Lecture notes
ACTSC 632 - Data Science with Actuarial Applications

June 16, 2023

The objective of this course is to learn various (**computer-based) statistical learning methods** and consider their **applications in contexts of interest in actuarial science and finance**. This include tasks that actuaries have traditionally been involved in such as the pricing and risk assessment/management of a block of insurance policies.

ACTSC 632 has strong connections with other courses in the Master of Actuarial Science (MACTSC) program:

- ACTSC 612 & 622: to price life insurance products, mortality/survival models are needed;
- ACTSC 613 & 623: some statistical models introduced in these courses will be examined in more detail in an insurance context;
- ACTSC 625: more modern techniques (such as applications of generalized linear models for loss modelling purposes) will be examined for pricing non-life insurance policies.

The course will be organized into the following 4 modules:

Module 1: Survival models

Module 2: Generalized linear models (GLMs) for non-life insurance

Module 3: Classification

Module 4: Tree-based methods

A short introduction to modelling and learning will first be presented with a special emphasis on their applications in actuarial science.

0 Introduction to modelling and learning in Actuarial Science

0.1 Introduction

Actuaries need to build models for various quantities/processes that are random/stochastic (such as models for the future lifetime rv T_x in life contingency, loss severity and frequency models, interest rates, stock returns and many others) to quantify the risks inherent in a block of insurance business. In general, models are built by collecting appropriate data which are then used for inference and then prediction. The field of statistics has helped actuaries with these tasks for a very long time.

What about *data science*, *predictive modelling* and *predictive analytics*?

- *Data science*: converts data into knowledge using (i) appropriate database management tools for transforming and organizing data; (ii) statistics and machine learning and (iii) distributed and parallel systems for providing computational infrastructure;
- *Predictive modelling*: involves the use of data to forecast future events. It relies on capturing relationships between explanatory variables and the predicted variables from past occurrences and exploiting them to predict future outcomes (if similar dynamics are expected in the future);
- *Predictive analytics*: use of data, statistical algorithms and machine learning techniques to identify the likelihood of future outcomes based on historical data.

So it is clear that *predictive modelling* describes (in a modern way) what actuaries do. Data science is the field giving them modern tools to reach these goals (i.e., capturing relationships, predicting future outcomes,...).

We now introduce key concepts and terminology which will be called upon throughout this course.

- Output variable (also known as **response** or dependent variable): variable of interest we aim to predict or model, often denoted by Y.
- Input variables (also known as **predictors**, independent variables or features): variables used to explain or predict the output variable. For a set of p predictors, we write $\mathbf{X} = (X_1, X_2, ..., X_p)$.

We assume that there exists some relationship between $\mathbf{X}$ and Y, which we write very generally as

$$Y = f(\mathbf{X}) + \varepsilon,$$

where f is fixed but an unknown function linking the predictors to the response, and ε is a random error term. The field of statistical learning refers to a set of approaches for estimating f which is crucial for prediction and inference purposes.

0.2 Prediction

Once an functional approximation $\widehat{f}$ for f is identified, we can use it to predict the response for a given set of predictors, i.e.

$$\widehat{Y} = \widehat{f}(\mathbf{X}).$$

Example 1 *For a non-life insurance policy, the insurer may want to estimate the total claim amount from a policyholder for the coming year given some characteristics of the policyholder (e.g., age, gender, location of principal residence,...). In this case,*

- Y *corresponds to the total claim amount for the policyholder;*
- $\mathbf{X}$ *corresponds to the set of the policyholder's characteristics deemed relevant to explain/predict* Y*.*

It is important to measure the accuracy of $\widehat{Y}$ as a predictor of Y to determine how well the model performs and how confident we should be with the prediction. Ideally, we want $\widehat{Y}$ to be close to Y. Often, we measure this accuracy by the quadratic distance $\left(Y-\widehat{Y}\right)^2$ via the mean square error

$$\begin{aligned} MSE\left(\widehat{Y}\right) &= \mathbb{E}\left[\left(Y-\widehat{Y}\right)^2\right] \\ &= \mathbb{E}\left[\left(Y-\widehat{f}(\mathbf{X})\right)^2\right]. \end{aligned}$$

Two components of the error are:

- *reducible error*: introduced by using $\widehat{f}$ instead of f
- *irreducible error*: introduced by ε.

Using the quadratic variation, we can write

$$\begin{aligned} MSE\left(\widehat{Y}\right) &= \mathbb{E}\left[\left(Y-\widehat{f}(\mathbf{X})\right)^2\right] \\ &= \mathbb{E}\left[\left(f(\mathbf{X})+\varepsilon-\widehat{f}(\mathbf{X})\right)^2\right] \\ &= \underbrace{\left(f(\mathbf{X})-\widehat{f}(\mathbf{X})\right)^2}_{\text{reducible}} + \underbrace{\mathbb{E}\left[\varepsilon^2\right]}_{\text{irreducible}} + \underbrace{2\mathbb{E}\left[\varepsilon\left(f(\mathbf{X})-\widehat{f}(\mathbf{X})\right)\right]}_{\substack{=0 \\ \text{by independence} \\ \text{and the fact that } \mathbb{E}[\varepsilon]=0}} \end{aligned}$$

0.3 Inference

The goal is to understand how Y is related to $\mathbf{X}$, i.e. how Y changes as a function of the predictor set $\mathbf{X}=(X_1,...,X_p)$? Examples of more specific questions we may want to answer:

- which predictors have the greatest influence (i.e., stronger association) on the response?
- what is the relationship between the response and each predictor? Is this relationship linear (or not)?

