partial \mu^2} = \frac{2y}{\mu^2}$, so $\frac{\partial^2 d}{\partial \mu^2}(\mu;\mu) = 2/\mu$. The variance function is μ. If $\sigma^2 = 1$, as it does when the Poisson random variable isn't overdispersed or underdispersed, the variance equals the expectation.

There is a nice lemma that provides the mathematician flexibility in computing the variance function.

Proposition 1.2. Let d be a regular unit deviance. $\frac{\partial^2 d}{\partial y^2}(\mu;\mu) = \frac{\partial^2 d}{\partial \mu^2}(\mu;\mu) = -\frac{\partial^2 d}{\partial y \partial \mu}(\mu;\mu)$, $\forall \mu \in \Omega$.

Proof. By definition, we know that $d(\mu;\mu) = 0$ and $d(y;\mu) > 0 \ \forall y \neq \mu$. This is sufficient to claim that $d(\cdot;\mu)$ has a local minimum at μ. Thus, $\frac{\partial d}{\partial y}(\mu;\mu) = 0$ and $\frac{\partial d}{\partial \mu}(\mu;\mu) = 0$. Adding by "0" gives that $\frac{\partial d}{\partial \mu}(\mu;\mu)+\frac{\partial d}{\partial y}(\mu;\mu) = 0$. Since d is regular, it is twice continuously differentiable with respect to (y,μ). Take partial derivatives:

$$\frac{\partial^2 d}{\partial \mu^2}(\mu;\mu) + \frac{\partial^2 d}{\partial \mu \partial y}(\mu;\mu) = 0;$$
$$\frac{\partial^2 d}{\partial y^2}(\mu;\mu) + \frac{\partial^2 d}{\partial y \partial \mu}(\mu;\mu) = 0.$$

Subtract by $\frac{\partial^2 d}{\partial \mu \partial y}(\mu;\mu)$. Thus,

$$\frac{\partial^2 d}{\partial \mu^2}(\mu;\mu) = -\frac{\partial^2 d}{\partial \mu \partial y}(\mu;\mu) = \frac{\partial^2 d}{\partial y^2}(\mu;\mu). \qquad \square$$

So far, we have established a vocabulary to talk about the EDMs, identified some members in the family, and figured out how to compute the variance function. What is remarkable about the exponential family is that dissimilar distributions fit into the same framework. As a demonstration, we simulate realizations of some random variables belonging to the exponential family. Using the ggplot2 package in R, we plot the empirical distributions of these random variables. Consider the following figures.

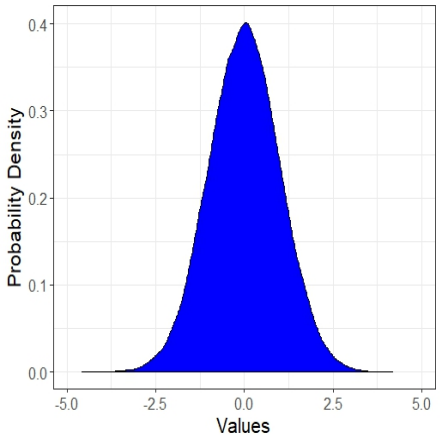

Figure 2: 100,000 Simulated N(0,1) Random Variables

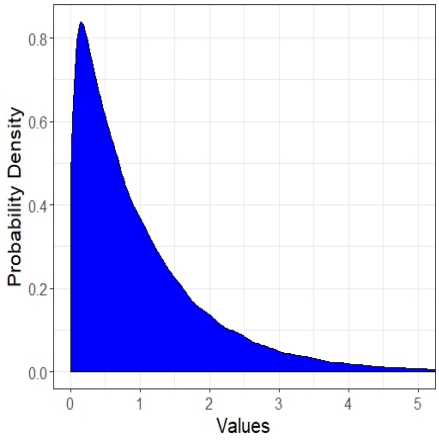

Figure 3: 100,000 Simulated Gamma($\alpha = 1$, $\beta = 1$) Random Variables

The gamma distribution and the normal distribution look similar; you could argue that the gamma distribution is just a left-skewed normal distribution. Pay attention to the x-axis labels. Gamma random variables take only positive real values whereas normal random variables take positive and negative real values.

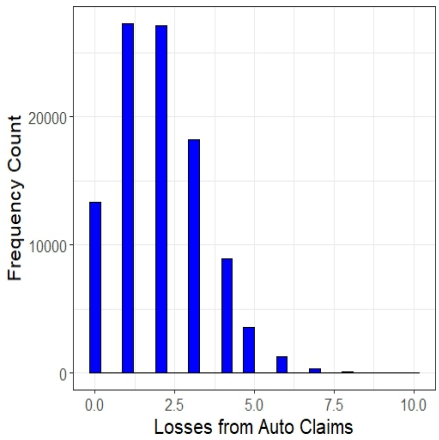

Figure 4: 100,000 Simulated Poisson($\lambda = 2$) Random Variables

Certainly this graph looks quite dissimilar from the previous two graphs. Poisson random variables are discrete; that is, they only have positive probability for countably many values. In contrast, normal and gamma random variables are continuous. It is nontrivial work to relate these dissimilar distributions by just two parameters. The diversity of this family makes it a good candidate to describe a variety of datasets.

The Tweedie Family

Tweedie models are exponential dispersion models closed under scale transformations and translations. We denote a Tweedie model as $Tw_p(\mu, \sigma^2)$. Notice how this notation includes a subscript p. Throughout this text, we will refer to p as either the power parameter or as an indexing parameter. MCK Tweedie presented a paper titled "An index which distinguishes between some important exponential families" at the Indian Statistical Institute's Golden Jubilee Conference in 1984, hence the language indexing parameter [26]. The parameter p can be seen as an index that identifies the type of Tweedie random variable. Its role in the variance functions of Tweedie models earns it the designation as the power parameter. Tweedie models have variance functions of the form $V(\mu) = \mu^p$, hence the language power parameter. Below is a table of the Tweedie models and their powers [11, 15].

Distribution	**Domain**	p **value**
Stable*	$\mathbb{R}$	$p < 0$
Normal (Gaussian)	$\mathbb{R}$	0
Poisson	$\mathbb{N}$	1
Compound Poisson-gamma	$\mathbb{R}_{0+}$	$1 < p < 2$
Gamma	$\mathbb{R}_{+}$	2
Stable*	$\mathbb{R}_{+}$	$2 < p < 3$
Inverse Gaussian	$\mathbb{R}_{+}$	3
Stable*	$\mathbb{R}_{+}$	$p > 3$

Table 1: Tweedie models based on indexing parameter p. * indicates that the model is only stable for some parameter choices.

Before we give a theorem that specifies what makes Tweedie models distinct from other EDMs, we must clarify a definition and argue two lemmas.

Definition 1.6. Let X be a random variable. We say X is *infinitely divisible* if $\forall n \in \mathbb{N}$ we can write X as the sum of n independent, identically distributed (i.i.d.) random variables.

Lemma 1.1. Let X be a random variable. Then, $\mathrm{Var}(cX) = c^2\mathrm{Var}(x)$.

Proof. From Ross [21], we use that $\mathrm{Var}(X) = \mathrm{E}[X^2] - \mathrm{E}[X]^2$.

$$\begin{aligned}\mathrm{Var}(cX) =& \mathrm{E}[(cX)^2] - (\mathrm{E}[cX])^2 \\ =& c^2\mathrm{E}[X^2] - c^2\mathrm{E}[X]^2 \\ =& c^2(\mathrm{E}[X^2] - \mathrm{E}[X]^2) \\ =& c^2\,\mathrm{Var}(X).\end{aligned}$$

□

Lemma 1.2. Let g be some function that is at least once differentiable, takes positive real inputs, and for which $g(xy) = g(x)g(y)$ holds. Then $g(x) = x^a$ where a is some constant.

Proof. Consider a function f such that $f(x) = \log g(e^x)$. So,

$$f(x+y) = \log g(e^{x+y}) = \log g(e^x e^y) = \log g(e^x)g(e^y) = \log g(e^x) + \log g(e^y) = f(x) + f(y).$$

Also, $f(c \cdot x) = c \cdot f(x)$ for some constant c. Thus, f is linear. If f is linear, then g must have been exponential. □

Theorem 1.1. Let X be an $EDM(\mu, \sigma^2)$ such that $V(1) = 1$ and V is at least once differentiable. If there exists a function $f : \mathbb{R}_+ \times \Sigma \to \Sigma$ for which

$$cX = EDM(c\mu, f(c, \sigma^2)) \ \forall c > 0$$

holds, then:

(i.) X is a Tweedie model;

(ii.) $f(c, \sigma^2) = c^{2-p}\sigma^2$;

(iii.) X is infinitely divisible.

Proof. According to Lemma 1.1, $\mathrm{Var}(cX) = c^2\mathrm{Var}(X) = c^2\sigma^2 V(\mu)$. Our assumption gives that $\mathrm{Var}(cX)$ equals $f(c, \sigma^2)V(c\mu)$ as well. Set these two expressions to be equal to

one another.

$$c^2\sigma^2 V(\mu) = f(c, \sigma^2)V(c\mu).$$

Divide by $f(c, \sigma^2)$. This division is legal because the codomain of f is $\mathbb{R}_+$. Now,

$$\frac{c^2\sigma^2}{f(c, \sigma^2)}V(\mu) = V(c\mu).$$

Take $\sigma^2 = 1$. Then

$$\frac{c^2}{f(c^2, 1)} = V(c)$$

because $V(1) = 1$ from our initial assumptions. This maneuver suggests that $\frac{c^2\sigma^2}{f(c,\sigma^2)}$ is $V(c)$, though we don't yet know what this $V(\cdot)$ function is. In any event, we have the relation

$$V(c\mu) = V(c)V(\mu).$$

We use Lemma 1.2 to claim that $V(\mu) = \mu^p$ for some $p \in \mathbb{R}$. Next, we compare

$$\frac{c^2\sigma^2}{f(c, \sigma^2)} = V(c) = c^p$$

and conclude that $f(c, \sigma^2) = c^{2-p}\sigma^2$. For $p \neq 2$, $f(c, \sigma^2)$ varies in $\mathbb{R}_+$ because σ^2 and c vary in $\mathbb{R}_+$. This scalability allows us to construct X as a sum of i.i.d. cX random variables; that is, X is infinitely divisible for $p \neq 2$. When $p = 2$, X is a gamma random variable. It is well-known that gamma random variables are infinitely divisible. Thus, X is infinitely divisible in the $p = 2$ case as well. □

Corollary 1.1. Tweedie models are closed under scale transformations. For positive real constant $c > 0$ and a Tweedie model $Tw_p(\mu, \sigma^2)$,

$$c \cdot Tw_p(\mu, \sigma^2) = Tw_p(c\mu, c^{2-p}\sigma^2).$$

Proof. This corollary follows immediately from Theorem 1.1. □

Corollary 1.1 provides a useful trick for scaling Tweedie models. Scaling $X \sim N(\mu, \sigma^2)$ by c results in $cX \sim N(c\mu, c^2\sigma^2)$. Likewise, $c \cdot Ga(\alpha, \beta)$ results in a $Ga(\alpha, \beta/c)$ random variable because $Tw_2(\alpha/\beta, 1/\alpha) = Ga(\alpha, \beta)$. These examples showcase how easy it is to scale Tweedie models. Another useful property of Tweedie models is that we can translate them, i.e. add or subtract by a constant value.

Theorem 1.2. Let X be an $EDM(\mu, \sigma^2)$ closed under translation and with differentiable unit variance function. This closure means that there exists a function $h(c, \sigma^2)$ such that

$$c + X = EDM(c + \mu, h(c, \sigma^2)), \ \forall c \in \mathbb{R}.$$

Such an EDM is infinitely divisible and has an exponential unit variance function.

Proof. Evaluate the variances on both sides.

$$\sigma^2 V(\mu) = h(c, \sigma^2) V(c + \mu).$$

Thus $V(c + \mu) = g(c)V(\mu)$ where $g(c) = \sigma^2/h(c, \sigma^2)$. Because $V(\cdot)$ is differentiable and positive, g is differentiable and positive. c varies in $\mathbb{R}$, so we can differentiate with respect to it. Differentiate with respect to c at 0. We get $V'(\mu) = g'(0)V(\mu)$. Next we solve the differential equation.

$$\int \frac{1}{V(\mu)} V'(\mu) = \int g'(0)$$

$$\log V(\mu) = \mu g'(0) + c_0$$

$$V(\mu) = c_0 \exp\{\mu g'(0)\}$$

Here $c_0 > 0$ stands for a constant. This constant is a consequence of indefinite integration. Clearly $V(\mu)$ is an exponential function. Lastly, use substitution to solve for $h(c, \sigma^2)$ in the equation

$$h(c, \sigma^2) V(c + \mu) = \sigma^2 V(\mu).$$

$h(c, \sigma^2) = \sigma^2 \exp\{-g'(0)c\}$. When $g'(0) = 0$, the unit variance function corresponds to that of a normal random variable. It is well-known that a sum of i.i.d. normal random variables is a normal random variable. When $g'(0) \neq 0$, $h(c, \sigma^2)$ varies with $c \in \mathbb{R}$, so we can construct a sum of i.i.d. random variables. Thus X is infinitely divisible. □

Remark. Technically speaking, the unit variance function in Theorem 1.2 does not have the form $V(\mu) = \mu^p$ for some $p \in \mathbb{R}$. Describing Tweedie models as EDMs that have unit variance $V(\mu) = \mu^p$ makes it easy to memorize, but we must relax this definition to include closure under translation. Tweedie models are EDMs that are infinitely divisible and closed under translation and scale transformations.

Up to now, we have stated that Tweedie models have unit variance $V(\mu) = \mu^p$ for some $p \in \mathbb{R}$, and we have provided a few examples of possible values for the indexing parameter. But, the real number line is infinitely large. Surely there must be some values for the power parameter which do not work.

Proposition 1.3. There are no Tweedie models with index parameter $0 < p < 1$.

The proof of this proposition involves moment-generating functions and cumulant-generating functions. Because the main body of this text targets an undergraduate audience, we reserve the proof for the Appendix. The argument synthesizes many pages of Jorgensen's book *The Theory of Dispersion Models* [15].

Since Tweedie models are EDMs, they have unit deviances. Proposition 1.1 gives the unit deviances for the normal, Poisson, and gamma cases. But, what about when $p \notin \{0, 1, 2\}$? To justify the unit deviance for the general case, we again require advanced facility with EDMs. As a result, we leave the proof of the following proposition for the Appendix.

