Chapter 1
Introduction

1.1 The Statistical Modeling Cycle

We consider statistical modeling of insurance problems. This comprises the process of data collection, data analysis and statistical model building to forecast insured events that (may) happen in the future. This problem is at the very heart of statistics and statistical modeling. Our goal here is to present and provide the statistical tools that are useful in daily actuarial practice, in particular, we aim at describing the mathematical foundation behind these statistical concepts and how they can be applied. Statistical modeling has a wide range of applications, and, depending on the application, the theoretical aspects may be weighted differently. In insurance pricing we are mainly interested in optimal predictions, whereas economists often use statistical tools to explain observations, and in medical fields one is interested in causal effects that medications have on patients. Therefore, statistical theory is wide ranging, and one should always keep the corresponding application in mind. Shmueli [338] nicely discusses the difference between prediction and explanation; our focus here is mainly on prediction.

Box–Jenkins [49] and McCullagh–Nelder [265] distinguish three processes in statistical modeling: (i) model identification/selection, (ii) estimation, and (iii) prediction. In our statistical modeling cycle these three points are slightly modified and extended:

(1) Data collection, cleaning and pre-processing:
This item takes at least 80% of the total time in statistical modeling. It includes exploratory data analysis, data visualization and data pre-processing. This part of the modeling cycle does not seem to be very scientific, however, it is a highly important step because only extended data analysis allows the modeler to fully understand the data. Based on this knowledge the modeler can formulate her/his research question, her/his model, etc.

M. V. Wüthrich, M. Merz, *Statistical Foundations of Actuarial Learning and its Applications*, Springer Actuarial, https://doi.org/10.1007/978-3-031-12409-9_1

(2) Selection of a model class:
Based on the knowledge collected in the first item, the modeler has to select a suitable model class that is able to answer her/his research question. This model class can be in the sense of a data model (proper stochastic model), but it can also be an algorithmic model; we refer to the discussion on the "two modeling cultures" by Breiman [53].
(3) Choice of an objective function:
Once the modeler has specified a model class, she/he needs to define a decision rule how a particular member of the model class is selected for the collected data. Often this is in terms of an objective function, e.g., a scoring rule or a loss function that quantifies misspecification.
(4) Solving a (non-convex) optimization problem:
Once the first three items are completed, one is left with an optimization problem that tries to find the best model within the selected model class w.r.t. the given objective function and the collected data. In simple cases this optimization problem is a convex minimization problem for which numerical tools are in place. In more complex cases the optimization problem is neither convex nor concave, and the 'best' solution can often not be found explicitly. In that case, also the meaning of solution needs to be discussed.
(5) Model validation:
In the final/next step, the selected and fitted model needs to be validated. That is, does the model fit to the data, does it serve at predicting new data, does it answer the research question adequately, is there any better model/process choice, etc.?
(6) Possibly go back to (1):
If the answers in item (5) are not satisfactory, one typically goes back to (1). For instance, data pre-processing needs to be done differently, etc.

Especially, the two modeling cultures discussion of Breiman [53], after the turn of the millennium, has shaken up the statistical community. Having predictive performance as the main criterion, the data modeling culture has gradually shifted to the algorithmic culture, where the model itself plays a secondary role as long as the prediction is accurate. The latter is often in the form of a point predictor which can come from an algorithm. Lifting this discussion to a more scientific level, providing prediction uncertainty will slowly merge the two modeling cultures. There is an other interesting discussion by Efron [116] on prediction, estimation (of model parameters) and attribution (predictor selection), that is very much at the core of statistical modeling. In these notes we want to especially emphasize the one modeling culture view of Yu–Barter [397] who expect the two modeling cultures of Breiman [53] to merge much closer than one would expect. Our goal is to demonstrate how all these different techniques and views can be seen as a unified modeling framework.

Concluding, the purpose of these notes is to discuss and illustrate how the different statistical techniques from the data modeling culture and the algorithmic modeling culture can be combined to solve actuarial questions in the best possible way. The main emphasis in this discussion lies on the statistical modeling tools,

and we present these tools along with actuarial examples. In actuarial practice one often distinguishes between life and general insurance. This distinction is done for good reasons. There are legislative reasons that require to legally separate life from general insurance business, but there are also modeling reasons, because insurance products in life and general insurance can have rather different features. In this book, we do not make this distinction because the statistical methods presented here can be useful in both branches of insurance, and we are going to consider life and general insurance examples, e.g., the former considering mortality forecasting and the latter aiming at insurance claims prediction for pricing.