Note that the best choice of $\widehat{f}$ often depends on the goal of the exercise, i.e. inference vs prediction. For example, a linear model (i.e., $\widehat{f}(\mathbf{X}) = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p$) may be best for inference (very intuitive), but a more complicated $\widehat{f}$ might be better for prediction. One shall strike a good balance between these two important factors.

More generally, there is always a danger of overfitting the model (e.g., by including too many predictors in the model). Indeed, a very complicated model can provide a false sense of accuracy to the observed data, but may do poorly to predict future observations. There are well known caveats in overfitting the model. In general, one should aim to find as simple a model as possible that captures the important relationships between the response and the predictors. This is known as the principle of *parsimony*.

0.4 How do we estimate f?

In general, it is advisable to separate the data set in hand into two disjoint sets:

- *training set:* use to build the model
- *test set*: use to check how the model does on data points not used to build the model

In this case, the verification of the fit through the test set is more indicative of what to expect from future predictions.

More concretely, consider a data set $\{(\mathbf{x}_1, y_1), ..., (\mathbf{x}_{n+m}, y_{n+m})\}$ where $\mathbf{x}_i = (x_{i1}, x_{i2}, ..., x_{ip})$. We note that p stands for the number of predictors and $n + m$ corresponds to the total number of points (observations) in the data set. One can proceed by randomly choosing (without replacement) n data points from the data set $\{(\mathbf{x}_1, y_1), ..., (\mathbf{x}_{n+m}, y_{n+m})\}$ to form the training set. The remaining m data points will form the test set. Usually, n is chosen much larger than m (i.e., a majority of the data points are used to build the model).

For illustration purposes, we suppose for the rest of this section that

- training set: $\{(\mathbf{x}_1, y_1), ..., (\mathbf{x}_n, y_n)\}$
- test set: $\{(\mathbf{x}_{n+1}, y_{n+1}), ..., (\mathbf{x}_{n+m}, y_{n+m})\}$

Remark 2 *It is worth mentioning that in some cases though, the whole data set is used for both training and testing purposes. In that case, we would expect the model to do well to predict the data points in the data set as these same points were also used to build the model. We will revisit this idea later in the course.*

To estimate f using the training set, we have the option to choose either a parametric or a non-parametric approach.

- **Parametric methods**: we make an assumption that f has a certain parametric form and choose a method to estimate these parameters (e.g., MLE). For instance, we may assume that f is linear, i.e.

$$f(\mathbf{x}) = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p,$$

and then use least square regression to estimate $\beta_0, \beta_1, ..., \beta_p$.

There are advantages and disadvantages to proceed in this manner. Among others, the specification of the form of f greatly simplifies the problem (reducing the problem to estimating the model's parameters), but the chosen form may be far from the true one leading to the model's poor predictive power.

- **Non-parametric methods**: we do not make an explicit assumption about the functional form of f. As such, the method can fit a wider range of possible shapes for f. However, a much larger training set (i.e., with more observations) is required to get a more accurate approximation for f.

Generally speaking, we need to distinguish between cases where the response variable is *quantitative* or *qualitative* (the latter is also known as *categorical* - where the categories are arbitrarily assigned). Typically, we refer to *regression* for problems when the response is quantitative and to *classification* when the response is qualitative.

As an illustration, a common classification method is the so-called Bayes classifier. For Bayes classifier, for a set of predictor $\mathbf{x}$, we assign (or predict) the response to be from the category j that maximizes

$$\mathbb{P}\left(Y = j \,|\mathbf{X} = \mathbf{x}\right),$$

among all possible response type j. In other words, the classification is done by assigning a response to the category that gives the largest probability of generating this outcome (for a given set of predictors). For the predictor $\mathbf{x}$, the error rate of Bayes classifier is

$$1 - \max_j \mathbb{P}\left(Y = j \,|\mathbf{X} = \mathbf{x}\right),$$

and the overall error rate is

$$1 - \mathbb{E}\left[\max_j \mathbb{P}\left(Y = j \,|\mathbf{X}\right)\right].$$

Note that the above expectation is taken over the whole range of values for the predictors $\mathbf{X}$.

There exists other classifiers (such as the K-nearest neighbors) that will be introduced in Module 3.

0.5 Assessing the model accuracy

In order to decide which statistical learning methods to privilege for a given problem (and underlying data set), we clearly need a way to assess how good $\widehat{f}$ is in relation to the true, unknown f. We again distinguish between cases where the response is quantitative and qualitative.