Proposition 1.4. For a Tweedie model with index power $p \notin \{0, 1, 2\}$, the unit deviance $d(y; \mu)$ is

$$2\left\{\frac{[\max\{y, 0\}]^{2-p}}{(1-p)(2-p)} - \frac{y\mu^{1-p}}{1-p} + \frac{\mu^{2-p}}{2-p}\right\}.$$

This proposition concludes our introduction to the Tweedie family. We now know some useful properties about Tweedie models that distinguish them as a special class of EDMs. In the following sections, we discuss how and why actuaries use Tweedie random variables in generalized linear modeling.

A Special Look at Tweedie Distributions Used in Insurance Ratemaking

You don't have to read too many scientific papers before you encounter a normal, Poisson, or gamma random variable. These Tweedie models are very applicable to real-life data. Actuaries also use Poisson-gamma sums in insurance ratemaking. For these Tweedie models, the densities have domain $\mathbb{R}_{0+}$; that is, the random variable can take on any nonnegative real value, including zero. Within a given time period, most insurance policyholders don't file a claim. Policyholders that do file claims could have small claims, medium-sized claims, or large claims. $Tw_p(\mu, \sigma^2)$ random variables with $p \in (1, 2)$ and large σ^2 describe the empirical data well. We use ggplot2 in R to simulate some compound Poisson-gamma random variables. As you look over these graphs, treat the x-axis values as dollar values and consider the costs an insurer covers for its policyholders in a given time period.

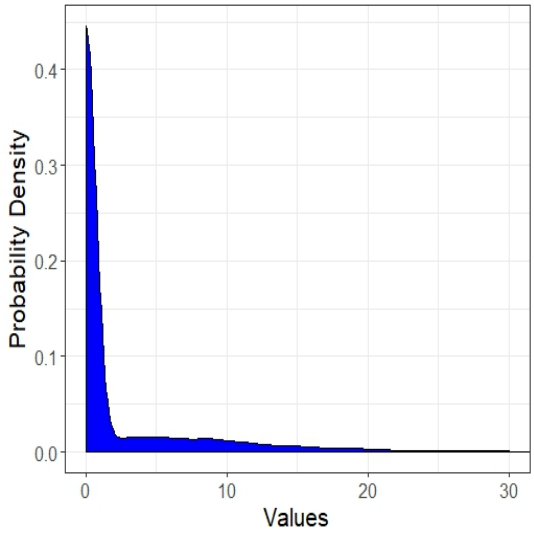

Figure 5: 10,000 Simulated $Tw_{1.33}(2, 10)$ Random Variables

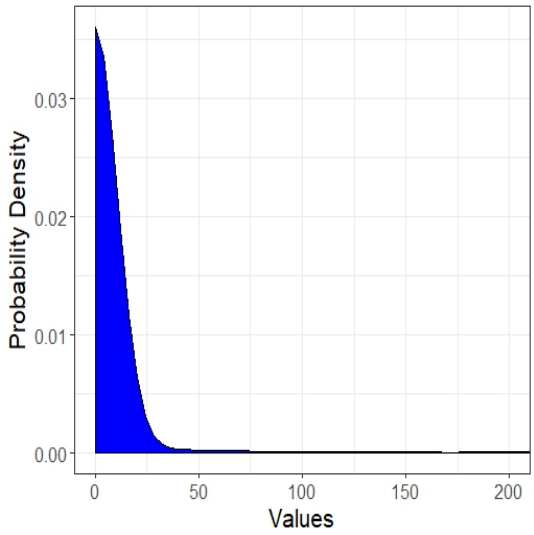

Figure 6: 10,000 Simulated $Tw_{1.66}(10, 100)$ Random Variables

Most of the realizations take value 0 or values close to 0, but some realizations take large positive values. This behavior is exactly what we expect from claims data: mostly no claims, but a few large losses. Compound Poisson-gamma models do a good job describing data that contain both zeros and continuous positive values: daily rainfall in a month, fish caught in an hour, defunct lightbulbs in a large building in a week, et cetera.

Viewing Poisson-gamma sums as Tweedie models does not provide us with any intuition about how these distributions arise. To gain some intuition, we will introduce renewal processes. Renewal theory is a topic in stochastic processes. A stochastic (random) process is a sequence of random variables. Sidney Resnick explains that "[r]enewal processes model occurrences of events happening at random times where the times between the events can be approximated by independent, identically distributed random variables" [20]. Much of this section follows the writings of Resnick and Sheldon Ross [20, 22]. To define a renewal process, we must equip ourselves with some definitions.

Definition 1.7. Let $\{X_k, k \geq 1\}$ be a sequence of independent random variables that take only nonnegative values. We call the random variables X_k *interarrival times*. The cumula-

tive distribution $F(x)$ of a given X_k is $Pr(X_k < x)$. We say that the interarrival distribution F is *proper* if $F(\infty) = 1$. (We assume proper distributions in this section.)

Definition 1.8. Let $S_n = X_1 + \cdots + X_k + \cdots + X_n$. We call the sequence $\{S_n, n \geq 1\}$ a *renewal sequence*, or, equivalently, a *renewal process*.

Note. Think of a random variable X_k as the time until some event happens and the renewal process S_n as the time until n such events happen in sequence.

Renewal processes involve time passing and random events happening. We would like to count how many of these random events happen. A *counting process* $N(t)$ maintains the following properties:

(i.) $N(t) \in \mathbb{N}$;

(ii.) $N(s) \leq N(t)$ if $s \leq t$.

Ross presents two definitions for the counting process generated from a renewal sequence. We will use both definitions interchangeably to make proofs clearer. First, let the counting process $N(t) = \sum_{n=0}^{\infty} I_{[0,t]}(S_n)$ where the indicator random variable $I_{[0,t]}$ means

$$I_{[0,t]}(S_n) = \begin{cases} 1 & S_n \leq t \\ \\ 0 & \text{otherwise} \end{cases} .$$

This notation with the indicator random variable stresses the act of counting when renewals occur. Equivalently, let the counting process $N(t) = \max\{n : S_n \leq t\}$. This notation illustrates finding the maximum of the events observed prior to time t. Now, we exercise our expanded toolkit to prove a proposition from renewal theory.

Proposition 1.5. With probability 1, $N(\infty) = \infty$. That is, almost surely, infinite events occur in infinite time.

Proof.

$$\begin{aligned} Pr(N(\infty) < \infty) &= Pr\left(\bigcup_{n=1}^{\infty} \{X_n = \infty\} \right) \\ &\leq \sum_{n=1}^{\infty} Pr(X_n = \infty) \\ &= 0. \end{aligned}$$

We justify the above argument. First, the probability of the arbitrary union expresses the probability that some interarrival time never arrives. This is logically equivalent to saying that the counting process is finite. Second, the inequality is Boole's Inequality [21]. Third, $Pr(X_n = \infty) = 0$ because X_n is a proper interarrival time. That is, $Pr(X_n < \infty) = 1$.

Now, use the complement rule. We get that $Pr(N(\infty) = \infty) = 1$. □

If we consider some random variable Y_k happening at each renewal, we can find many applications of counting processes. For instance, suppose that at each renewal there is a random positive cost that accrues. This example fits well to insurance claims. Interarrival times correspond to when claims occur and the matching Y_k correspond to the cost of the claim. Compound Poisson-gamma processes are renewal processes where the Y_k correspond to gamma random variables and the counting process $N(t)$ corresponds to a Poisson random variable. (A Poisson point process $N(t)$ is constructed from exponential interarrival times [7].) Before we prove a result for these compound processes, we define two things.

Definition 1.9. Let X_n be a random variable for all $n \in \mathbb{N}$. A *compound point process* is $\sum_{n=1}^{N(t)} X_n$ where $N(t)$ is a counting process.

Definition 1.10. We say that T is a *stopping time* if the occurrence of an event can be determined by looking at values of a process up to that time.

Example 1.3. $\{N(t) = n\}$ is a stopping time because it is independent of interarrival times $X_k,\ k \geq n + 1$.

Theorem 1.3 (Wald's Equation)**.** Suppose that the interarrival times are i.i.d. random variables with finite expectation and that $N(t)$ is a stopping time with finite expectation. Then,

$$\mathrm{E}\left[\sum_{n=1}^{N(t)} X_n\right] = \mathrm{E}[X] \cdot \mathrm{E}[N(t)].$$

Proof. Let the indicator variable I_n be 1 if $N(t) \geq n$ and 0 otherwise. It follows that $\sum_{n=1}^{N(t)} X_n = \sum_{n=1}^{\infty} X_n I_n$. I_n is independent of X_n because I_n is completely determined by $N(t)$ not stopping until after the observation of $X_1, \cdots, X_{n-1}$. Thus,

$$\mathrm{E}\left[\sum_{n=1}^{N} X_n\right] = \mathrm{E}\left[\sum_{n=1}^{\infty} X_n I_n\right] = \sum_{n=1}^{\infty} \mathrm{E}[X_n I_n] = \sum_{n=1}^{\infty} \mathrm{E}[X_n] \cdot \mathrm{E}[I_n].$$

Now, consider the following algebra:

$$\begin{aligned}
\sum_{n=1}^{\infty} \mathrm{E}[X_n] \cdot \mathrm{E}[I_n] &= \mathrm{E}[X] \sum_{n=1}^{\infty} \mathrm{E}[I_n] \\
&= \mathrm{E}[X] \sum_{n=1}^{\infty} Pr(N(t) \geq n) \\
&= \mathrm{E}[X] \cdot \mathrm{E}[N(t)] \ .
\end{aligned}$$

What requires further justification in this lengthy sequence of equality statements is that $\sum_{n=1}^{\infty} Pr(N(t) \geq n) = \mathrm{E}[N(t)]$. Recall that $N(t)$ is a discrete random variable. We can formulate $Pr(N(t) \geq n)$ as $Pr(N(t) = n) + Pr(N(t) = n+1) + Pr(N(t) = n+2) + \cdots$. Thus, the summation is

$$\begin{aligned}\sum_{n=1}^{\infty} Pr(N(t) \geq n) &= Pr(N(t)=1) + Pr(N(t)=2) + Pr(N(t)=3) \cdots \\ &\quad + Pr(N(t)=2) + Pr(N(t)=3) + \cdots \\ &\quad + Pr(N(t)=3) + \cdots \\ &= \sum_{n=1}^{\infty} nPr(N(t)=n) \\ &= \mathrm{E}[N(t)]. \end{aligned}$$ □

Remark. This proof assumes that $\mathrm{E}[N(t)]$ is finite. We justify this in the Appendix.

Example 1.4. Let $X_i \sim Ga(\alpha, \beta)$ and let $N(t)$ be a Poisson random variable with mean λt. Use Wald's equation to find the expectation.

$$\mathrm{E}\left[\sum_{i=1}^{N(t)} X_i\right] = \mathrm{E}[X_i] \cdot \mathrm{E}[N(t)] = \frac{\alpha\lambda t}{\beta}.$$

Regression Analysis and Generalized Linear Models

Psychologist and winner of a Nobel Prize in Economics Daniel Kahnemann eloquently explains regression to the layperson in his 2011 book ***Thinking Fast and Slow***. He describes the phenomenon of "regression to the mean" as when a golfer performs poorly the day after a stellar performance and vice versa [16]. In statistical modeling, regression analysis means using algorithms to estimate the relationship between an average output and other variables. With the widespread availability of data in the early 21st century, regression analysis has become a popular form of machine learning in many scientific fields. Actuaries are just some of the many professionals who rely on regression models to do their routine work.

We introduce regression models with graphs. For demonstration purposes, we

picked three sets of 25 numbers ranging from 0 to 100. Note that the linear models have no predictive value. We handpicked vectors to demonstrate positive and negative correlations.

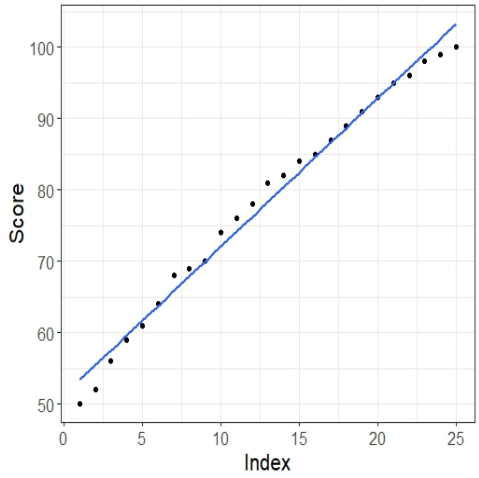

Figure 7: Best Fit Line for Positively Correlated Data

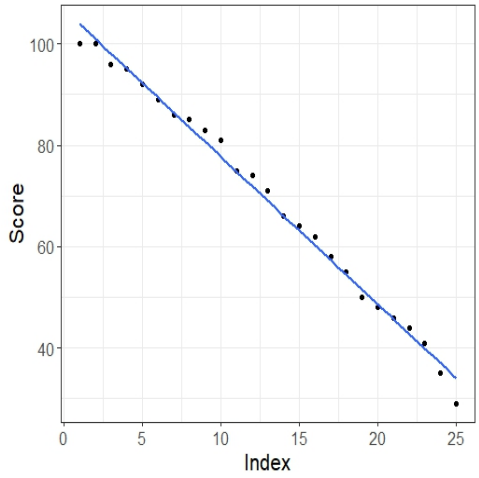

Figure 8: Best Fit Line for Negatively Correlated Data

Figures 7 and 8 should remind you of your high school statistics courses. The model draws a line that minimizes the total distance between the cluster of points. Statistical programs like R, SAS, STATA, and Python offer functions to plot such relationships. For Figure 9, we use the default method for geom_smooth. This mapping performs locally-weighted scatterplot smoothing. We do not discuss local regression

here. We point it out to say that there are many regression models in use, though many people are only familiar with linear regression.

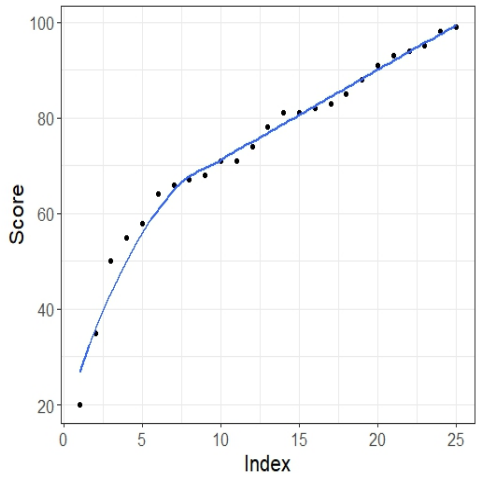

Figure 9: Locally-Weighted Regression Curve for Positively Correlated Data

Researchers in the social sciences, the biologicial sciences, the financial sector, and many other domains use regression models. The latter half of this thesis estimates the Tweedie index parameter and evaluates the estimator's performance in the context of generalized linear models (GLMs). We introduce the framework of regression analysis and its measurements of accuracy so that you can make sense of these case studies.