1.2 Preliminaries on Probability Theory

The modern axiomatic foundation of probability theory was introduced in 1933 by the famous mathematician Kolmogoroff [221] in his book called "Grundbegriffe der Wahrscheinlichkeitsrechnung". We give a brief introduction to probability theory and random variables; this introduction follows the lecture notes [387]. Throughout we assume to work on a sufficiently rich probability space $(\Omega, \mathcal{A}, \mathbb{P})$, meaning that this probability space should be able to carry all objects that we study. We denote (real-valued) random variables on this probability space by capital letters $Y, Z, \ldots$, and random vectors use boldface capital letters, e.g., we have a random vector $\boldsymbol{Y} = (Y_1, \ldots, Y_q)^\top$ of dimension $q \in \mathbb{N}$, where each component Y_k, $1 \le k \le q$, is a random variable. Random variables Y are characterized by (cumulative) distribution functions[1] $F : \mathbb{R} \to [0, 1]$, for $y \in \mathbb{R}$

$$F(y) = \mathbb{P}[Y \le y],$$

being the probability of the event that Y has a realization of less or equal to y. We write $Y \sim F$ for Y having distribution function F. Similarly random vectors $\boldsymbol{Y} \sim F$ are characterized by (cumulative) distribution functions $F : \mathbb{R}^q \to [0, 1]$ with

$$F(\boldsymbol{y}) = \mathbb{P}\left[Y_1 \le y_1, \ldots, Y_q \le y_q\right] \qquad \text{for } \boldsymbol{y} = (y_1, \ldots, y_q)^\top \in \mathbb{R}^q.$$

In insurance modeling, there are two important types of random variables, namely, discrete random variables and absolutely continuous random variables:

- The distribution function F of a discrete random variable Y is a step function with countably many steps in discrete points $k \in \mathfrak{N} \subset \mathbb{R}$. A discrete random variable has probability weights in these discrete points

$$f(k) = \mathbb{P}[Y = k] > 0 \qquad \text{for } k \in \mathfrak{N},$$

[1] Cumulative distribution functions F are right-continuous, non-decreasing with $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to \infty} F(x) = 1$.

satisfying $\sum_{k\in\mathfrak{N}} f(k) = 1$. If $\mathfrak{N} \subseteq \mathbb{N}_0$, the integer-valued random variable Y is called count random variable. Count random variables are used to model the number of claims in insurance. A similar situation occurs if Y models nominal outcomes, for instance, if Y models gender with female being encoded by 0 and male being encoded by 1, then $f(0)$ is the probability weight of having a female and $f(1) = 1 - f(0)$ the probability weight of having a male; in this case we identify the finite set $\mathfrak{N} = \{0, 1\} = \{\text{female, male}\}$.

- A random variable $Y \sim F$ is said to be absolutely continuous[2] if there exists a non-negative (measurable) function f, called density of Y, such that

$$F(y) = \int_{-\infty}^{y} f(x)\, dx \qquad \text{for all } y \in \mathbb{R}.$$

  In that case we equivalently write $Y \sim f$ and $Y \sim F$. Absolutely continuous random variables are often used to model claim sizes in insurance.

More generally speaking, discrete and absolutely continuous random variables have densities $f(\cdot)$ w.r.t. a σ-finite measure ν on $\mathbb{R}$. In the former case, this σ-finite measure ν is the counting measure on $\mathfrak{N} \subset \mathbb{R}$, and in the latter case it is the Lebesgue measure on $\mathbb{R}$. In actuarial science we also consider mixed cases, for instance, Tweedie's compound Poisson random variable is absolutely continuous on $(0, \infty)$ having an additional point mass in 0; this model will be studied in Sect. 2.2.3, below.