- For a quantitative response, a common measure of fit is the observed mean-squared error (MSE) given by

$$MSE = \frac{1}{n+m}\sum_{i=1}^{n+m}\left(y_i - \widehat{f}\left(\mathbf{x}_i\right)\right)^2.$$

As mentioned above, it is common to break the data set into two disjoint sets (known as the *training* data set and the *test* data set). The training data set is used to construct $\widehat{f}$ (by minimizing the training MSE)

$$\text{training } MSE = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \widehat{f}\left(\mathbf{x}_i\right)\right)^2,$$

while the test data is used to assess the fit via the test MSE

$$\text{test } MSE = \frac{1}{m}\sum_{i=1}^{m}\left(y_{n+i} - \widehat{f}\left(\mathbf{x}_{n+i}\right)\right)^2,$$

A model that overfits the training set will generally yield a small training MSE, but likely result in much larger test MSE.

- For a qualitative response, the MSE cannot be used as the categories are arbitrarily assigned and cannot be ordered. Instead, we use the classification error defined as

$$\text{classification error} = \frac{\sum_{i=1}^{n+m} 1_{\{y_i \neq \widehat{y}_i\}}}{n+m}$$

with a general goal to minimize this error. When the data set is divided into a training and a test set, we build the model by minimizing the training classification error

$$\text{training classification error} = \frac{\sum_{i=1}^{n} 1_{\{y_i \neq \widehat{y}_i\}}}{n}$$

and measure its performance against the test set via the test classification error

$$\text{test classification error} = \frac{\sum_{i=1}^{m} 1_{\{y_{n+i} \neq \widehat{y}_{n+i}\}}}{m}.$$

Once again, a model that overfits the training set will yield a small training classification error, but likely result in much larger test classification error.

0.6 Cross-validation

What is cross-validation? It consists in methods to separate the data set into a training set and a test set. It does it by holding out a subset of the data from the fitting process and then assessing the fit via those held out observations. Next, we discuss 3 cross-validation approaches.

(1) The Validation Set approach

It consists in randomly splitting the available data set into two parts: training set and test set (method described above). The test error rate can be highly variable (depends on the split).

(2) Leave-one-out cross validation (LOOCV)

For a data set of size $n+m$, train the model with a total of $(n+m-1)$ observations and leaves the remaining observation to measure the error rate. Repeat this process $n+m$ times leaving a different observation out of the training process. Average out the errors (MSE or classification error) to estimate the test error rate. It leads to a more robust approach than the method described in (1).

(3) K-fold cross validation

For a data set of size $n+m$, we randomly split the data into K disjoint groups of approximately equal size. At each iteration, we leave one group of data out of the model train step and use this group to assess the model error. We assess the test MSE by averaging the error rates over the K iterations. This method

is less intensive than LOOCV as LOOCV can be viewed as a special case of K-fold cross validation where $K = n + m$. In practice, it is common to choose $K = 5$ or $K = 10$ when using the K-cross validation method.

We will look at applications of these cross-validation methods later on in the course.

1 Survival models

As extensively discussed in ACTSC 612 & 622, the pricing, reserving and profit testing exercises of life insurance policies require the use of mortality/survival models. So far, the mortality/survival model was assumed as given. In general, this is not the case and special care needs to be given on the construction, parameter estimation and overall fit of the chosen model to the mortality data. Note that this task needs to be handled with care as past data may not always be representation of future patterns. Indeed, it is well documented that human mortality has greatly improved over the last few decades (a concept known as longevity risk).

As such, the main goal of this module is to learn how to construct mortality/survival models to be used for actuarial applications. The discussion aims to provide insight on how to use data to build these models. This will be achieved by going over the following 5 topics:

- Estimation of lifetime distribution using nonparametric models (e.g., Kaplan-Meier and Nelson-Aalen estimators)
- Estimation of lifetime distribution using semi-parametric models (e.g., Cox proportational hazard model)
- Estimation of multiple-state models
- Poisson **(skip)** and binomial mortality models
- Graduation: steps required to construct a statistically sound and reasonably smooth mortality/survival model **(skip)**

We will not cover the concept of longevity risk in this course.

1.1 Preliminaries - notation and review

Initially, we are trying to model T_x, the future lifetime of a life aged x. As we know, this is the pivotal rv in all life insurance calculations. For simplicity, we use the notation $T := T_0$, and denote by ω the limiting age, i.e, $T \in (0, \omega]$.

Distributional quantities of T_x

- distribution function: $F_x(t) = \mathbb{P}(T_x \leq t) = {}_tq_x$ where
$$F_x(t) = \frac{F_0(x+t) - F_0(x)}{1 - F_0(x)}$$
- survival function: $S_x(t) = \mathbb{P}(T_x > t) = {}_tp_x$ where
$$S_x(t) = \frac{S_0(x+t)}{S_0(x)}.$$
- density function (if it exists):
$$f_x(t) = \frac{\mathrm{d}}{\mathrm{d}t} F_x(t)$$

- force of mortality:

$$\mu_x(t) := \mu_{x+t} = \frac{f_x(t)}{1 - F_x(t)} = \frac{f_x(t)}{{}_tp_x}$$

As we know, the force of mortality is a fundamental quantity for mortality/survival models. Among other relationships, this can be shown that

$${}_tp_x = e^{-\int_0^t \mu_{x+s} \mathrm{d}s},$$

and

$${}_tq_x = \int_0^t {}_sp_x \mu_{x+s} \mathrm{d}s,$$

for $t \geq 0$.