Definition 1.11. Let Y be a $n \times 1$ vector. Each row in the vector Y contains a random variable y_i. Let X be a $n \times p$ matrix. Each column vector refers $x_j, 0 \leq j \leq p-1$, contains observations for some explanatory variable. We call X the design matrix or the matrix of observations. Let β be a $p \times 1$ vector with the β_j unknown weight parameters, and let ε be a $n \times 1$ vector of unknown error terms. Consider some link function g that expresses the relationship

$$g(Y) = X\beta + \varepsilon \, ,$$

or, equivalently,

$$g(Y) = \beta_0 x_0 + \cdots + \beta_p x_p + \varepsilon \, .$$

This relationship is a *generalized linear model*.

In supervised learning, we know the link function g, the target values y_i, and the design matrix X. Although we have the realized values for each y_i, the y_i are still random variables with mean parameter μ_i. The regression problem involves solving the equation

$$g(\mu_i) = \beta_0 x_{i,0} + \cdots + \beta_p x_{i,p}, \ \forall 1 \leq i \leq n \ .$$

Actuaries often use a log link function because it gives a multiplicative rating structure. That is,

$$\log \mu_i = \beta_0 x_{i,0} + \cdots + \beta_p x_{i,p} \ .$$

Notice the transformation

$$\mu_i = e^{\beta_0 x_{i,0}} \times \cdots \times e^{\beta_p x_{i,p}} \ .$$

The first explanatory variable x_0 is often an intercept term; in other words, x_0 is a $n \times 1$ column vector $(1, \cdots, 1)'$. Therefore, e^{β_0} acts as a starting rate. It is then modified by multiplication. One reason actuaries use this structure is because multiplication is easy to explain to customers and regulators.

Besides the link function g, the statistical modeler must also specify what distribution explains the random variables y_i. When Nelder and Wedderburn developed the software that enables modern-day statistical modeling, they established exponential dispersion models as response distributions. Gordon Smyth and Peter Dunn have since written code in R to implement the whole spectrum of Tweedie models in as response distributions for generalized linear models. This history explains one reason why we care so much about EDMs and the Tweedie family: their ubiquitous use in real-world modeling projects.

The objective of regression analysis is to determine the weights β_j that minimize the residual error. Maximum likelihood estimation (MLE) is one popular method to execute this task. MLE finds the β_j parameter values that maximize the likelihood that the observations Y occur, given the observations of the design matrix X. Once we have β_j weights, we can predict the target variables y_i based on the weights and the matrix of observations. We expect our model to imperfectly predict the values y_i. Hence, we record measurements that tell us how well our model fits the actual data.

Definition 1.12. Let Y be a random variable parameterized by θ with density function f. Given a vector of observations $(y_0, \cdots, y_n)$, the *log-likelihood* is $\log\Big(\prod_i \ f(y_i \,|\, \theta)\Big)$.

Definition 1.13. Suppose we have a generalized linear model A that uses some explanatory variables. Define the saturated model S to be the model that uses as many explanatory variables as there are data points. *Deviance* for the model A is

$$2 \cdot (ll_S - ll_A)$$

where ll stands for log-likelihood. Equivalently, the deviance is

$$2 \cdot \sum_{i=1}^{n} \left[\log\Big(f(y_i \,|\, \mu_i = y_i)\Big) - \log\Big(f(y_i \,|\, \mu_i = \mu_i)\Big) \right].$$

Deviance is one measurement that we use to evaluate the performance of a model. However, using only deviance as a measurement of model performance leads to problems. A model always fits better to the training dataset with more explanatory variables. But, the purpose of predictive analytics is to forecast results for a dataset different from the training dataset. We worry that we might overfit to the training data if we use too many explanatory variables in the model. As a result, we also record a measurement of model accuracy that penalizes the model for using too many explanatory variables.

Definition 1.14. The *Akaike Information Criterion* (AIC) is

$$-2 \cdot ll_A + 2 \cdot p \,,$$

where p is the number of explanatory variables used in model A.

With these metrics in our toolkit, a working definition of generalized linear models, and background knowledge of Tweedie models, we are almost ready to estimate the Tweedie power. The next chapter discusses a major obstacle we face in estimating the Tweedie power and proposed solutions.

Endnotes

- Tweedie random variables have a unit variance function. According to Definitions 1.4 and 1.5, the unit deviances of Tweedie random variables must be continuous and differentiable.

- Given an exponential dispersion model, are the $a(y;\sigma^2)$ function and the unit deviance $d(y;\mu)$ unique? Of course, we can make simple linear changes like adding or subtracting a constant, but these changes to the a and d functions are trivial. The question at hand is whether the structure of the a and d functions are special and specific to the random variable. Future research could evaluate this question of uniqueness.

- Exponential dispersion models provide a framework to describe random variables. We could propose functions $a(y;\sigma^2)$ and $d(y;\mu)$ as a method to generate new probability distributions. Would probability distributions generated from such an approach have any practical use? Do they occur naturally in empirical distributions of raw data? Future research could address this question of existence.

- Locally-weighted scatterplot smoothing combines the ideas of k-nearest neighbors modeling and multiple regression modeling. You can find statistical literature about the technique (and about k-nearest neighbors modeling) in libraries or on the Internet.

- Only use deviance to compare models if the models are nested [11]. In this situation, performing an analysis of variance (ANOVA) is a viable option.

- You can use AIC as a metric to compare models that are not nested. Modelers use multiple metrics to assess the performance of a model. It makes sense

to analyze both deviance and AIC. Deviance analysis helps you find the most probable model and AIC analysis addresses the concern of overfitting.

- Bayesian Information Criterion (BIC) is another common penalized measure of fit. AICc is a form of AIC that corrects for small sample sizes.
- This idea of a penalty term comes from information theory.
- Uniform, negative binomial, hypergeometric, and binomial random variables are examples of non-Tweedie random variables. The figure below shows 10 million random draws of a uniform random variable. Contrast this distribution to Tweedie distributions.

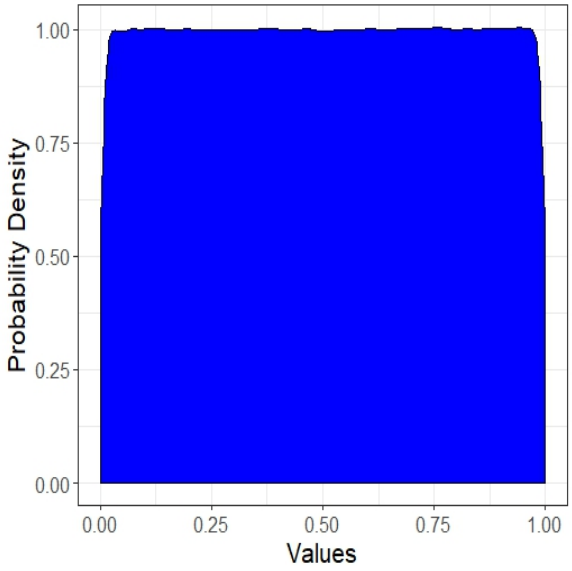

Figure 10: 10 Million Simulated Uniform Random Variables

2 Numerically Approximating Tweedie Densities

Maximum likelihood estimation is a popular way to estimate parameters for a distribution. To perform MLE, we require a closed form for the density function. Unfortunately, most Tweedie models do not have closed form densities. We have been writing $f(y; \mu, \sigma^2)$ as a succinct expression, but the $a(y; \sigma^2)$ part hides much of the baffling behavior of the Tweedie model. The $a(y; \sigma^2)$ functions have a lot of freedom to take on many forms. Besides some special cases, the $a(y; \sigma^2)$ functions are complicated series objects that can't be written down in a closed form. As a result, we must approximate the Tweedie densities before we do MLE.

Peter Dunn and Gordon Smyth propose algorithms to numerically approximate the Tweedie densities [4, 5, 6]. Peter Dunn created and maintains the tweedie R package [3]; Gordon Smyth contributed to and maintains the statmod R package [25]. Because I coded in R to do the applied work in Chapters 3, 4, and 5, Dunn and Smyth's collaborative research on numerical approximation methods is important to this paper. We discuss their methods in the proceeding subsections.

Fourier Inversion

Random variables are defined by their characteristic functions. Moreover, if a random variable has a density function, the characteristic function is the Fourier transform of its density. Dunn and Smyth use these well-known facts to determine the Tweedie densities.

Definition 2.1. Let X be a random variable. Its *characteristic function* is $\mathrm{E}[e^{itX}]$, where i is the imaginary unit.

An integral transform maps a function from its original domain to another domain where it is easier to evaluate the problem. Two popular integral transforms are the Laplace transform and the Fourier transform. (The Laplace transform is widely

applied in electrical engineering.) STEM students should be familiar with Fourier series. Math students see them in linear algebra courses, and physics students use them to deconstruct deconstruct a wave into a linear combination of sine and cosine waves. The Fourier transform draws inspiration from this idea of superposition. It expresses a continuous superposition of functions whereas a Fourier series expresses discrete superpositions.

Definition 2.2. Let $f(x)$ be some function. The *Fourier transform* of f is

$$\mathscr{F}(t) = \frac{1}{2\pi} \int_{\mathbb{R}} e^{-itx} f(x) dx \ .$$

The *Fourier inversion* of $\mathscr{F}(t)$ is

$$f(x) = \frac{1}{2\pi} \int_{\mathbb{R}} e^{itx} \mathscr{F}(t) dt \ .$$

Note. We present the Fourier transform and its inverse with the notation you see in many math texts, on Wikipedia, and on Wolfram MathWorld. However, Dunn and Smyth and other probabilists sometimes switch the i and $-i$. What is important to remember is that we have an integral transform and its inverse. Do not worry too much about the nomenclature.

If a random variable X has a density function $f(x)$, then the characteristic function is the Fourier transform of f. That is,

$$\mathrm{E}[e^{itX}] = \int_{\mathbb{R}} e^{itx} f(x) dx \ .$$

You may object that this format differs from our definition. We need only change the parameter t. Let $t = -2\pi s$ and manipulate the integral with u-substitution.

Likewise, we can find the density if we have the characteristic function. We evaluate the Fourier inversion of the characteristic function to find the density (if it exists). Besides Poisson random variables, Tweedie random variables have probability densities. However, their densities include $a(y; \sigma^2)$ functions that, in general, have

no closed form. One way that Dunn and Smyth implement Tweedie densities in R's tweedie package is by approximating the Fourier inversion of a model's characteristic function [4].

Recall that the density for a Tweedie model is

$$f(y;\mu,\sigma^2) = a(y;\sigma^2)\exp\Big\{-\frac{1}{2\sigma^2}d(y;\mu)\Big\}\,.$$

Dunn and Smyth set μ equal to 1, and they scale y to 1. Notice that y and μ are 1, so $d(y;\mu) = 0$. Thus, $a(1;\sigma^2/y^{2-p}) = f(1;1,\sigma^2/y^{2-p})$. We can get $a(1;\sigma^2/y^{2-p})$ by computing the Fourier inversion of the characteristic function. In the Appendix, we justify that $a(y;\sigma^2) = f(y;y,\sigma^2)$. Because Tweedie models are closed under scaling, we rescale by y to get $a(y;\sigma^2)$. Substitute this $a(y;\sigma^2)$ into the density and compute given μ. (Dunn and Smyth actually propose three different methods to get $a(y;\sigma^2)$ by Fourier inversion [4].)

First, I recognize that this explanation is unsatisfactory. I encourage readers to review the Appendix and read Dunn and Smyth's papers to understand the full picture. Ultimately, their method involves manipulating both the axiomatic and constructive definitions of the Tweedie densities. Moreover, the characteristic function they integrate involves a cumulant-generating function. As this thesis targets a wider audience, I have resigned the details to the Appendix. Essentially, Dunn and Smyth use the Fourier inversion and the rescaling property of Tweedie models to determine $a(y;\sigma^2)$. It is easy to compute the other part of the density.

On another note, it is difficult to compute the Fourier inversion of the characteristic function. This task boils down to evaluating a highly oscillatory integral. We provide figures to illustrate how difficult this computation can be. Figure 11 shows a damped cosine wave that oscillates slowly; Figure 12 shows a damped cosine wave that oscillates faster; Figure 13 shows a damped cosine wave that oscillates very quickly.

Adding up the areas under the curves for Figures 11 and 12 appears feasible. But, for highly oscillatory integrands like that in Figure 13, we must rely on a numerical approximation.

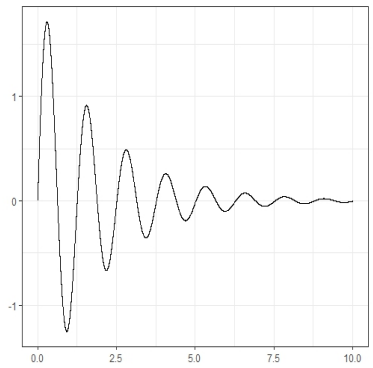

Figure 11: Simple Damped Cosine Wave

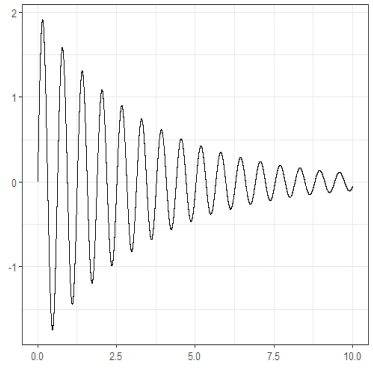

Figure 12: More Complex Damped Cosine Wave

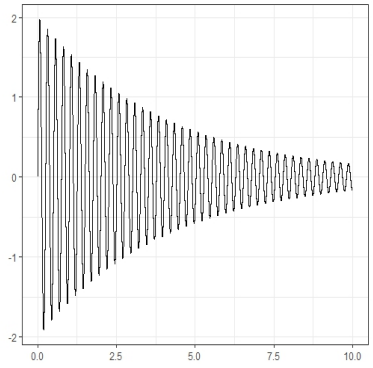

Figure 13: Highly Oscillatory Damped Cosine Wave

Dunn and Smyth modify an algorithm proposed by Sidi to approximate highly oscillatory integrals [4]. First, the algorithm finds exact zeros where the curve crosses the x-axis. Second, it defines linear equations, where the linear equations equal the integral from 0 to the most recent root plus the area under the curve up to the next root. Next, the algorithm solves the linear equations. That solution is an approximation for the integral. This approximation increases in accuracy the more linear equations we define because we incorporate in more areas under the curve. Dunn and Smyth's algorithm terminates when the relative error between performing one more iteration is less than 10^{-10}.