Choose a random variable $Y \sim F$ and a measurable function $h : \mathbb{R} \to \mathbb{R}$. The expected value of $h(Y)$ is defined by (upon existence)

$$\mathbb{E}[h(Y)] = \int_{\mathbb{R}} h(y)\, dF(y).$$

We mainly focus on the following important examples of function h:

- expected value, mean or first moment of $Y \sim F$: for $h(y) = y$

$$\mu = \mathbb{E}[Y] = \int_{\mathbb{R}} y\, dF(y);$$

- k-th moment of $Y \sim F$ for $k \in \mathbb{N}$: for $h(y) = y^k$

$$\mathbb{E}\left[Y^k\right] = \int_{\mathbb{R}} y^k\, dF(y);$$

[2] Absolutely continuous is a stronger property than continuous.

- moment generating function of $Y \sim F$ in $r \in \mathbb{R}$: for $h(y) = e^{ry}$

$$M_Y(r) = \mathbb{E}\left[e^{rY}\right] = \int_{\mathbb{R}} e^{ry}\, dF(y);$$

always subject to existence.

The moment generating function $M_Y(\cdot)$ is sufficient for identifying distribution functions of random variables Y. The following statements are elementary and their proofs are based on Section 30 of Billingsley [34], for more details we also refer to Chapter 1 in the lecture notes [387]. Assume that the moment generating function of $Y \sim F$ has a strictly positive radius of convergence $\rho_0 > 0$ around the origin implying that $M_Y(r) < \infty$ for all $r \in (-\rho_0, \rho_0)$. In this case we can write $M_Y(r)$ as a power series expansion

$$M_Y(r) = \sum_{k=0}^{\infty} \frac{r^k}{k!}\, \mathbb{E}\left[Y^k\right] \qquad \text{for all } r \in (-\rho_0, \rho_0).$$

As a consequence we can differentiate $M_Y(\cdot)$ in the open interval $(-\rho_0, \rho_0)$ arbitrarily often, term by term under the sum. The derivatives in $r = 0$ provide the k-th moments (which all exist and are finite)

$$\frac{d^k}{dr^k}\, M_Y(r)|_{r=0} = \mathbb{E}\left[Y^k\right] \qquad \text{for all } k \in \mathbb{N}_0. \tag{1.1}$$

In particular, in this case we immediately know that all moments of Y exist, and these moments completely determine the moment generating function M_Y of Y. Another consequence is that for a random variable Y, whose moment generating function M_Y has a strictly positive radius of convergence around the origin, the distribution function F is fully determined by this moment generating function. That is, if we have two such random variables Y_1 and Y_2 with $M_{Y_1}(r) = M_{Y_2}(r)$ for all $r \in (-r_0, r_0)$, for some $r_0 > 0$, then $Y_1 \overset{(\mathrm{d})}{=} Y_2$.[3] Thus, these two random variables have the same distribution function. This statement carries over to the limit, i.e., if we have a sequence of random variables $(Y_n)_n$ whose moment generating functions converge on a common interval $(-r_0, r_0)$, for some $r_0 > 0$, to the moment generating function of Y, also being finite on $(-r_0, r_0)$, then $(Y_n)_n$ converges in distribution to Y; such an argument is used to prove the central limit theorem (CLT).

[3] The notation $Y_1 \overset{(\mathrm{d})}{=} Y_2$ is generally used for equality in distribution meaning that Y_1 and Y_2 have the same distribution function.

In insurance, we often deal with so-called positive random variables Y, meaning that $Y \geq 0$, almost surely (a.s.). In that case, the statements about moment generating functions and distributions hold true without the assumption of having a positive radius of convergence around the origin, see Theorem 22.2 in Billingsley [34]. Note that for positive random variables the moment generating function $M_Y(r)$ exists for all $r \leq 0$.