1.2 Estimation of nonparametric lifetime models (no censoring)

The objective is to use mortality data to build an estimate of $\{S_0(t), t \geq 0\}$ (or equivalently of $\{\mu_t, t \geq 0\}$). We do not want to choose a parameterized family of distributions to do so (such as assuming that T_x has an exponential distribution).

Remark 3 *For simplicity, we drop the argument* 0 *to the survival function* $S_0(t)$ *in what follows.*

1.2.1 Naive approach (empirical distribution)

We could observe a group of n newborns until they die. Let t_i be the realized time of death of the ith person in the sample. We can then construct the approximation

$$\begin{aligned} \widehat{S}_e(t) &= \frac{\text{nb of people still alive at time } t}{n} \\ &= \frac{\sum_{i=1}^n 1_{\{t_i > t\}}}{n}, \end{aligned} \tag{1}$$

where the indicator 1_A is 0 if A is true and 0 otherwise. Therefore, $\{\widehat{S}_e(t), t \geq 0\}$ will be a step function. This naive approach amounts to estimate $S(t)$ using the *empirical survival function.*

Example 4 *Assume we observe 10 lives, and times of death are 2.5, 23, 47, 66, 69, 72, 80, 82, 82, 90. Represent graphically the empirical survival function* $\{\widehat{S}_e(t), t \geq 0\}$ *based on this data.*

t	$\widehat{S}_e(t)$
$0 \leq t < 2.5$	1
$2.5 \leq t < 23$	0.9
$23 \leq t < 47$	0.8
$47 \leq t < 66$	0.7
$66 \leq t < 69$	0.6
$69 \leq t < 72$	0.5
$72 \leq t < 80$	0.4
$80 \leq t < 82$	0.3
$82 \leq t < 90$	0.1
$t \geq 90$	0

There are some shortcomings in using the naive approach:

- we need to observe all times of death, which may take a long time and/or be logistically challenging (as e.g., people may leave the study)
- the estimated survival function $\{\widehat{S}_e(t), t \geq 0\}$ is a step function, and therefore does not provide a direct estimator for $\{\mu_t, t \geq 0\}$.

As an alternative, we can instead model the *cumulative hazard rate function* given by

$$H(t) = -\ln S(t).$$

The following estimation method is based on the idea of directly estimating $H(t)$. As we will see, the cumulative hazard rate involves conditional probabilities (based on survival up to the present time), which suggests estimation may only require to know who is still present in the sample (particularly useful when we will introduce the concept of censoring in the next sub-section). Also, from a modelling standpoint, we may have some intuition for how the hazard rate function should behave, thus enabling diagnostic tools when analyzing data.

In summary, we can either estimate $S(t)$ directly or estimate $S(t)$ through its cumulative hazard rate $H(t)$.

1.2.2 Nelson-Aalen (NA) estimator

Notation: assume a sample $\{t_i\}_{i=1}^n$ of size n. Denote the k **distinct** values of this sample by $y_1 < y_2 < ... < y_k$. Let s_i be the number of observations with values y_i in the sample with $\sum_{i=1}^k s_i = n$.

We define the *risk set* r_i for a given value y_i as the number of observations at risk immediately before time y_i, i.e.

$$r_i = \sum_{j=i}^{k} s_j, \qquad i = 1, 2, ..., k.$$

Note that the sequence $\{r_i\}_{i=1}^k$ can be defined recursively as $r_1 = n$ and $r_i = r_{i-1} - s_{i-1}$ for $i = 2, 3, ..., k$.

Intuition: if the y_i's represent the distinct times of death, then we can think of r_i as the number of people who are at risk to die just before time y_i.

The Nelson-Aalen (NA) estimate of the cumulative hazard rate function is

$$\widehat{H}_{NA}(t) = \begin{cases} 0, & t < y_1 \\ \sum_{i=1}^{j-1} \frac{s_i}{r_i}, & y_{j-1} \leq t < y_j, \qquad j = 2, 3, ..., k, \\ \sum_{i=1}^{k} \frac{s_i}{r_i}, & t \geq y_k. \end{cases}$$

which leads to the following NA estimate of the survival function

$$\begin{aligned}\widehat{S}_{NA}(t) &= e^{-\widehat{H}_{NA}(t)} \\ &= \begin{cases} 1, & t < y_1 \\ \exp\left\{-\sum_{i=1}^{j-1} \frac{s_i}{r_i}\right\}, & y_{j-1} \le t < y_j, \qquad j = 2,3,...,k, \\ \exp\left\{-\sum_{i=1}^{k} \frac{s_i}{r_i}\right\}, & t \ge y_k. \end{cases}\end{aligned} \tag{2}$$

Notes:

- $\widehat{H}_{NA}(t)$ is a step function, so we cannot take the derivative to get an estimate for μ_t;
- The NA survival function $\widehat{S}_{NA}(t)$ in (2) can be compared with the empirical survival function given in (1) which, using the above notation, can be written as

$$\widehat{S}_e(t) = \begin{cases} 1, & t < y_1 \\ \frac{r_j}{n}, & y_{j-1} \le t < y_j, \qquad j = 2,3,...,k, \\ 0, & t \ge y_k. \end{cases}$$

Intuition behind the definition of the NA estimator

To explain where the NA estimator comes from, we first introduce the following definition:

Definition 5 *For T a discrete rv with mass points at $y_1 < y_2 < ... < y_k$, we define the* discrete hazard rate function *as*

$$\lambda_i = \mathbb{P}(T = y_i \,|T \ge y_i),$$

for $i = 1,...,k$. Its cumulative hazard rate function is defined as

$$H(t) = \sum_{i=1}^{j} \lambda_i,$$

for $t \in [y_j, y_{j+1})$.