Another nontrivial part of this algorithm is how we compute the areas under the curve. The integrands are not as simple to solve as the integrands we see in college calculus courses. The algorithm uses 512-point Gaussian quadrature to compute the areas under the curve. Recall Riemann sums and the trapezoidal method from your college calculus course. These methods find points and weights, and then evaluate a weighted average. Conveniently, the points and weights correspond to the geometric idea of summing up rectangular/trapezoidal areas. Gaussian quadrature uses this same approach, but the choice of points and weights is more ingenious.

Depending on y, σ^2, and p, this algorithm could be computationally intensive or inaccurate. For R's tweedie package, this method of approximating a Fourier inversion is the default way to get the Tweedie densities. Other methods perform better for certain combinations of y, σ^2, and p.

Infinite Series Evaluation

One such method involves approximating an infinite series expansion [5]. We can write the $a(y;\sigma^2)$ function as a series expansion [5, 15]. For a Poisson-gamma model,

$$a(y;\sigma^2) = \frac{1}{y} W(y;\sigma^2, p),$$

where $W(y;\sigma^2,p) = \sum_{j=1}^{\infty} W_j$. For $p > 2$ and $y > 0$,

$$a(y;\sigma^2) = \frac{1}{\pi y} V(y;\sigma^2,p)$$

with $V(y;\sigma^2,p) = \sum_{k=1}^{\infty} V_k$.

Both W_j and V_k are complicated fractions with many parameters, exponents, factorials, and gamma functions. (See Dunn and Smyth's paper [5] for the exact expression.) Nevertheless, we want to evaluate W and V. Dunn and Smyth determine the indices j and k for which W_j and V_k reach maximums. They do this task by treating j and k as continuous, differentiating with respect to them, and setting the derivatives to zero. In other words, they find the maximum in the usual way taught in college calculus. Next, Dunn and Smyth find upper and lower limits for j and k around $j_{\max}$ and $k_{\max}$. These bounds are found computationally by finding W_j and V_k sufficiently small relative to $W_{\max}$ and $V_{\max}$. In the end, their algorithm sums a finite number of W_j and V_k to serve as an approximation for W and V.

This series approach aims to add up only terms in the series that contribute significantly. We find the term that contributes the most, and we include it and its neighbors in the computation. Like the Fourier inversion approach, this algorithm could be computationally intensive or inaccurate depending on the parameters y, σ^2, and p.

Saddlepoint Approximation

The third way Dunn and Smyth approximate $a(y;\sigma^2)$ is by the saddlepoint approximation. This technique is well-known in the statistical community. You can find a more thorough treatment of the saddlepoint approximation in the mathematical literature. In any event, the saddlepoint approximation gives that $a(y;\sigma^2) \approx \sqrt{2\pi\sigma^2 y^p}$, where p is the Tweedie power [6]. For my projects, I do not use the saddlepoint

method to approximate the Tweedie densities.

Comparing Methods

The series and inversion methods perform poorly in different parameter spaces [4, 5, 6]. Dunn and Smyth's series approach requires extensive computing to reach an accurate approximation in the following cases:

- p close to 2, y large, or σ^2 small for Poisson-gamma models;
- p close to 2, y small, or σ^2 small for models with $p > 2$;
- y small for inverse-Gaussian (p=3) models.

Meanwhile, the inversion approach performs well for p close to 2. It runs slow for $p > 3$ and y large. Do not try to memorize which methods perform best for which parameters.

We can use if-else control structures so that our algorithm uses the most efficient and accurate method for the given parameters. In R, dtweedie does this automatically, and tweedie.profile does this if you set the method parameter as "interpolation". I experimented with methods "series", "inversion", and "interpolation" and found that I get quicker and more reliable results with "interpolation" as my method. We will use these methods in Chapters 4 and 5 to approximate Tweedie densities and estimate the Tweedie power.

Endnotes

- Due to my limited exposure to programming languages, I only considered using R and Python for my applied projects. They are the most commonly used programming languages for data science. I am aware of SAS having some ability to make Tweedie models. I don't know if there are any other programming languages that enable Tweedie modeling.

- I briefly looked into using Python. From this research, I know that Python can perform Tweedie modeling. However, I don't consider Python's implementation to be as functional or well-documented as R's tweedie package. I determined that R would best suit my purposes. After all, Dunn and Smyth have written extensively on at least three different numerical approximation methods, and the tweedie R package provides four different numerical approximation methods.

- Dunn uses Fortran to speed up the numerical computations in the inversion algorithm. A good software development project could be to translate Dunn's tweedie package to Python. This project would likely require the programmer to be familiar with Fortran.

3 Estimating the Power Parameter

The Tweedie power parameter acts an an index, telling us the distribution of our model. This feature is why MCK Tweedie named his paper "An Index which Distinguishes between Some Important Exponential Families." We know that the index encodes important information. Choose $p = 1$, and we get the discrete Poisson distribution. Select $p = 3$, and we get the sharp-peaked, wide-tailed continuous inverse-Gaussian distribution. There is a clear difference between Tweedie random variables with $1 < p < 2$ and $p \geq 2$. What I want to look into is the differences between Tweedie random variables in the same subclass. That is, how different are different compound Poisson-gamma models? Can we improve a loss cost model by estimating $p \in (1, 2)$? How different are gamma random variables to inverse-Gaussian random variables? Can we improve a severity model by estimating $p \in (2, 3)$? I hypothesize that we can improve our insurance models by estimating the power parameter.

We estimate the power parameter by computing log-likelihoods [3]. The technique used in the tweedie.profile function does the following:

1. Take as input a vector of data and a list of possible power parameters.
2. Given the vector data, compute the log-likelihood of the vector for each power parameter.
3. Return the power parameter that maximizes the log-likelihood.
4. (Optional) Plot the log-likelihoods on a graph.
5. (Optional) Draw a smooth curve that passes through the points on the graph.
6. (Optional) Return the power that maximizes log-likelihood on the curve.

Before I begin modeling with this new tool, I test how well it performs on simulated data. Using rtweedie, rinvgauss, and rgamma, I generate realizations of random

variables. For instance, I create 10,000 realizations of a $Tw_{1.33}$ model, 10,000 realizations of a $Tw_{1.5}$ model, and 10,000 realizations of a $Tw_{1.75}$ model. tweedie.profile accurately estimates the power parameter within two decimal places and the dispersion parameter within single digits. Moreover, rtweedie and tweedie.profile execute in less than a minute. Likewise, I receive promising results in estimating a gamma random variable and an inverse-Gaussian random variable. (I recommend you use rinvgauss in the statmod package to generate these realizations. rtweedie executes slowly for $p > 2$.)

I struggle to continue these tests with $p > 3$. rtweedie and tweedie.profile execute slowly for large p. I don't worry too much about this behavior. Applications to loss cost and severity modeling will focus on $1 < p < 2$ and $2 \leq p \leq 3$. Moreover, it is difficult to describe Tweedie models for $p > 3$. At this time, we recognize them only for their theoretical purposes. All in all, this approach to estimating the Tweedie power performs well for the parameter spaces we care about.

I practiced using these packages by modeling with data from the insuranceData R package [31]. I built severity models and loss cost models. In both cases, tweedie.profile returned Tweedie power estimates that fell into the range I suspected they would. These results give me additional confidence that tweedie.profile estimates the power parameter well.

We can get a good maximum likelihood estimate for the Tweedie power, but that doesn't mean that the estimated Tweedie power will provide significant improvements to our insurance models. Models are built with many features. The Tweedie power parameter is one feature out of many. In the next two chapters, I optimize the Tweedie power for two insurance case studies. I design my experiments as follows:

1. Split the data into 75 percent training data and 25 percent test data.

2. Build a Tweedie model using the training data. I look at deviance change, the significance of a variable, and AIC, and then I arbitrate on if I include the

variable in the model. For the most part, I include a variable if it decreases the AIC.

3. Ceteris paribus, change the Tweedie power from a default value to an estimated value. Severity models use default value $p = 2$ and loss cost models use default values $p \in (1.33, 1.5, 1.66)$.

4. See how the models perform on the test data.

5. (Optional) Use ggplot2 to visualize the results.

6. (Optional) Consider other metrics to examine model performance.

This design highlights the Tweedie power's effect as a model parameter. Of course, this step in model production takes time and energy. As a result, I will also comment on the economic value this parameter tuning brings to an insurance company.

Endnotes

- David A. Freedman discusses maximum likelihood estimation in depth in his book ***Statistical Models: Theory and Practice***.

- R's glm object performs the well-documented iteratively re-weighted least squares algorithm from Nelder and Wedderburn.

- I don't try to build a perfect model. The models I build are on outdated data and will never be put in production. I try to build decent models that include some predictive features. Ultimately, the aim of this modeling is to perform a scientific experiment, not to provide an insurance product.

- Be cautious when using the do.smooth hyper parameter in the tweedie.profile function. When smoothing is on, the list of Tweedie powers you pass to the function can affect the estimate. When smoothing is off, the Tweedie power in the input list that maximizes log-likelihood becomes your estimate. If the spacing between Tweedie powers is large, you can get a poor estimate.

4 Case Study: Automobile Bodily Injury Claims

This chapter showcases the applied work I did to model severity with an optimized Tweedie power. For severity datasets, actuaries usually use a gamma or inverse-Gaussian distribution to describe the target variable's distribution. Gamma models correspond to a Tweedie power of 2 and inverse-Gaussian models correspond to a Tweedie power of 3. I want to evaluate if we can improve the likelihood of a severity model by estimating the true Tweedie power.

Our severity data was collected by the Insurance Research Council in 2002. We access the data from the R package insuranceData. The dataset is titled AutoBi, standing for Automobile Bodily Injury [31]. Bodily injury coverage pays for the costs to people injured in a car accident. Such costs may involve medical expenses, legal fees, funeral costs, and loss of income. Given that medical care and legal services cost a lot in America, you can imagine that bodily injury claims have high payouts. Besides the column data on claim costs, AutoBi contains six other categorical variables. These variables explain if the claimant had an attorney represent him or her, if the claimant wore a seatbelt, if the claimant had insurance, the age of the claimant, the sex of the claimant, and the marital status of the claimant. In total, the dataset contains 1340 observations.

The model we build uses AGE, ATTORNEY, and SEATBELT as explanatory variables. ATTORNEY is a binary variable, SEATBELT has categories Yes, No, Not Applicable, and AGE is split roughly into 10-20 year age groups. Only these three variables decreased the Akaike Information Criterion. We expect ATTORNEY to be a strongly predictive covariate in my model because it captures the costs associated with legal representation. Using a seatbelt should decrease the risk of death and serious injury, so it makes sense why the SEATBELT is predictive. I am less confident putting AGE in our model. It provides less significance and only a marginal decrease in AIC. Nevertheless, we want to have at least a few covariates in the model to differentiate

between policyholders. Modeling with only three explanatory variables is not ideal. This paucity of variables fails to capture enough signal about the target. I would want to model with more explanatory variables for actual insurance products.

In any event, this study is meant to assess the utility of tuning the Tweedie power. Below we display the profile log-likelihood plot made to estimate the Tweedie power. According to this technique, the Tweedie power is most likely 2.3 and the dispersion parameter is most likely 1.110. We make a Tweedie model with these parameter choices and a gamma model based on 75 percent training data.

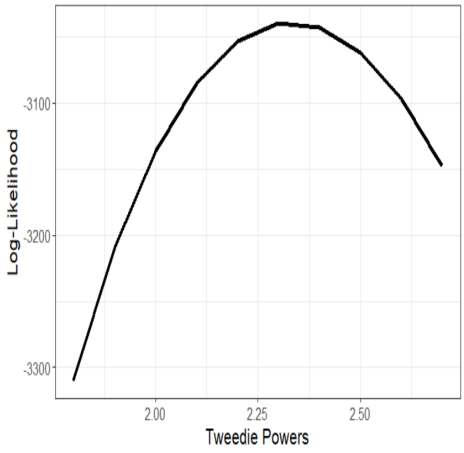

Figure 14: Profile Log-likelihood Plot for AutoBi Data

We compare the two models. The Akaike Information Criterion for the Tweedie model evaluates to 4499.53, whereas it evaluates to 4611.25 for the gamma model. These results say that the Tweedie model fits better to the data than the gamma model. We underline that AIC talks about the quality of a model relative to another model. It is possible that both models fit the data poorly.

Based on the AIC metrics, we conclude that we get a better model when we estimate the Tweedie power. However, we want to quantify this improvement. Table 2 presents the weights for the two models. Observe that the coefficients change with order of magnitude 10^{-2}. Table 3 records predictions for the first 5 cases in the training set. We see differences in the tens and hundreds of dollars.

Explanatory Variable	Tweedie Model	Gamma Model
Intercept	1.8699	1.8382
ATTORNEY	-1.4757	-1.4772
SEATBELT 1	0.4490	0.4887
SEATBELT 2	1.3231	1.4061
AGE 1	-1.4032	-1.3988
AGE 2	-0.6246	-0.6544
AGE 3	-0.1750	-0.1814
AGE 4	-0.0666	-0.0669
AGE 5	-0.6338	-0.6296

Table 2: Comparison of linear weights between the two AutoBi severity models.

Policyholder Index	Actual Loss	Tweedie Model	Gamma Model
7	3538	5443.06	5325.41
9	874	1232.93	1257.50
22	230	1955.63	1951.10
23	26262	8554.29	8546.74
33	603	1244.36	1215.72

Table 3: Loss predictions for first 5 rows in AutoBi testing data.

The training set accounts for about 5.1 million dollars in losses. We calculate that the Tweedie model predicts about 100,000 dollars less in losses than the gamma model. I credit this difference to how the Tweedie model situates itself in between the gamma and inverse-Gaussian cases. Inverse-Gaussian models have smaller peaks and wider tails than gamma models. Because most of the claims reside in the peak area, I suspect the smaller peak in the Tweedie model explains the discount. This information suggests that estimating the Tweedie power could improve accuracy by approximately 2 percent.

We receive less encouraging results when we look at the testing set. These losses account for about 2.8 million dollars in loss. We predict losses based on the two models and observe that the difference in predicted losses only varies by about two thousand dollars. This computation indicates that the trained model doesn't translate over as well to the testing data. Moreover, both models fail to predict a huge amount of the total losses. The ratio of predicted aggregate losses over actual aggregate losses is

approximately 57 percent. If we had access to more explanatory variables, I reckon we could enhance my models and fix this issue.

Next, we plot the model predictions and the target's empirical distribution. Notably, we see multiple humps and thin tails in Figures 15 and 16, whereas the actual distribution has one hump and a thicker tail. Neither model predicts the large losses very well. Additionally, we see little difference between Figures 15 and 16. The shapes of the two distributions look almost identical. This observation calls into question how different two Tweedie models are when they both fall into the same subclass.