Existence of the moment generating function $M_Y(r)$ for some positive $r > 0$ can also be interpreted as having a light-tailed distribution function. Observe that if $M_Y(r)$ exists for some positive $r > 0$, then we can choose $s \in (0, r)$ and Chebychev's inequality gives us (we assume $Y \geq 0$, a.s., here)

$$\mathbb{P}[Y > y] = \mathbb{P}\big[\exp\{sY\} > \exp\{sy\}\big] \leq \exp\{-sy\} M_Y(s). \tag{1.2}$$

The latter tells us that the survival function $1 - F(y) = \mathbb{P}[Y > y]$ decays exponentially for $y \to \infty$. Heavy-tailed distribution functions do not have this property, but the survival function decays slower than exponentially as $y \to \infty$. This slower decay of the survival function is the case for so-called subexponential distribution functions (an example is the log-normal distribution, we refer to Rolski et al. [320]) and for regularly varying survival functions (an example is the Pareto distribution). Regularly varying survival functions $1 - F$ have the property

$$\lim_{y \to \infty} \frac{1 - F(ty)}{1 - F(y)} = t^{-\beta} \qquad \text{for all } t > 0 \text{ and some } \beta > 0. \tag{1.3}$$

These distribution functions have a polynomial tail (power tail) with tail index $\beta > 0$. In particular, if a positively supported distribution function F has a regularly varying survival function with tail index $\beta > 0$, then this distribution function is also subexponential, see Theorem 2.5.5 in Rolski et al. [320].

We are not going to specifically focus on heavy-tailed distribution functions, here, but we will explain how light-tailed random variables can be transformed to enjoy heavy-tailed properties. In these notes, we are mainly interested in studying different aspects of regression modeling. Regression modeling requires numerous observations to be able to successfully fit these models to the data. By definition, large claims are scarce, as they live in the tail of the distribution function and, thus, correspond to rare events. Therefore, it is often not possible to employ a regression model for scarce tail events. For this reason, extreme value analysis only plays a marginal role in these notes, though, it has a significant impact on insurance prices. For more on extreme value theory we refer to the relevant literature, see, e.g., Embrechts et al. [121], Rolski et al. [320], Mikosch [277] and Albrecher et al. [7].

1.3 Lab: Exploratory Data Analysis

Our theory is going to be supported by several data examples. These examples are mostly based on publicly available data. The different data sets are described in detail in Chap. 13. We highly recommend the reader to use these data sets to gain her/his own modeling experience.

We describe some tools here that allow for a descriptive and exploratory analysis of the available data; exploratory data analysis has been introduced and promoted by Tukey [357]. We consider the observed claim sizes of the Swedish motorcycle data set described in Sect. 13.2. This data set consists of 656 (positive) claim amounts y_i, $1 \le i \le n = 656$. These claim amounts are illustrated in the boxplots of Fig. 1.1.

Typically in insurance, there are large claims that dominate the picture, see Fig. 1.1 (lhs). This results in right-skewed distribution functions, and such data is better illustrated on the log scale, see Fig. 1.1 (rhs). The latter, of course, assumes that all claims are strictly positive.

Figure 1.2 (lhs) shows the empirical distribution function of the observations y_i, $1 \le i \le n$, which is obtained by

$$\widehat{F}_n(y) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{y_i \le y\}} \qquad \text{for } y \in \mathbb{R}.$$

If this data set has been generated by i.i.d. random variables, then the Glivenko–Cantelli theorem [64, 159] tells us that this empirical distribution function $\widehat{F}_n$ converges uniformly to the (true) data generating distribution function, a.s., as the number n of observations converges to infinity, see Theorem 20.6 in Billingsley [34].

Figure 1.2 (rhs) shows the empirical density of the observations y_i, $1 \le i \le n$. This empirical density is obtained by considering a kernel smoother of a given

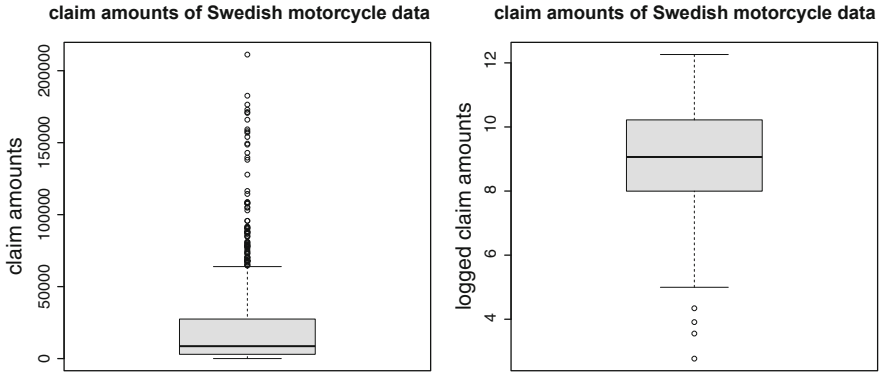

Fig. 1.1 Boxplot of the claim amounts of the Swedish motorcycle data set: (lhs) on the original scale and (rhs) on the log scale