When using a nonparametric approach to model T, we are making an implicit assumption that T is a discrete rv. Hence, we shall estimate all the λ_i's to come up with an estimate of $H(t)$. Intuitively, it should be clear that

$$\widehat{\lambda}_i = \frac{s_i}{r_i},$$

i.e. the probability of death at time y_i given that r_i individuals are at risk at time y_i is s_i/r_i. Hence, we can estimate $H(t)$ for $t \in [y_j, y_{j+1})$ by

$$\widehat{H}_{NA}(t) = \sum_{i=1}^{j} \frac{s_i}{r_i}.$$

Example 6 *For the data set* $2.5, 23, 47, 66, 69, 72, 80, 82, 82, 90$*, we have* $k = 9$ *and* $s_i = 1$ *for all* i *except* $s_8 = 2$*. More precisely we have*

i	y_i	r_i	s_i
1	2.5	10	1
2	23	9	1
3	47	8	1
4	66	7	1
5	69	6	1
6	72	5	1
7	80	4	1
8	82	3	2
9	90	1	1

Therefore,

$$\widehat{H}_{NA}(t) = \begin{cases} 0, & t < 2.5, \\ \frac{1}{10}, & 2.5 \le t < 23, \\ \frac{1}{10} + \frac{1}{9}, & 23 \le t < 47, \\ \vdots & \vdots \\ \frac{1}{10} + \frac{1}{9} + ... + \frac{1}{4}, & 80 \le t < 82, \\ \frac{1}{10} + \frac{1}{9} ... + \frac{1}{4} + \frac{2}{3}, & 82 \le t < 90, \\ \frac{1}{10} + \frac{1}{9} ... + \frac{1}{4} + \frac{2}{3} + \frac{1}{1}, & t \ge 90. \end{cases}$$

Next, we compare the two estimators we have discussed for $\{S(t), t \ge 0\}$, namely the empirical survival function and the NA estimator:

t	$\widehat{S}_e(t)$	$\widehat{S}_{NA}(t)$
$0 \le t < 2.5$	1	1
$2.5 \le t < 23$	0.9	0.9048
$23 \le t < 47$	0.8	0.8097
$47 \le t < 66$	0.7	0.7145
$66 \le t < 69$	0.6	0.6194
$69 \le t < 72$	0.5	0.5243
$72 \le t < 80$	0.4	0.4293
$80 \le t < 82$	0.3	0.3343
$82 \le t < 90$	0.1	0.1716
$t \ge 90$	0	0.0631

As we can see, the two estimates for the survival function $S(t)$ are close. However, we note a more significant difference in the estimates for large values of t.

1.3 Estimation of nonparametric lifetime models (with censoring)

We now look at how to construct nonparametric estimators for the survival function in the presence of censoring. Censoring is (almost always) an inevitable component of any data collection of survival study. In this course, we only consider *right-censoring*, which we introduce below along with other types of censoring (given for completeness).

- *Right-censoring*: (also called "censored from above") this refers to cases where there exists a threshold u such that if an observation is larger than u it gets recorded as u. An example would be when a survival study ends and a person is still alive at that point or if a participant voluntarily leaves a study.
- *Left-censoring*: (also called "censored-from-below") this refers to cases where there exists a threshold d such that if an observation is smaller than d it gets recorded as d. An example would be when there is a delay between the start of the study (time 0) and when the subjects are first observed to determine their status. E.g., someone is already dead by the time we first look at the patients in an hospital ward to determine their time of survival until death. This is less common.
- *Left-truncation:* (also called "truncated-from-below") this refers to cases where there exists a threshold d such that if an observation is smaller than d it is not recorded. An example would be in P&C insurance when we have a deductible d, losses below d are not recorded.
- *Right-truncation:* (also called 'truncated-from-above") this refers to cases where there exists a threshold u such that if an observation is larger than u, it is not recorded. This is quite uncommon. An example would be when we want to model the time between when a virus is contracted and when symptoms appear. If we recruit in our sample people who may have been exposed to the virus but did not necessarily got infected with it, then those who don't have symptoms at the end of the study are right-truncated, because the fact they don't have symptoms is probably because they didn't get the virus in the first place and therefore, they should not be included in our data.
- *Interval-censoring*: this refers to when we don't have the exact value of each observation, but instead know it lies within a certain interval. An example would be when we only record an integer age for the age at death.
- *Random censoring:* this refers to cases where the threshold u (or d) associated with censoring is a random variable. An example would be when we want to model the time-until-death rv for people who have a life insurance policy, and someone surrenders their policy after 3 years (so we lose track of them). Their decision to surrender and its timing was not known ahead of time and thus the censoring time (which here, takes values 3 years) should be modeled as a rv.
- *Non-informative censoring:* this refers to cases when the censoring process doesn't provide any information on the observation. For example, we may have a random threshold U_i for censoring that is independent of the observation rv T_i.
- *Type I censoring:* refers to when study ends after a pre-determined amount of time (known at time 0) and anyone for whom the event of interest didn't happen by the end of the study is right-censored at that value.
- *Type II censoring:* refers to when a study goes on until a fixed number of events occur, and the censoring times caused by the end of the study are therefore not known in advance.