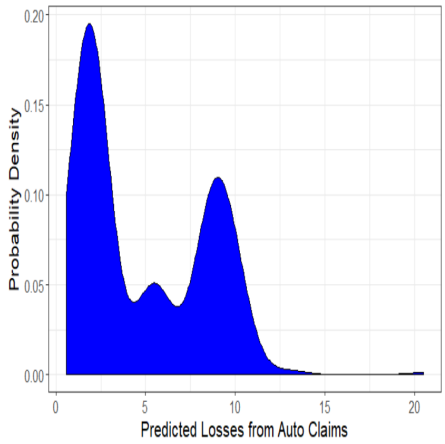

Figure 15: Distribution of $Tw_{2.3}$ Severity Model

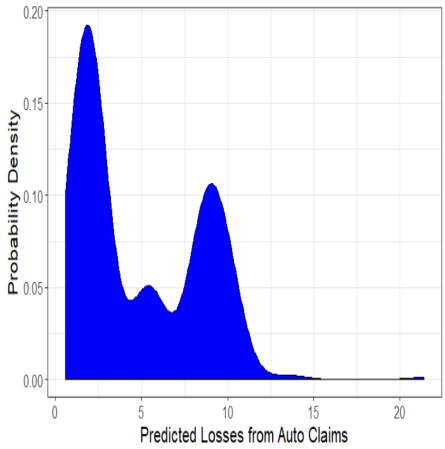

Figure 16: Distribution of Gamma Severity Model

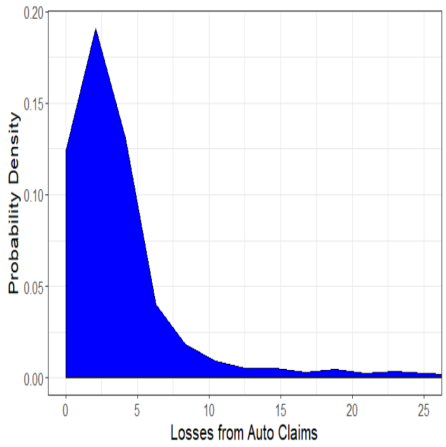

Figure 17: Distribution of AutoBi Losses

This experiment considered severity modeling with a Tweedie power $p \in (2, 3)$. Actuaries normally encounter the name Tweedie in terms of compound Poisson-gamma random variables and loss cost modeling. In this sense, this case study provides some novelty. We made a Tweedie model for claims severity, and we argued its advantages over a gamma model. Unfortunately, our parameter tuning of the Tweedie power did little to change the model predictions.

5 Case Study: Liberty Mutual Home Fire Perils

The first case study concerned automobile claims and severity data. Now, we consider home insurance claims and loss cost data. We did a poor job modeling automobile bodily injury claims, in large part because we didn't have many explanatory variables available. This next case study involves more policyholders and more explanatory variables. Liberty Mutual provided me with a large dataset concerning fire coverage home insurance policies. The data contains 300 explanatory variables and 450,065 observations. The explanatory variables fall into descriptive classes. We have some generic variables that I suspect refer to marital status, sex, age, et cetera. Geodemographic variables discuss the statistics about the policyholder's neighborhood. For example, average commute to work and the number of fire stations in an area could be geodemographic variables. Weather variables refer to regional statistics like annual rainfall. Crime variables measure statistics like the number of police officers in a community. (Note: I guess at possible interpretations for the variables because Liberty Mutual encrypted the variable names to safeguard their proprietary practices.)

Since the target variable is loss cost, many rows hold zeros for their target. That is, most policyholders do not file a claim for fire damage. This scenario calls for a compound Poisson-gamma model to describe the discrete mass at 0 and the otherwise continuous positive density. We build models with 2 generic variables, 2 geodemographic variables, and 7 weather variables. We estimate the Tweedie power to be 1.5. We compare the $Tw_{1.5}$ model to $Tw_{1.33}$ and $Tw_{1.66}$ models.

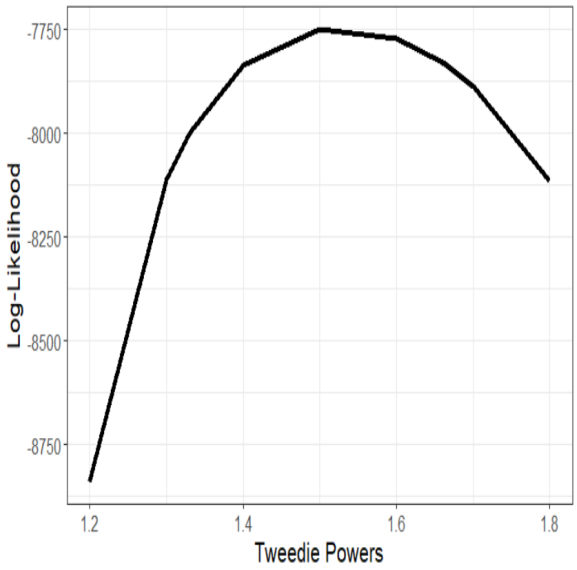

Figure 18: Profile Log-likelihood Plot for LM Home Fire LC Data

Again, we observe the best AIC measurement from the model with the estimated Tweedie power. For the models with power 1.33, 1.5, and 1.66, we measure AIC as 16018, 15518, and 15681. These results indicate that the model with the estimated Tweedie power provides a better fit. It makes sense that we get a better model by estimating a parameter value rather than arbitrarily selecting it.

We provide some tables to quantify the results. Table 4 summarizes the linear weights for the three models. Table 5 reports loss cost predictions for the first 5 rows. Similar to the automobile bodily injury claims, the weights only vary by an order of magnitude 10^{-2}. Due to some issues with the target variable, we do not report on the business value of our optimized model. (See the chapter endnotes.) Our models give a ratio of predicted losses over the true losses that is approximately 85 percent. I suspect this improvement from our first case study has to do with the increase in the number of variables used.

Explanatory Variable	Tweedie 1.33 Model	Tweedie 1.5	Tweedie 1.66 Model
Variable 10	-0.8316	-0.8395	-0.8493
Variable 13	-0.3864	-0.4015	-0.4206
Geodem 24	-0.7472	-0.5915	-0.4208
Geodem 37	0.1056	0.0750	0.0425
Weather 7	0.2223	0.1853	0.1502
Weather 10	-0.8446	-0.8730	-0.9013
Weather 72	-0.1523	-0.1650	-0.1793
Weather 102	0.0167	0.0168	0.0170
Weather 104	0.1804	0.1783	0.1760
Weather 118	0.0309	0.0314	0.0318
Weather 173	-0.0121	-0.0118	-0.0121

Table 4: Comparison of linear weights between LM Home Fire LC models.

Policyholder Index	Actual Loss	Tweedie Model 1.5	Tweedie Model 1.66
8	0	0.0037	0.0039
10	0	0.0060	0.0060
19	0	0.0014	0.0013
20	0	0.0033	0.0032
23	0	0.0033	0.0032

Table 5: Loss predictions for first 5 rows in LM Home Fire testing data.

We conclude this case study with some graphs. You can't see changes in prediction between the different models. Moreover, the models predict the loss costs poorly. All three models predict a distribution ostensibly similar to the distribution shown in Figure 19. Notice that the models uniformly predict loss ratios close to 0. Meanwhile, Figure 20 shows the actual count of loss ratios greater than 0. Essentially, all three models fail to identify risky policies.

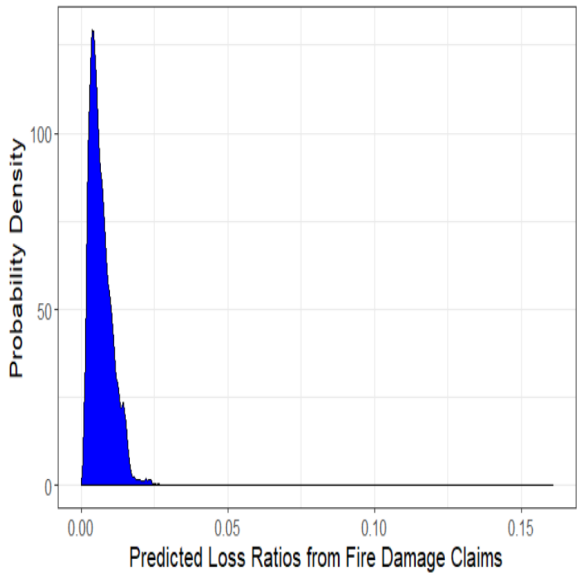

Figure 19: Distribution of Tw_p Loss Ratio Predictions

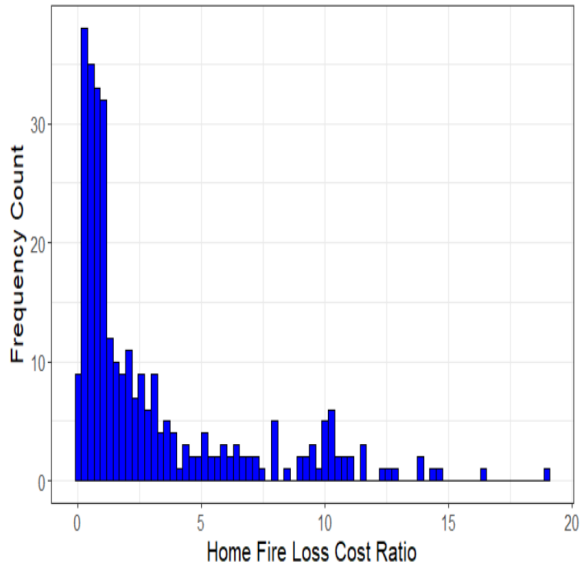

Figure 20: Non-zero Loss Ratios for LM Home Fire Policies

All things considered, I evaluate that tuning the Tweedie power provides little to no enhancement to the model. This parameter tuning mainly tinkers with individual predictions. For the two case studies presented in this thesis, I would advise alternative feature engineering to enhance the models.

Endnotes

- Our model fits a Tweedie distribution to empirical data. Implicitly, we are assuming that the data is well-described by a Tweedie distribution. But, this assumption could be flawed. Fitting a distribution to data is a difficult and imprecise science.

- The target variable for the Liberty Mutual dataset is not actually loss cost. It is a transformed ratio where the loss cost is divided by the premium. From Chapter 1, we know that Tweedie models are closed under scalar multiplication. However, this transformed ratio involves multiplication that is not constant. Policyholders pay different premiums. We accept this imperfection in order to use a larger dataset. The target still appears to make sense in the context of Tweedie models. However, we can't report on total loss costs in this scenario because the target is a ratio

- In both case studies, we see heavier losses predicted as the Tweedie power nears 2. Hypothetically, the actuary could observe this trend and pick a power value that begets the predictions he or she desires. Such an action calls into question the actuary's ethics. In this regard, estimating the Tweedie power via an algorithm appears to conform with reasonable ethical standards.

- I don't consider crime variables in modeling because they included many missing values. For the sake of time, I don't wrangle the empties. I choose to not exclude the large number of observations that have these empties. Hence, I don't use crime variables in my models.

- I don't include Tweedie powers close to 1 and 2 in the profile log-likelihood analysis. Dunn and Smyth's algorithm misbehaves in these situations.

6 Conclusion

One main motivation for this research is to enhance actuarial models via tuning the Tweedie power. In other words, is it worthwhile to estimate the Tweedie power in actuarial ratemaking? Both my case studies indicate that using an estimated Tweedie power can decrease the Akaike Information Criterion. However, it remains unclear how this change modifies the model and how much benefit this parameter tuning provides to the model.

There are advantages and disadvantages to estimating the Tweedie power. The case studies I provided suggest that estimation leads to a more probable model. It is good practice to use the power value that best characterizes the underlying data. Actuaries could justify their parameter selections to regulators by citing the statistical literature. On the other hand, estimating the Tweedie power could slow down model production and/or insert new costs for software acquisition and technical training. The actuary should balance practical and theoretical considerations to provide the best insurance product possible given the available resources.

This research started in earnest when my manager Tommy Wright told me that we arbitrarily set the Tweedie power to be 1.5 for loss cost models. I spent over a year researching for this project, at least nine of which I dedicated to studying the Tweedie family. Initially, the prospect of improving actuarial models by estimating the Tweedie power excited me greatly. Yet, I conclude this research by seeing reason behind the thing I had hoped to fix. That is, I now feel comfortable arbitrarily setting the Tweedie power to 1.5.

This conclusion notwithstanding, I developed my research skills, learned about mathematical topics at the graduate-level, and improved my facility with R programming. Second, this document makes accessible to actuaries and other STEM professionals some advanced statistical theory. Third, I expect researchers to continually refine tools used in generalized linear modeling. Fourth, a company's technical

infrastructure changes over time. Maybe someday it will make practical sense for actuaries to estimate the Tweedie power. Lastly, Tweedie models find application in other fields besides actuarial science. For other fields, accuracy could matter more than business considerations. In the end, I hope that you have found some value in this applied introduction to Tweedie models.

Appendix

Chapters 1 and 2 discussed Tweedie models and numerical approximations of their densities without introducing generating functions and measure theory. Speaking about Tweedie models at a high level made the thesis palatable for a larger audience. However, some readers want to look "under the hood" and see the base parts that make up Tweedie models. This appendix is for those readers.

One big advantage of this continued exploration is that we will add to our toolkit as probability theorists. With this expanded field knowledge, we will be able to prove results that we left unjustified in Chapters 1 and 2. Moreover, topics covered in this appendix have value for the graduate student in probability or statistics. I encourage you to take these appendix items as jumping-off points for further research.

Elementary Measure Theory

Some math students see ring theory and topology in their undergraduate schooling. To explain the basics of measure theory, I'll try to draw analogies to ring theory and topology. For instance, a topology defines what it means for a set to be open. A topology for X is a set of sets $\mathcal{T} \subset \mathcal{P}(X)$ for which the following is true:

(i.) $\emptyset$, $X \in \mathcal{T}$;

(ii.) arbitrary unions of sets in the topology are in the topology;

(iii.) finite intersections of sets in the topology are in the topology.

Measure theory involves a mathematical object similar to a topology: a σ-algebra. This object has three rules that determine what is in the collection of subsets.

Definition. A *σ-algebra* $\mathcal{A} \subset \mathcal{P}(X)$ is a collection of subsets of X for which the following is true:

(i.) $\emptyset$, $X \in \mathcal{A}$;

(ii.) countable unions of sets in $\mathcal{A}$ are in $\mathcal{A}$;

(iii.) complements of sets in $\mathcal{A}$ are in $\mathcal{A}$.

Informally, an algebra is a ring closed under scalar multiplication. A ring is a set of elements coupled with two operations. On the other hand, σ means summand (add things together) in mathematics. Thus, a σ-algebra is a ring for which we can perform an operation countably many times. The two operations are complement $'$ and union $\cup$.