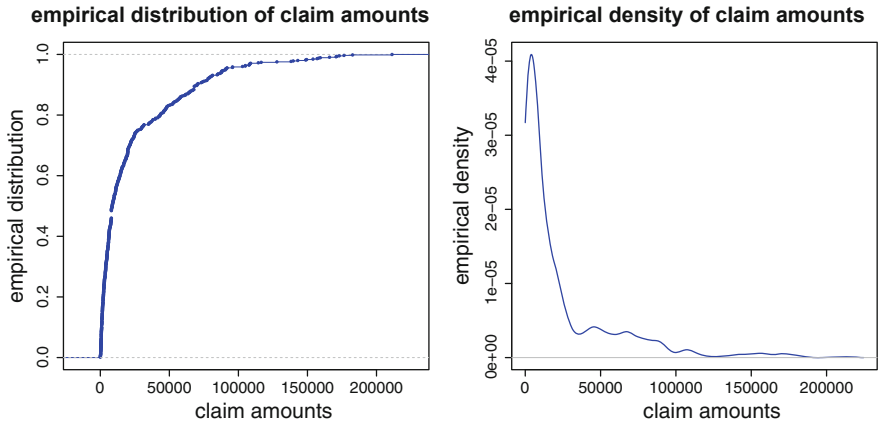

Fig. 1.2 (lhs) Empirical distribution and (rhs) empirical density of the observed claim amounts y_i, $1 \le i \le n$

bandwidth around each observation y_i. The standard choice is the Gaussian kernel, with the bandwidth determining the variance parameter $\sigma^2 > 0$ of the Gaussian density,

$$y \mapsto \widehat{f_n}(y) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{1}{2} \frac{(y - y_i)^2}{\sigma^2} \right\}.$$

From the graph in Fig. 1.2 (rhs) we observe that the main body of the claim sizes is below an amount of 50'000, but the biggest claim exceeds 200'000. The latter motivates to study heavy-tailedness of the claim size data. Therefore, one usually benchmarks with a distribution function F that has a regularly varying survival function with a tail index $\beta > 0$, see (1.3). Asymptotically a regularly varying survival function behaves as $y^{-\beta}$; for this reason the log-log plot is a popular tool to identify regularly varying tails. The log-log plot of a distribution function F is obtained by considering

$$y > 0 \mapsto (\log y, \log(1 - F(y))) \in \mathbb{R}^2.$$

Figure 1.3 gives the log-log plot of the empirical distribution function $\widehat{F}_n$. If this plot looks asymptotically (for $y \to \infty$) like a straight line with a negative slope $-\beta$, then the data shows heavy-tailedness in the sense of regular variation. Such data cannot be modeled by a distribution function for which the moment generating function $M_Y(r)$ exists for some positive $r > 0$, see (1.2). Figure 1.3 does not suggest a regularly varying tail as we do not see an obvious asymptotic straight line for increasing claim sizes.

These graphs give us a first indication what the claim size data is about. Later on we are going to introduce explanatory variables that describe the insurance

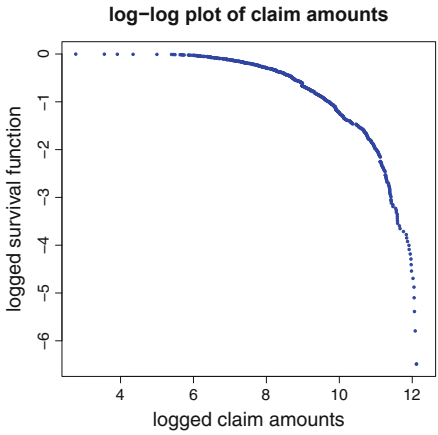

Fig. 1.3 Log-log plot of the empirical distribution function $\widehat{F}_n$

policyholders behind these claims. These explanatory variables characterize the policyholder and the general goal is to get a better description of the claim sizes as a function of these explanatory variables, e.g., older policyholders may cause larger claims than younger ones, etc. Such patterns are called *systematic effects* that can be explained by explanatory variables.