We need some new notation to deal with censoring. For the j-th observation time, we let

$$x_j = \text{observation itself (if observed)}$$
$$u_j = \text{censoring value (if censored)}$$

For a given observation, we either get an x_j or a u_j (depending on whether the value was observed or censored). As before, $y_1 < .... < y_k$ are the k distinct observed values, with $s_i =$ nb of times y_i was observed. Note that here, we may have $\sum_{i=1}^{k} s_i < n$ (as some data points may be censored).

Next, we need to adjust how to compute the risk set r_j for the j-th (ordered) observation y_j which corresponds to the number of individuals who are still under observation at y_j. Hence,

$$r_j = (\#\text{ of } x_i \geq y_j) + (\#\text{ of } u_i \geq y_j) \tag{3}$$

Equivalently, we can compute the r_j's recursively as follows:

$$r_j = r_{j-1} - s_{j-1} - (\#\text{ of } u_i\text{'s in } [y_{j-1}, y_j))$$

for $j = 2, 3, ..., k$ where the starting point of this recursion is r_1 given by (3).

Finally, it is common practice in the present context to add a '+' next to a right-censored value u_j to indicate that the observation has been right-censored.

Example 7 *Suppose we have data giving us the time (in weeks) until a patient dies after surgery on a malignant melanoma. Some observations are censored (denoted by a '+') as the patient may have fully recovered and left the study, or died from another cause. The data is as follows: 10, 13+, 18+, 19, 23+, 30, 36 , 38+, 54+, 56+, 59, 75, 93, 97, 104+, 107, 107+, 107+. Compute the r_j's and the s_j's for this example.*

j	y_j	r_j	s_j
1	10	18	1
2	19	15	1
3	30	13	1
4	36	12	1
5	59	8	1
6	75	7	1
7	93	6	1
8	97	5	1
9	107	3	1

1.3.1 Kaplan-Meier (product limit) estimator

We start by giving a simple intuitive derivation for the estimator, and then later explain how this estimator arises as a MLE. The notation we use to denote the Kaplan-Meier estimator for $S(t)$ is $\widehat{S}_{KM}(t)$.

So we start with $\widehat{S}_{KM}(t) = 1$ for $t < y_1$. Then we have that

$$\widehat{S}_{KM}(y_1) = \frac{r_1 - s_1}{r_1} = \frac{\#\text{ of survivors beyond } y_1}{\text{risk set at } y_1}$$

How do we get $\widehat{S}_{KM}(y_2)$? We can write

$$S(y_2) = S(y_1)\,\frac{S(y_2)}{S(y_1)},$$

and as for the first step, now estimate $\frac{S(y_2)}{S(y_1)}$ by $\frac{r_2 - s_2}{r_2}$. By repeating this process, we get the following Kaplan-Meier estimator of the survival function:

$$\widehat{S}_{KM}(t) = \begin{cases} 1, & 0 \leq t < y_1, \\ \prod_{i=1}^{j-1} \frac{r_i - s_i}{r_i}, & y_{j-1} \leq t < y_j, \qquad j = 2, 3, ..., k, \\ \prod_{i=1}^{k} \frac{r_i - s_i}{r_i}, & t \geq y_k. \end{cases}$$

Remark 8 *If some censored and observed times are identical, we adopt the convention that the censoring time is larger by an infinitesimally small quantity so that the observation occurs first.*

Example 9 *The Kaplan-Meier estimate for the melanoma data set is*

t	$\widehat{S}_{KM}(t)$
$[0, 10)$	1
$[10, 19)$	$\frac{17}{18} = 0.944$
$[19, 30)$	$\frac{17}{18}\frac{14}{15} = 0.8815$
$[30, 36)$	$\frac{17}{18}\frac{14}{15}\frac{12}{13} = 0.8137$
$[36, 59)$	0.7459
$[59, 75)$	0.6526
$[75, 93)$	0.5594
$[93, 97)$	0.4662
$[97, 107)$	0.3729
$[107, \infty)$	0.2486

Note: The programming language R, along with the "survival" package (which you can get access to by typing LIBRARY(SURVIVAL)), can be used to compute the KM estimator. You can use the function SURV to create the data to be handled. In the vector called CENSORED, we indicate with a '1' a data point that was actually observed, and with '0' a data point that is censored.

```
library(survival)
times <- c(10,13,18,19,23,30,36,38,54,56,59,75,93,97,104,107,107,107)
censored <- c(1,0,0,1,0,1,1,0,0,0,1,1,1,1,0,1,0,0)
Sdata <- Surv(times, censored)
KMdsurv <- survfit(Sdata ~1, type = "kaplan-meier" , conf.type= "plain" )
```

You can then call `plot(KMdsurv)` to get a plot of the survival function, or `summary(KMdsurv)` to get numerical values for the r_i's, s_i's and corresponding KM estimates.