How do σ-algebras relate to probability theory? σ-algebras make concrete the idea of a probability space. There are random events in a probability space. We want to be able to detect if an event is in our probability space. That is, if we take the union of countably many events, we can detect the event in our experiment. We also want to be able to identify the event in our probability space. If we take the intersection of countably many events, we can narrow in on the event in our experiment.

Definition. A nonempty set X equipped with a σ-algebra $\mathcal{A}$ is called a *measurable space*.

Definition. A *measure* is a countably additive, nonnegative, extended real value function defined on a σ-algebra. That is, $\nu : \mathcal{A} \to [0, \infty]$ such that

(i.) $\nu(\emptyset) = 0$;

(ii.) $\nu\Big(\bigcup_i^n A_i\Big) = \sum_i^n \nu(A_i)$ for mutually exclusive events $\Big\{A_i : i \in \mathbb{N}\Big\} \subset \mathcal{A}$.

Notice the relationship between the third axiom of probability and the definition of a measure. This is no coincidence. Probability is a measure. Measure theory is an abstraction of basic probability.

Example. Let $\{H, T\}$ be a set. Define the σ-algebra to be $\mathcal{P}(\{H, T\})$. Consider the measure $\nu : \mathcal{P}(\{H, T\}) \to [0, \infty]$ where $\nu(\{H\}) = 0.5$ and $\nu(\{T\}) = 0.5$. Then, $(\{H, T\}, \mathcal{P}(\{H, T\}), \nu)$ is a measure space. This measure space corresponds to the probability of flipping fair coins.

Remark. The power set of a set is always a σ-algebra.

Example. Let $X = \{1, 2, 3, 4, 5, 6\}$ and $\mathcal{P}(X)$ be a σ-algebra. Define the measure $\nu(i) = 1/6 \ \ \forall i \in X$. The measure space $(X, \mathcal{P}(X), \nu)$ corresponds to the probability of rolling a fair dice.

Example. Let X be an arbitrary nonempty set and $\mathcal{P}(X)$ be a σ-algebra. Define the measure

$$\nu_C(A) = \begin{cases} |A|, & A \text{ is finite} \\ \infty, & \text{otherwise} \end{cases} .$$

This measure ν_C is called the *counting measure*. We use it to define discrete probabilities in a measure-theoretic way.

Example. Another important measure is the *Lebesgue measure* ν_L. Given an open set, $\mathcal{O} = \coprod_k (a_k, b_k)$, where $\forall k \ \ (a_k, b_k)$ is an open interval. ($\coprod$ denotes the disjoint union.) Then, $\nu_L = \sum_k (b_k - a_k)$. The spirit of the Lebesgue measure is to measure length in $\mathbb{R}^n$.

There are two more measure-theoretic definitions relevant to this thesis. First, we say a measure ν is finite if $\nu(X)$ is finite. Second, we say a measure ν is σ-finite if the set X is the countable union of measurable sets with finite measure ν_i. It is immediate from the definitions that a finite measure is σ-finite. However, a measure being σ-finite doesn't imply that it is finite.

Integration

Most STEM students are familiar with integration from their calculus courses. This integration is Riemann integration. We denote it

$$\int_a^b f(x) \, dx \ .$$

There are many different ways to integrate though. Abstractly speaking, integration is the process of adding up (infinitely) many tiny pieces. Another popular integral is the Riemann-Stieltjes integral. For the Riemann integral, we call the function f the *integrand* and we call x in the dx part the *integrator*. The Riemann-Stieltjes integral is similar to the Riemann integral, but we allow the integrator function to be different than the identity function $g(x) = x$. That is,

$$\int_a^b f(x)\, dg(x) \ ,$$

where f is a function defined on $[a, b]$ and g is a monotone increasing function defined on $[a, b]$, is the Riemann-Stieltjes integral. Steven Krantz provides a formal definition of the Riemann-Stieltjes integral and a detailed study of it in his paper "The Integral: A Crux for Analysis" [17]. While researching the Riemann-Stieltjes integral, I didn't find many examples online of solved practice problems. Below we offer two examples of Riemann-Stieltjes integration. The examples highlight some important properties. See the properties of the Riemann-Stieltjes integral and some more examples here [9].

Example. Compute $\int_3^{10} x^2 d(x^2 + 2x + 1)$.

$$\begin{aligned}
\int_3^{10} x^2\, d(x^2 + 2x + 1) &= \int_3^{10} x^2\, dx^2 + 2\int_3^{10} x^2\, dx + \int_3^{10} x^2\, d1 \\
&= \int_3^{10} x^2 \cdot 2x\, dx + 2\int_3^{10} x^2\, dx + \int_3^{10} x^2 \cdot (1)'\, dx \\
&= 2\int_3^{10} x^3\, dx + 2\int_3^{10} x^2\, dx + 0 \\
&\approx 5603.167 \ .
\end{aligned}$$

Example. Solve $\int_{-3}^{5} x^3 \, d[x]$.

$$\begin{aligned}\int_{-3}^{5} x^3 \, d[x] &= \sum_{i=-2}^{5} x^3 \\ &= -8 - 1 + 0 + 1 + 8 + 27 + 64 + 125 \\ &= 216 \, .\end{aligned}$$

The floor function is constant for intervals of length 1 and then jumps up 1 mark at a new integer. The change in the integrator is 0, except for at integers. This argument explains why we can express Riemann-Stieltjes integrals with a floor function integrator as summations.

The key idea behind Riemann-Stieltjes integration is that the integrator function does not have to be static. We can integrate with a measure as the integrator. For instance, we can integrate

$$\int_X f(x) \, \nu(dx),$$

where f is defined on measurable space X and ν is a measure. Jorgensen bases his constructive definition of exponential dispersion models on integrals of this kind.

Generating Functions

Moment-generating functions and cumulant-generating functions are essential tools in any probability theorist's toolkit.

Definition. Let X be a random variable with density function f. The moment-generating function of X is

$$M_X(t) = \mathrm{E}[e^{Xt}] = \int e^{xt} f(x) dx.$$

Definition. Let X be a random variable with density function f. The cumulant-generating

function of X is

$$K_X(t) = \log M_X(t) = \log \mathrm{E}[e^{Xt}] = \log \int e^{xt} f(x)dx.$$

Proposition.

$$M_{a+bX}(t) = e^{at} M_X(bt).$$

$$K_{a+bX}(t) = at + K_X(bt).$$

Proof.

$$\begin{aligned} M_{a+bX}(t) &= \int e^{(a+bx)t} f(x)dx \\ &= \int e^{at} e^{x(bt)} f(x)dx \\ &= e^{at} \int e^{x(bt)} f(x)dx \\ &= e^{at} M_X(bt). \end{aligned}$$

$$\begin{aligned} K_{a+bX}(t) &= \log \left[e^{at} \int e^{x(bt)} f(x)dx \right] \\ &= \log e^{at} + \log \int e^{x(bt)} f(x)dx \\ &= at + K_X(bt). \end{aligned}$$ □

Proposition. If X and Y are independent random variables, then

$$M_{X+Y}(t) = M_X(t) M_Y(t) \text{ and } K_{X+Y}(t) = K_X(t) + K_Y(t).$$

Proof. By independence, $M_{X+Y}(t) = \mathrm{E}[e^{(X+Y)t}] = \mathrm{E}[e^{Xt}]\mathrm{E}[e^{Yt}] = M_X(t)M_Y(t)$. Likewise, $K_{X+Y}(t) = \log M_{X+Y}(t) = \log M_X(t) + \log M_Y(t) = K_X(t) + K_Y(t)$. □

The jth moment is $\mathrm{E}[X^j] = M_X^{(j)}(0)$ where $^{(j)}$ denotes the jth derivative of the function. Similarly, the jth cumulant is $K_X^{(j)}(0)$. Cumulants are a useful way to compute expectations and variances; the first cumulant is the expectation and the second cumulant is the variance. Sometimes the 3rd cumulant is referred to as skewness and the 4th cumulant is referred to as kurtosis. People say that the third cumulant affects how the distributon leans and that the fourth cumulant impacts the sharpness of the peak in the distribution. In general, it is hard to describe how cumulants after these influence the shape of a distribution.

A Constructive Definition for EDMs

This subsection reiterates much of what Jorgensen formalizes in ***The Theory of Dispersion Models*** [15]. I cherry-pick definitions and theorems for discussion. We start with one-parameter models. Then, we complicate things by introducing a second parameter.

Definition. Let ν be a σ-finite measure on $\mathbb{R}$. Define the *cumulant function* $\kappa(\theta)$ for $\theta \in \mathbb{R}$ as

$$\kappa(\theta) = \log \int e^{\theta y} \, \nu(dy).$$

The domain of the cumulant function $\kappa(\theta)$ is

$$\Theta = \left\{ \theta \in \mathbb{R} : \left(\int e^{\theta y} \, \nu(dy) \right) < \infty \right\}.$$

This definition asks a lot of the reader all at once. In practice, we seldom compute the cumulant function $\kappa(\theta)$. Proofs and results we give substitute in $\kappa(\theta)$ as a convenient way to encapsulate this analytic expression. What is suspect about this definition is the σ-finite measure ν. We would like to know what ν is before we proceed.

Recall the $a(y; \sigma^2)$ functions present in the EDM densities. After Proposition 1.1, I remarked that these $a(y; \sigma^2)$ functions maintain incredible flexibility. We proposed

suitable $a(y;\sigma^2)$ functions for the normal, Poisson, binomial, and gamma cases. These candidates didn't look akin to one another. In general, these $a(y;\sigma^2)$ functions don't have closed forms. Let $b(y;\sigma^2)$ be a function like $a(y;\sigma^2)$ that takes input (y,σ^2). For now, let $\sigma^2 = 1$.

$$\nu(dy) = b(y;1)dy,$$

where dy is the Lebesgue measure. Similar to the $a(y;\sigma^2)$ function, this function $b(y;\sigma^2)$ has a lot of freedom.

For a random variable Y parameterized by θ and defined on a measurable sets A, the cumulative distribution is

$$Pr_\theta(Y \in A) = \int_A \exp\{y\theta - \kappa(\theta)\}\ \nu(dy).$$

Notice the density function $\exp\{y\theta - \kappa(\theta)\}b(y;1)$ inside this expression. Contrast this density with the axiomatic density we presented in Chapter 1:

$$f(y;\mu,1) = a(y;1)\exp\{-\frac{1}{2}d(y;\mu)\}.$$

We will soon argue that these two formulations are the same.

With this expression for the distribution, it is easy to determine the moment-generating function and the cumulant-generating function for the random variable Y.

$$\begin{aligned} M_Y(t;\theta) &= \int \exp\{yt\}\exp\{y\theta - \kappa(\theta)\}\ \nu(dy) \\ &= \exp\{-\kappa(\theta)\}\int \exp\{y\theta + yt\}\ \nu(dy) \\ &= \exp\{-\kappa(\theta)\}\exp\{\kappa(\theta+t)\} \\ &= \exp\{\kappa(\theta+t) - \kappa(\theta)\}. \end{aligned}$$

We get the third line in the derivation by citing the definition of the cumulant function. The result gives a neat expression for the moment-generating function. Since the cumulant-generating function $K_Y(t;\theta)$ is just the log of the moment-generating function, it follows immediately that

$$K_Y(t;\theta) = \kappa(\theta + t) - \kappa(\theta).$$

Using these simple expressions, we can compute the jth cumulant and the jth moment with respect to t. This evaluation justifies the jargon "cumulant function" $\kappa(\theta)$.

$$K^{(j)}(t;\theta) = \frac{\partial^{(j)} K(t;\theta)}{\partial t^j} = \kappa^{(j)}(\theta + t).$$

$$K^{(j)}(0;\theta) = \kappa^{(j)}(\theta).$$

Recall that the first cumulant is the mean μ. Observe that

$$\kappa'(\theta) = \mu.$$

Define a function $\tau : \Theta \to \Omega$ where Θ is the parameter space, Ω is the mean parameter space, and $\tau(\theta) = \kappa'(\theta)$. In other words, we have a function that maps the parameter θ to the position parameter μ. Let $\tau^{-1} : \Omega \to \Theta$ be the inverse function of τ. This function τ^{-1} maps a position parameter μ to the parameter θ. With functions τ and τ^{-1}, we will soon show that the axiomatic defintion and the constructive definition of an exponential dispersion model express the same idea. First, we formalize the constructive definition of an exponential dispersion model.

Definition. We call $\{Pr_\theta : \theta \in \Theta\}$ a *one-parameter exponential family* if

(i) the distribution functions do not map to a constant value to 1;

(ii) Θ contains more elements than just 0.

These two conditions say that the distribution functions describe random behavior and that the family includes at least two members. We generalize this one-parameter family into a two-parameter family: exponential dispersion models. Let Σ be a set containing elements $\sigma^2 > 0$. Given a one-parameter exponential family,

$$\frac{\kappa(\theta)}{\sigma^2} = \log \int \exp\left\{\theta \frac{y}{\sigma^2}\right\} \nu_{\frac{1}{\sigma^2}}(dy)$$

holds for some σ-finite measure $\nu_{\frac{1}{\sigma^2}}$. Essentially, we have scaled the original one-parameter exponential family by a second parameter (the dispersion parameter).

The distribution function for a member Y of the two-parameter exponential family has a familiar form:

$$Pr_{(\theta,\sigma^2)}(Y \in A) = \int_A \exp\left\{\frac{y\theta - \kappa(\theta}{\sigma^2}\right\} \nu_{\frac{1}{\sigma^2}}(dy).$$

We evaluate the moment-generating function and cumulant-generating function with this distribution function.

$$\begin{aligned} M_Y(t;\theta,\sigma^2) &= \int \exp\{yt\} \exp\left\{\frac{y\theta - \kappa(\theta)}{\sigma^2}\right\} \nu_{\frac{1}{\sigma^2}}(dy) \\ &= \exp\left\{ -\frac{\kappa(\theta)}{\sigma^2}\right\} \int \exp\left\{\frac{y}{\sigma^2}(\theta + t\sigma^2)\right\} \nu_{\frac{1}{\sigma^2}}(dy) \\ &= \exp\left\{ -\frac{\kappa(\theta)}{\sigma^2})\right\} \exp\left\{\frac{\kappa(\theta + t\sigma^2)}{\sigma^2}\right\} \\ &= \exp\left\{\frac{\kappa(\theta + t\sigma^2) - \kappa(\theta)}{\sigma^2}\right\}. \end{aligned}$$

$$K_Y(t;\theta,\sigma^2) = \log\left(\exp\left\{\frac{\kappa(\theta + t\sigma^2) - \kappa(\theta)}{\sigma^2}\right\}\right) = \frac{\kappa(\theta + t\sigma^2) - \kappa(\theta)}{\sigma^2}.$$

Find the first and second cumulants by differentiating with respect to t at $t = 0$.

$$\begin{aligned}\frac{\sigma^2\kappa'(\theta)}{\sigma^2} &= \kappa'(\theta)\\ &= \tau(\theta)\\ &= \mu.\end{aligned}$$

$$\begin{aligned}\frac{(\sigma^2)^2\kappa''(\theta)}{\sigma^2} &= \sigma^2\kappa''(\theta)\\ &= \sigma^2\tau'(\mu).\end{aligned}$$

Recall that the second cumulant $K''(0)$ gives variance for the random variable. Moreover, $\sigma^2 \cdot V(\mu)$ is the variance of an EDM. Therefore, $\tau'(\theta) = \kappa''(\theta) = V(\mu)$. We have three ways to describe the variance function for an exponential dispersion model.