1.4 Outline of This Book

This book has eleven chapters (including the present one), and it has two appendices. We briefly describe the contents of these chapters and appendices.

In Chap. 2 we introduce and discuss the exponential family (EF) and the exponential dispersion family (EDF). The EF and the EDF are by far the most important classes of distribution functions for regression modeling. They include, among others, the Gaussian, the binomial, the Poisson, the gamma, the inverse Gaussian and Tweedie's models. We introduce these families of distribution functions, discuss their properties and provide several examples. Moreover, we introduce the Kullback–Leibler (KL) divergence and the Bregman divergence, which are important tools in model evaluation.

Chapter 3 is on classical statistical decision theory. This chapter is important for historical reasons, but it also provides the right mathematical grounding and intuition for more modern tools from data science and machine learning. In particular, we discuss maximum likelihood estimation (MLE), unbiasedness, consistency and asymptotic normality of MLEs in this chapter.

Chapter 4 is the core theoretical chapter on predictive modeling and forecast evaluation. The main problem in actuarial modeling is to forecast and price future claims. For this, we build predictive models, and this chapter deals with assessing and ranking these predictive models. We therefore introduce the mean squared

error of prediction (MSEP) and, more generally, the generalization loss (GL) to assess predictive models. This chapter is complemented by a more decision-theoretic approach to forecast evaluation, it discusses deviance losses, proper scoring, elicitability, forecast dominance, cross-validation, Akaike's information criterion (AIC) and we give an introduction to the bootstrap simulation method.

Chapter 5 discusses state-of-the-art statistical modeling in insurance which is the generalized linear model (GLM). We discuss GLMs in the light of claim count and claim size modeling, we present feature engineering, model fitting, model selection, over-dispersion, zero-inflated claim counts problems, double GLMs, and insurance-specific issues such as the balance property for having unbiasedness.

Chapter 6 summarizes some techniques that use Bayes' theorem. These are classical Bayesian statistical models, e.g., using the Markov chain Monte Carlo (MCMC) method for model fitting. This chapter discusses regularization of regression models such as ridge and LASSO regularization, which has a Bayesian interpretation, and it concerns the Expectation-Maximization (EM) algorithm. The EM algorithm is a general purpose tool that can handle incomplete data settings. We illustrate this for different examples coming from mixture distributions, censored and truncated claims data.

The core of this book are deep learning methods and neural networks. Chapter 7 considers deep feed-forward neural (FN) networks. We introduce the generic architecture of deep FN networks, and we discuss universality theorems of FN networks. We present network fitting, back-propagation, embedding layers for categorical variables and insurance-specific issues such as the balance property in network fitting and network ensembling to reduce model uncertainty. This chapter is complemented by many examples on non-life insurance pricing, but also on mortality modeling, as well as tools that help to explain deep FN network regression results.

Chapters 8 and 9 consider recurrent neural (RN) networks and convolutional neural (CN) networks. These are special network architectures that are useful for time-series and spatial data modeling, e.g., applied to image recognition problems. Time-series and images have a natural topology, and RN and CN networks try to benefit from this additional structure (over tabular data). We introduce these network architectures and provide insurance-relevant examples.

Chapter 10 discusses natural language processing (NLP) which deals with regression modeling of non-tabular or unstructured text data. We explain how words can be embedded into low-dimension spaces that serve as numerical word encodings. These can then be used for text recognition, either using RN networks or attention layers. We give an example where we aim at predicting claim perils from claim descriptions.

Chapter 11 is a selection of different topics. We mention forecasting under model uncertainty, deep quantile regression, deep composite regression or the LocalGLMnet which is an interpretable FN network architecture. Moreover, we provide a bootstrap example to assess prediction uncertainty, and we discuss mixture density networks.

Chapter 12 (Appendix A) is a technical chapter that discusses universality theorems for networks and sieve estimators, which are useful for studying asymptotic normality within a network framework. Chapter 13 (Appendix B) illustrates the data used in this book.

Finally, we remark that the book is written in a typical mathematical style using the structure of Lemmas, Theorems, etc. Results and statements which are particularly important for applications are highlighted with gray boxes.