Variance of the KM estimator

As for any estimator, we may want to estimate the variance of the KM estimator. This is obviously useful to quantify the uncertainty related to the estimator. To derive the variance of the KM estimator, it is useful to first describe the KM estimator as an MLE. More precisely, the KM estimator has a nice interpretation as a nonparametric MLE. The idea is to write out the likelihood of our sample based on an unknown survival function (and corresponding hazard function) and then find the value of the discrete hazard rate function that maximizes the likelihood.

To construct the likelihood, we recall that we are interested in building a nonparametric model for T with only mass points at times $y_1 < y_2 < ... < y_k$. Hence, for the MLE the λ_i's are the "parameters" that need to be estimated. It is important to point out that the times $y_1 < y_2 < ... < y_k$ are assumed fixed here.

We consider the contribution to the likelihood of each data point in the sample. We make a distinction between data points that are censored or observed:

- for an (uncensored) observation at time y_i, its contribution to the likelihood is

$$\underbrace{\left\{\prod_{j=1}^{i-1}(1-\lambda_j)\right\}}_{\text{survival through } y_1,\ldots,y_{i-1}} \quad \underbrace{\lambda_i}_{\text{death at } y_i},$$

  for $i = 1, 2, ..., k$.

- for a censored observation at time $u_j \in [y_i, y_{i+1})$, its contribution to the likelihood is

$$\underbrace{\left\{\prod_{j=1}^{i}(1-\lambda_j)\right\}}_{\text{survival through } y_1,\ldots,y_i}.$$

For notational convenience, define

$$\begin{aligned} c_i &= \#\text{ of observations censored in } [y_i, y_{i+1}) \\ &= r_i - r_{i+1} - s_i, \end{aligned}$$

for $i = 1, 2, ..., k$. Aggregating the individual contributions of each data point in the sample, the likelihood is given by

$$L(\lambda_1, ..., \lambda_k) = \left\{\prod_{i=1}^{k}\left(\lambda_i \prod_{j=1}^{i-1}(1-\lambda_j)\right)^{s_i}\right\}\left\{\prod_{i=1}^{k}\left(\prod_{j=1}^{i}(1-\lambda_j)\right)^{c_i}\right\}$$

Therefore, we have

$$\begin{aligned} L(\lambda_1, ..., \lambda_k) &= \prod_{i=1}^{k}\lambda_i^{s_i}(1-\lambda_i)^{c_i+\ldots+c_k+s_{i+1}+\ldots+s_k} \\ &= \prod_{i=1}^{k}(\lambda_i)^{s_i}(1-\lambda_i)^{r_i-s_i}, \end{aligned}$$

(since $r_i = s_i + \ldots + s_k + c_i + \ldots + c_k$). To find the MLE for λ_j we need to solve (for each j)

$$\begin{aligned} \frac{\partial}{\partial\lambda_j}\left(s_j \ln\lambda_j + (r_j - s_j)\ln(1-\lambda_j)\right) &= 0 \\ \Leftrightarrow \frac{s_j}{\lambda_j} - \frac{r_j - s_j}{1-\lambda_j} &= 0 \\ \Leftrightarrow s_j(1-\lambda_j) = \lambda_j(r_j - s_j) &\Leftrightarrow s_j = \lambda_j r_j. \end{aligned}$$

Hence,

$$\boxed{\hat{\lambda}_j = \tfrac{s_j}{r_j}}$$

Therefore, the MLE for

$$S(t) = \prod_{i=1}^{j}(1-\lambda_i),$$

for $t \in [y_j, y_{j+1})$ is

$$\widehat{S}_{KM}(t) = \prod_{i=1}^{j} \left(1 - \widehat{\lambda}_i\right) = \prod_{i=1}^{j} \left(1 - \frac{s_i}{r_i}\right),$$

namely the KM estimator for $S(t)$.

Now we can explain how to get an approximation for the variance of the KM estimator. We first need to recall the so-called *delta method*.

Theorem 10 *For an estimator X_n based on a sample of size n, if $X_n \to N\left(\theta, \frac{\sigma^2}{n}\right)$ in distribution as $n \to \infty$ and g is a differentiable function, then*

$$g(X_n) \to N\left(g(\theta), \left(\frac{\partial g}{\partial \theta}\right)^2 \frac{\sigma^2}{n}\right).$$

So we want to apply this result in order to approximate

$$Var\left(\widehat{S}_{KM}(t)\right) = Var\left(\prod_{i=1}^{m} \left(1 - \widehat{\lambda}_j\right)\right), \qquad m \leq k. \tag{4}$$

We proceed by going through the following steps:

1. We first rewrite (4) as

$$Var\left(\prod_{j=1}^{m} \left(1 - \widehat{\lambda}_j\right)\right) = Var\left(\exp\left(\ln \prod_{j=1}^{m} \left(1 - \widehat{\lambda}_j\right)\right)\right) = Var\left(\exp\left(\sum_{j=1}^{m} \ln\left(1 - \widehat{\lambda}_j\right)\right)\right).$$

   We will first apply the delta method to the ln function inside the summation, and then we apply it again to the exponential function.