Return to the distribution function $Pr_{(\theta,\sigma^2)}(Y \in A)$. We defined this function with a σ-finite measure $\nu_{\frac{1}{\sigma^2}}(y)$, but we didn't explain what the measure is. Let $\nu_{\frac{1}{\sigma^2}}(y)$ equal $b(y;\sigma^2)dy$. It follows that the density is

$$f(y;\mu,\sigma^2) = b(y;\sigma^2)\exp\left\{\frac{y\theta - \kappa(\theta)}{\sigma^2}\right\} = b(y;\sigma^2)\exp\left\{\frac{y\tau^{-1}(\mu) - \kappa(\tau^{-1}(\mu))}{\sigma^2}\right\}.$$

Compare with the axiomatic density

$$f(y;\mu,\sigma^2) = a(y;\sigma^2)\exp\left\{-\frac{1}{2\sigma^2}d(y;\mu)\right\}.$$

In ***The Theory of Dispersion Models***, Jorgensen gives that the unit deviance $d(y;\mu)$ for an exponential dispersion model as

$$2\Big[\sup_{\theta\in\Theta}\Big(y\theta - \kappa(\theta)\Big) - y\tau^{-1}(\mu) + \kappa(\tau^{-1}(\mu))\Big].$$

Use calculus to find θ that maximizes $y\theta - \kappa(\theta)$.

$$\frac{\partial}{\partial\theta}\Big(y\theta - \kappa(\theta)\Big) = y - \kappa'(\theta) = y - \tau(\theta) = 0.$$

Suppose y is in the mean parameter space Ω. Observe that $\tau^{-1}(y)$ maximizes $y\theta - \kappa(\theta)$. Thus, when $y \in \Omega$, we get unit deviance $d(y;\mu)$ as

$$2\Big[y\tau^{-1}(y) - \kappa(\tau^{-1}(y)) - y\tau^{-1}(\mu) + \kappa(\tau^{-1}(\mu))\Big].$$

Recall that the mean parameter space Ω is the interior of the convex support C. Table 1 in Chapter 1 records the convex supports for Tweedie models. Most of the time, the convex support is an open interval like $\mathbb{R}$. As a result, y is almost always in the mean parameter space. Only for $Tw_p(\mu,\sigma^2)$ with $1 < p < 2$ is the convex support the half open interval $[0,\infty)$. Jorgensen handles this special case when $y = 0$ in his book [15]. To keep things simple, we assume $y \in \Omega$. The following proposition draws the connection between the axiomatic definition and the constructive definition.

Proposition. Let f_a be the density function from the axiomatic definition and f_c be the density function from the constructive definition. Suppose $a(y;\sigma^2) = f_c(y;y,\sigma^2)$. Then, for Tweedie models,

$$f_a(y;\mu,\sigma^2) = f_c(y;\mu,\sigma^2).$$

That is,

$$a(y;\sigma^2)\exp\left\{-\frac{1}{2\sigma^2}d(y;\mu)\right\} = b(y;\sigma^2)\exp\left\{\frac{y\theta - \kappa(\theta)}{\sigma^2}\right\}.$$

Proof. By assumption, we have a Tweedie model. As a result, $y \in \Omega$ and the unit deviance is

$$2\Big[y\tau^{-1}(y) - \kappa(\tau^{-1}(y)) - y\tau^{-1}(\mu) + \kappa(\tau^{-1}(\mu))\Big].$$

Substitute in $a(y;\sigma^2) = f_c(y;y,\sigma^2)$.

$$\begin{aligned} f_a(y;\mu,\sigma^2) &= f_c(y;y,\sigma^2)\exp\left\{-\frac{1}{2\sigma^2}d(y;\mu)\right\} \\ &= b(y;\sigma^2)\exp\left\{\frac{y\tau^{-1}(y)-\kappa(\tau^{-1}(y))}{\sigma^2}\right\}\exp\left\{-\frac{1}{2\sigma^2}d(y;\mu)\right\} \\ &= b(y;\sigma^2)\exp\left\{\frac{y\theta-\kappa(\theta)}{\sigma^2}\right\} \\ &= f_c(y;\mu,\sigma^2). \end{aligned}$$ □

Remark. We only discuss the convex supports of Tweedie models. In general, the axiomatic and constructive definitions of exponential dispersion models are the same. See *The Theory of Dispersion Models* [15].

To summarize, we now have a different way to think about EDM (and Tweedie) densities in terms of cumulant functions, moment-generating functions, and cumulant-generating functions. This version is identical to what we discussed in Chapter 1. Besides expanding our toolkit with moments and cumulants, we revealed how EDMs are constructed by scaling one-parameter exponential families.

Tweedie Family Proofs

We asserted some propositions for Tweedie models in Chapter 1 without justification. Furthermore, we suggested in a footnote that some index parameters p correspond to stable distributions. I am happy to now give arguments for these results. I hope that the inquisitive reader who has persevered through this not-so-gentle introduction receives some satisfaction from this subsection.

Recall that the variance function for a Tweedie model is $V(\mu) = \mu^p$. We determine an expression for parameter θ and cumulant function $\kappa(\theta)$ in terms of μ and p. Observe

that

$$\kappa''(\theta) = \frac{\partial \tau(\theta)}{\partial \theta} = \frac{\partial \mu}{\partial \theta} = \mu^p.$$

Ignoring the arbitrary constants we get from indefinite integration,

$$\theta = \begin{cases} \frac{\mu^{1-p}}{1-p} & p \neq 1 \\ \log \mu & p = 1 \end{cases}.$$

Jorgensen writes μ in terms of θ and p as well [15]. He introduces a parameter α that is related to p in that

$$\alpha = \frac{p-2}{p-1}.$$

The inverse relationship says that

$$p = \frac{\alpha - 2}{\alpha - 1}.$$

Consider that

$$p - 1 = \frac{\alpha - 2}{\alpha - 1} - 1 = \frac{\alpha - 2}{\alpha - 1} - \frac{\alpha - 1}{\alpha - 1} = \frac{-1}{\alpha - 1}.$$

Now, compute μ in terms of θ and α by finding the inverse of θ in terms of μ and p.

$$\mu = \begin{cases} \left(\frac{\theta}{\alpha - 1}\right)^{\alpha - 1} & p \neq 1 \\ e^{\theta} & p = 1 \end{cases}.$$

Next, we find the cumulant function $\kappa(\theta)$ by solving the differential equation $\kappa'(\theta) = \tau(\theta) = \mu$. For $p \neq 1, 2$, integrate to get

$$\kappa(\theta) = \frac{\alpha - 1}{\alpha} \left(\frac{\theta}{\alpha - 1}\right)^{\alpha}.$$

For $p = 1$, we integrate to see that $\kappa(\theta) = e^{\theta}$. The tricky case is for $p = 2$. When

$p = 2$, $\alpha = 0$. Thus, we get

$$\kappa'(\theta) = \frac{-1}{\theta}.$$

The antiderivative is $-\log(-\theta)$. In summary,

$$\kappa(\theta) = \begin{cases} \frac{\alpha-1}{\alpha}\left(\frac{\theta}{\alpha-1}\right)^{\alpha} & p \neq 1, 2 \\ -\log(-\theta) & p = 2 \\ e^{\theta} & p = 1 \end{cases}.$$

I understand that it is difficult to remember these many relationships between p, α, μ, θ, $\kappa(\theta)$, and $V(\mu)$. Math is trying sometimes because it challenges your mental faculties, and sometimes it is trying because it requires you to regurgitate and synthesize a bulk of information. The task at hand speaks to the latter. For the proceeding proofs, we will use these expressions. Refer back to this page and the preceding page when you need to recall the relationships between Tweedie parameters and functions.

Proposition. There are no Tweedie models with index parameter $0 < p < 1$.

Proof. Assume there exists a $Tw_p(\mu, \sigma^2)$ where $p \in (0, 1)$. We know that

$$\kappa(\theta) = \frac{\alpha - 1}{\alpha}\left(\frac{\theta}{\alpha - 1}\right)^{\alpha}.$$

Differentiate twice with respect to θ to find that

$$\kappa''(\theta) = \left(\frac{\theta}{\alpha - 1}\right)^{\alpha-2}.$$

Before we take the next step, we must consider what α is. The relationship between α and p is best shown graphically. Observe that $\lim_{p\to 0^+} \frac{p-2}{p-1} = 2$ and that the mapping from p to α is monotone increasing. Therefore, $\alpha - 2 > 0$.

Recall that $0 \in \Theta$ for all exponential dispersion models. For $\theta = 0$, the variance of the Tweedie random variable is 0. That is, the Tweedie random variable is not stochastic at all. This result contradicts the fact that Tweedie models are random objects. We conclude that there exist no Tweedie models indexed by $p \in (0, 1)$. □

We can derive the general unit deviance for Tweedie models as well. In Chapter 1, we saw the unit deviances for the normal, Poisson, and gamma cases. These Tweedie models are special cases.

Proposition. For a Tweedie model with index power $p \notin \{0, 1, 2\}$, the unit deviance $d(y; \mu)$ is

$$2\left\{ \frac{\max(y, 0)^{2-p}}{(1-p)(2-p)} - \frac{y\mu^{1-p}}{1-p} + \frac{\mu^{2-p}}{2-p} \right\}.$$

Proof. Recall that Tweedie models have unit deviance

$$2\left\{ \sup_{\theta \in \Theta} \Big[y\theta - \kappa(\theta) \Big] - \Big(y\tau^{-1}(\mu) + \kappa(\tau^{-1}(\mu)) \Big) \right\}.$$

We know that $\tau^{-1}(\mu) = \theta$ and $\kappa(\tau^{-1}(\mu)) = \kappa(\theta)$. Verify that

$$y\tau^{-1}(\mu) = y\theta = \frac{y\mu^{1-p}}{1-p}.$$

This equality follow directly from how we described θ in terms of μ and p. Writing $\kappa(\theta)$ in terms of μ and p requires more algebraic manipulation than just substitution. Check the algebra below. We start by simplifying what is inside of the parentheses, and then we simplify the multiplier outside of the parentheses. Observe that we apply the formula $(p-1)(\alpha-1) = -1$ multiple times.

$$
\begin{aligned}
\kappa(\theta) &= \frac{\alpha-1}{\alpha}\Big(\frac{\theta}{\alpha-1}\Big)^{\alpha} \\
&= \frac{\alpha-1}{\alpha}\Big(\frac{\mu^{1-p}}{(1-p)(\alpha-1)}\Big)^{\alpha} \\
&= \frac{\alpha-1}{\alpha}(\mu^{1-p})^{\alpha} \\
&= \frac{\alpha-1}{\alpha}(\mu^{1-p})^{\frac{p-2}{p-1}} \\
&= \frac{\alpha-1}{\alpha}\mu^{\frac{(p-2)(1-p)}{p-1}} \\
&= \frac{\alpha-1}{\alpha}\mu^{2-p} \\
&= \frac{(\alpha-1)(p-1)}{p-2}\mu^{2-p} \\
&= -\frac{\mu^{2-p}}{p-2} \\
&= \frac{\mu^{2-p}}{2-p}.
\end{aligned}
$$

Next, we concern ourselves with $\sup_{\theta\in\Theta}\Big[y\theta-\kappa(\theta)\Big]$. From calculus, we see that the maximum occurs when

$$
\begin{aligned}
y &= \Big(\frac{\theta}{\alpha-1}\Big)^{\alpha-1} \\
&= \Big(\frac{\mu}{(\alpha-1)(1-p)}\Big)^{(\alpha-1)(1-p)} \\
&= \mu^{(\alpha-1)(1-p)} \\
&= \mu.
\end{aligned}
$$

This calculation means that we only get a solution if $y \geq 0$. Substitute in $\max(y,0)$ for y.

Compute

$$\begin{aligned}\sup_{\theta\in\Theta}\Big[y\theta-\kappa(\theta)\Big] &= \max(y,0)\cdot\frac{\max(y,0)^{1-p}}{1-p}-\frac{\max(y,0)^{2-p}}{2-p}\\ &= \frac{\max(y,0)^{2-p}}{1-p}-\frac{\max(y,0)^{2-p}}{2-p}\\ &= \frac{\max(y,0)^{2-p}}{(1-p)(2-p)}.\end{aligned}$$

Bring all the pieces together. For $Tw_p(\mu,\sigma^2)$ with $p\notin\{0,1,2\}$, we conclude that the unit deviance is

$$2\left\{\frac{\max(y,0)^{2-p}}{(1-p)(2-p)}-\frac{y\mu^{1-p}}{1-p}+\frac{\mu^{2-p}}{2-p}\right\}. \qquad \square$$

The last interesting thing we will say about Tweedie models harkens back to a comment I made in Chapter 1. I remarked that some Tweedie models are stable distributions for some parameter choices θ. Stable distributions connect to limiting distributions and they find application in many financial settings. For example, James Weatherall talks at a high level about stable distributions in his science-popularizing book *The Physics of Wall Street* [29]. This topic is something I plan to study more during my professional career. For now, we introduce stable distributions in the context of Tweedie models.

Definition. Let $X_1,\cdots,X_n$ be independent, identically distributed random variables with distribution F. We say X is a *stable* random variable, if, for all $n\in\mathbb{N}$, there exist constant b and non-constants c_i such that

$$aX+b=\sum_{i=1}^{n}c_iX_i$$

also has distribution F. If $b=0$, we say that X is a *strictly stable* random variable.

Theorem. Let X be a Tweedie model with $p\in(-\infty,0]\cup(2,\infty)$ and with $\tau^{-1}(\mu)=0$. X

is a strictly stable random variable.

Proof. Observe that $\tau^{-1}(\mu) = \theta$, so $\theta = 0$. The cumulant-generating function of the Tweedie model is

$$K_X(t; 0, \sigma^2) = \frac{\kappa(t\sigma^2) - \kappa(0)}{\sigma^2}.$$

Recall that $\kappa(\theta) = \frac{\alpha-1}{\alpha}\left(\frac{\theta}{\alpha-1}\right)^{\alpha}$ for the given p. For $p \leq 0$, $\alpha \in (1, 2]$; for $p > 2$, $\alpha \in (0, 1)$. We see these mappings in Figures 21 and 22. Consequently, $|\alpha - 1| > 0$ and $\kappa(0) = 0$.

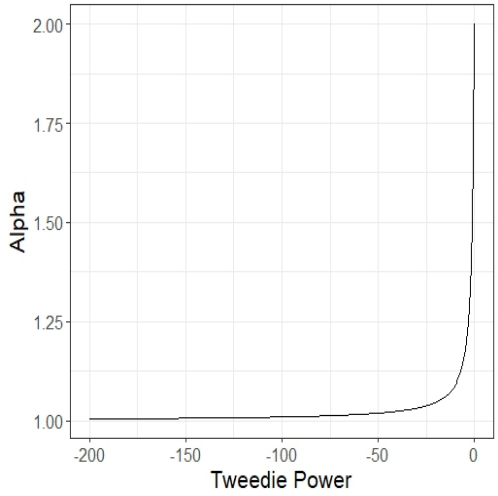

Figure 21: Relationship between α and p for $p \leq 0$

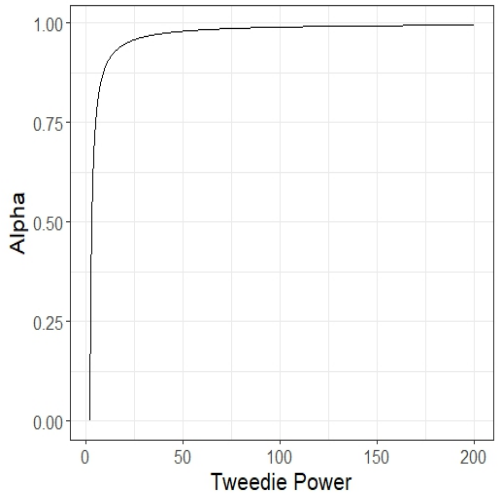

Figure 22: Relationship between α and p for $p > 2$

Consider $X_1, \cdots, X_n$ independent, identically distributed copies of X. Evaluate the cumulant-generating function of $\sum_{i=1}^{n} X_i$. That is,

$$\begin{aligned}\sum_{i=1}^{n} K_{X_i}(t; 0, \sigma^2) &= n \cdot K_X(t; 0, \sigma^2) \\ &= \frac{n(\alpha-1)}{\alpha\sigma^2}\Big(\frac{t\sigma^2}{\alpha-1}\Big)^{\alpha} \\ &= \frac{\alpha-1}{\alpha\sigma^2}\Big(\frac{n^{1/\alpha}t\sigma^2}{\alpha-1}\Big)^{\alpha} \\ &= K_X(n^{1/\alpha}t; 0, \sigma^2).\end{aligned}$$

$X_1 + \cdots + X_n$ has the same distribution as $n^{1/\alpha}X$. By definition, X is a strictly stable distribution. □

Besides the fact that some Tweedie random variables are stable in special cases, Jorgensen justifies the jargon "stable" by stating that Tweedie random variables have properties similar to stable distributions. For example, both Tweedie random variables and stable random variables are infinitely divisible. Moreover, Tweedie distributions appear as limiting distributions in a kind of generalized central limit theorem [15]. See *The Theory of Dispersion Models* to study this tangential idea.

Renewal Theory

This subsection covers some properties of renewal processes and proves some related propositions. Proofs in this section follow the approaches taken in Sheldon Ross' "Stochastic Process" [22].

A convolution expresses how one function modifies the integral of another function. We interpret a convolution visually as the overlap seen when shifting one function over another function.

Definition. $(f \star g)(t) = \int_0^t f(\tau)g(t-\tau)d\tau$ is the *convolution* of shifting g over f.

Remark. The bounds of integration do not need to be 0 and t. We define convolution with these bounds because most Tweedie models have the nonnegative reals as their domains. t is some number that is in the domain of both f and g.

Note. We denote the n-fold convolution as $f^{\star n} = f \star \cdots \star f$.

Here are some useful properties for convolutions:

- Associativity: $(f \star g) \star h = f \star (g \star h)$.
- Commutativity: $f \star g = g \star f$.

We use convolutions in probability to find the cumulative distribution function of a sum of two random variables. Suppose X and Y are random variables with cumulative distribution functions F_X and F_Y. Then, the new random variable $X+Y$ has cumulative distribution functions $F_X \star F_Y$. n-fold convolution comes in handy when we talk about infinitely divisible random variables. For example, consider n i.i.d. exponential random variables with cumulative distribution function F. The sum of these random variables is a gamma random variable with cumulative distribution function $F^{\star n}$. With this background in convolutions well-established, we now return to renewal theory.

Much of renewal theory involves the study of the renewal function. The renewal function $r(t)$ is the expectation of the counting process. That is, $r(t) = \mathrm{E}[N(t)]$. Another way to express the renewal function is as a n-fold convolution.

Proposition. $r(t) = \sum_{n=1}^{\infty} F_n(t)$ where F_n is the n-fold convolution of the interarrival distribution.

Proof. Evaluate the statement $N(t) \geq n \iff S_n \leq t$. In words, this statement means that the number of renewals by time t is greater than or equal to n if and only if n renewals occur before time t. Review this statement until you agree with its logic.

Next, consider the following:

$$\begin{aligned} Pr(N(t) = n) &= Pr(N(t) \geq n) - Pr(N(t) \geq n+1) \\ &= Pr(S_n \leq t) - Pr(S_{n+1} \leq t) \\ &= F_n(t) - F_{n+1}(t). \end{aligned}$$

Compute $\sum_{n=1}^{\infty} nP(N(t) = n) = \mathbb{E}[N(t)]$.

$$\begin{aligned} \sum_{n=1}^{\infty} nP(N(t) = n) &= \sum_{n=1}^{\infty} nF_n(t) - nF_{n+1}(t) \\ &= F_1(t) - F_2(t) + 2F_2(t) - 2F_3(t) + 3F_3(t) - 3F_4(t) + \cdots \\ &= F_1(t) + F_2(t) + F_3(t) + \cdots \\ &= \sum_{n=1}^{\infty} F_n(t). \end{aligned}$$

We conclude that the renewal function $r(t)$ is the infinite sum of the n-fold convolutions of the interarrival distribution F. □

In our proof of Wald's equation in the thesis body, we assumed that the expectation of $\mathrm{E}[N(t)]$ to be finite. This assumption requires a proof. For the next two propositions, we investigate the finite nature of the random variable $N(t)$.

Proposition. With probability 1, $\frac{N(t)}{t} \to \frac{1}{\mu}$ at $t \to \infty$.

Proof. Consider $S_{N(t)}$ and $S_{N(t)+1}$. $S_{N(t)}$ is the time of the last renewal prior to or at time t. Likewise, $S_{N(t)+1}$ is the time of the first renewal after time t. Thus, $S_{N(t)} \leq t \leq S_{N(t)+1}$. Now, divide by $N(t)$.

$$\frac{S_{N(t)}}{N(t)} \leq \frac{t}{N(t)} \leq \frac{S_{N(t)+1}}{N(t)} \ .$$

(If we define the renewal process with the zeroth renewal at time 0, then $N(t) \geq 1$. So it is legal to divide by $N(t)$.) Next, observe that S refers to the timing of events and N refers to the count of events. It would make sense if the average of the time elapsed until an event happened was the expectation of the interarrival distribution. From the strong Law of Large Numbers, we get that $\mu \leq \frac{t}{N(t)} \leq \mu$. Here, μ is $\mathrm{E}[X_i]$. Take the reciprocal to get $\mu \leq \frac{N(t)}{t} \leq \mu$. (Again, we exercise caution in dividing by μ. But, $\mu > 0$ from our definition of the interarrival times.) Therefore, $\frac{N(t)}{t} \to \frac{1}{\mu}$ with probability 1. □

Note. This proposition speaks to the rate at which the counting process grows whereas the next proposition states that the count of random events is finite in expectation.

Proposition. $\mathrm{E}[N(t)]$ is finite.

Proof. For interarrival time X_i, we know that $Pr(X_i = 0) < 1$ by definition. Thus, there exists $\alpha > 0$ such that $Pr(X_i \geq \alpha) > 0$. Let $\{\bar{X}_n, n \geq 1\}$ be a set of random variables where

$$\bar{X}_n = \begin{cases} 0 & X_n < \alpha \\ \\ \alpha & X_n \geq \alpha \end{cases}.$$

Observe that $\bar{X}_n \leq X_n$. Define the counting process $\bar{N}(t) = \max\{n : \bar{X}_1 + \cdots + \bar{X}_n \leq t\}$. The random variable $\bar{X}_1 + \cdots + \bar{X}_n$ takes on only discrete values $k\alpha$ in its support ($k \in \mathbb{N}$). Consider a renewal for the $\bar{X}_n$ random variable when $X_n \geq \alpha$. The counts of these renewals at the times $k\alpha$ are independent geometric random variables (getting a α is a success and getting a 0 is a failure). Their mean p is $Pr(X_n \geq \alpha)$. Formalize these independent geometric random variables as $G_1, \cdots, G_n$.

Suppose $n\alpha \leq t \leq (n+1)\alpha$. Compute $\mathrm{E}[\bar{N}(t)]$ as

$$\sum_{k=1}^{n} \mathrm{E}[G_k] = \frac{n}{p} \leq \frac{t/\alpha}{p} < \infty\,.$$

Because $\bar{X}_n \leq X_n$, it follows that $\bar{N}(t) \geq N(t)$. In English, we count more renewals when the interarrival time is shorter. Take expectations for the two counting processes. Clearly, $\mathrm{E}[\bar{N}(t)] \geq \mathrm{E}[N(t)]$. We conclude that the expectation of $N(t)$ is finite. (Equivalently, the renewal function is finite.) $\square$

Remark. The approach of this proof is to define a related renewal process, and then show that the corresponding counting process is both finite and greater than the original counting process.

References

[1] Stephen Abbott. *Understanding Analysis*. Vol. 2. Springer, 2015.

[2] *Beginning Statistics: Continuous Random Variables*. https://2012books.lardbucket.org/books/beginning-statistics/s09-continuous-random-variables.html. Accessed: 2018-05-03.

[3] Peter Dunn. *Package 'tweedie'*. R Documentation. Package for R programming. Dec. 2016.

[4] Peter K Dunn and Gordon K Smyth. "Evaluation of Tweedie Exponential Dispersion Model Densities by Fourier Inversion". In: *Statistics and Computing* 18.1 (2008), pp. 73–86.

[5] Peter K Dunn and Gordon K Smyth. "Series Evaluation of Tweedie Exponential Dispersion Model Densities". In: *Statistics and Computing* 15.4 (2005), pp. 267–280.

[6] Peter K Dunn and Gordon K Smyth. "Tweedie family densities: methods of evaluation". In: *Proceedings of the 16th International Workshop on Statistical Modelling, Odense, Denmark*. 2001, pp. 2–6.

[7] Richard Durrett. *Essentials of Stochastic Processes*. Vol. 2. Springer, 2012.

[8] Kirill Eremenko. *R Programming A-Z: R for Data Science with Real Exercises!* https://www.udemy.com/r-programming/. Advanced Visualization with GGPlot2. 2018.

[9] *Evaluating Riemann-Stieltjes Integrals*. http://mathonline.wikidot.com/evaluating-riemann-stieltjes-integrals. Accessed: 2018-02-15.

[10] David A Freedman. *Statistical Models: Theory and Practice*. Cambridge University Press, 2009.

[11] Mark Goldburd, Anand Khare, and Dan Tevet. "Generalized Linear Models for Insurance Rating". In: *Casualty Actuarial Society*. 2016.

[12] Md Masud Hasan and Peter K Dunn. "Two Tweedie distributions that are Near-optimal for Modelling Monthly Rainfall in Australia". In: *International Journal of Climatology* 31.9 (2011), pp. 1389–1397.

[13] John K Hunter. *Measure Theory Notes*. https://www.math.ucdavis.edu/~hunter/measure_theory/measure_notes.pdf. University of California Davis, 2011.

[14] Bent Jorgensen. "Exponential Dispersion Models". In: *Journal of the Royal Statistical Society. Series B (Methodological)* (1987), pp. 127–162.

[15] Bent Jorgensen. *The Theory of Dispersion Models*. Chapman and Hall, 1997.

[16] Daniel Kahneman. *Thinking Fast and Slow*. Farrar, Straus and Giroux, Oct. 2011.

[17] Steven G Krantz. "The Integral: A Crux for Analysis". In: *Synthesis Lectures on Mathematics and Statistics* 4.1 (2011), pp. 76–80.

[18] Richard J Larsen, Morris L Marx, et al. *An Introduction to Mathematical Statistics and Its Applications*. Vol. 5. Pearson, 2017.

[19] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2018. URL: https://www.R-project.org/.

[20] Sidney I Resnick. *Adventures in Stochastic Processes*. Springer Science & Business Media, 2013.

[21] Sheldon M Ross. *A First Course in Probability*. Pearson Education International, 2009.

[22] Sheldon M Ross. *Stochastic Processes*. John Willy & Sons, 1983.

[23] Sheng G Shi. "Direct Analysis of Pre-adjusted Loss Cost, Frequency or Severity in Tweedie Models". In: *Casualty Actuarial Society E-Forum*. 2010, pp. 1–13.

[24] Hiroshi Shono. "Application of the Tweedie Distribution to Zero-catch Data in CPUE Analysis". In: *Fisheries Research* 93.1 (2008), pp. 154–162.

[25] Gordon Smyth. *Package 'statmod'*. R Documentation. Package for R programming. June 2017.

[26] MCK Tweedie. "An Index which Distinguishes between Some Important Exponential Families". In: *Statistics: Applications and new directions: Proc. Indian statistical institute golden Jubilee International conference*. Vol. 579. 1984, p. 604.

[27] MCK Tweedie. "Functions of a Statistical Variate with Given Means, with Special Reference to Laplacian Distributions". In: *Mathematical Proceedings of the Cambridge Philosophical Society*. Vol. 43. 1. Cambridge University Press. 1947, pp. 41–49.

[28] Hans Van Vliet. *Software engineering: principles and practice*. Vol. 3. Wiley New York, 1993.

[29] James Owen Weatherall. *The Physics of Wall Street: A Brief History of Predicting the Unpredictable*. Houghton Mifflin Harcourt, 2013.

[30] Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2009. ISBN: 978-0-387-98140-6. URL: http://ggplot2.org.

[31] Alicja Wolny-Dominiak and Michal Trzesiok. *insuranceData: A Collection of Insurance Datasets Useful in Risk Classification in Non-life Insurance.* R package version 1.0. 2014. URL: https://CRAN.R-project.org/package=insuranceData.