2. Conditional on r_j, the number of observed values at time y_j, namely s_j, is a binomial rv with mean $r_j\lambda_j$ and variance $r_j\lambda_j(1-\lambda_j)$. It follows that

$$Var\left(1 - \widehat{\lambda}_j\right) = Var\left(\widehat{\lambda}_j\right) = Var\left(1 - \frac{s_j}{r_j}\right) = \frac{\lambda_j(1-\lambda_j)}{r_j}.$$

3. Using the delta method, we get

$$Var(\ln\left(1 - \widehat{\lambda}_j\right)) \approx \left(\frac{1}{1-\lambda_j}\right)^2 \frac{\lambda_j(1-\lambda_j)}{r_j} = \frac{\lambda_j}{(1-\lambda_j)\, r_j}.$$

4. Then, we use the approximation

$$Var\left(\sum_{j=1}^{m} \ln\left(1 - \widehat{\lambda}_j\right)\right) \approx \sum_{j=1}^{m} Var\left(\ln\left(1 - \widehat{\lambda}_j\right)\right) = \sum_{j=1}^{m} \frac{\lambda_j}{(1-\lambda_j)\, r_j}$$

   by assuming independence between the s_j's (viewing the r_j's as being fixed).

5. We then apply the delta method using the function $g(x) = e^x$ so that we get

$$\begin{aligned} Var\left(\widehat{S}_{KM}(t)\right) &= Var\left(\exp\left(\sum_{j=1}^{m}\ln\left(1-\widehat{\lambda}_j\right)\right)\right) \\ &\approx \left(e^{\sum_{j=1}^{m}\ln\left(1-\widehat{\lambda}_j\right)}\right)^2 \sum_{j=1}^{m}\frac{\lambda_j}{(1-\lambda_j)\,r_j} \\ &= \left(\widehat{S}_{KM}(t)\right)^2 \sum_{j=1}^{m}\frac{\lambda_j}{(1-\lambda_j)\,r_j}, \end{aligned}$$

which we shall approximate by

$$\begin{aligned} \widehat{Var}\left(\widehat{S}_{KM}(t)\right) &= \left(\widehat{S}_{KM}(t)\right)^2 \sum_{j=1}^{m}\frac{\widehat{\lambda}_j}{\left(1-\widehat{\lambda}_j\right)r_j} \\ &= \left(\widehat{S}_{KM}(t)\right)^2 \sum_{j=1}^{m}\frac{\frac{s_j}{r_j}}{\left(1-\frac{s_j}{r_j}\right)r_j} \\ &= \left(\widehat{S}_{KM}(t)\right)^2 \sum_{j=1}^{m}\frac{s_j}{r_j\,(r_j-s_j)} \end{aligned}$$

Summarizing, we get

$$\boxed{Var(\hat{S}_{KM}(y_m)) \approx (\hat{S}_{KM}(y_m))^2 \sum_{j=1}^{m}\frac{s_j}{r_j(r_j-s_j)}} \tag{5}$$

which is called *Greenwood's approximation.*

Example 11 *If we go back to our melanoma example, then we have*

j	y_j	$\hat{S}_{KM}(y_j)$	$\hat{\text{Var}}(\hat{S}_{KM}(y_j))$
1	10	0.944	$(0.944)^2\frac{1}{18\times17} = 0.0002915$
2	19	0.8815	$(0.8815)^2\left(\frac{1}{18\times17}+\frac{1}{15\times14}\right) = 0.006239$
3	30	0.8137	0.00956
4	36	0.7459	0.012248
5	59	0.6526	0.016983
6	75	0.5594	0.019928
7	93	0.4662	0.021083
8	97	0.3729	0.020447
9	107	0.2486	0.01939

Once we have an approximation for $Var(\hat{S}_{KM}(y_m))$, namely the Greenwood estimator (5), we can construct an approximate confidence interval (CI) for $S(t)$ using the normal approximation (as suggested by the delta method). That is, we use the CI $\hat{S}_{KM}(t) \pm z_{1-\alpha/2}\sqrt{(\hat{S}_{KM}(t))^2 \sum_{j=1}^{m}\frac{s_j}{r_j(r_j-s_j)}}$ for $t \in [y_m, y_{m+1})$. Note that this confidence interval can lie outside of the interval $[0,1]$ in which case the lower and/or upper bound are adjusted accordingly.

1.3.2 Nelson-Aalen estimator

We now introduce the NA estimator for the cumulative hazard rate $H(t)$ in the presence of censoring. As it turns out, the estimator remains defined in the exact same way as in the non-censoring case, with the understanding that the risk sets r_i are calculated to take into account censoring, i.e., the same way we compute them for the KM estimator in Section 1.3.1. In other words, in the presence of censoring we still have

$$\widehat{H}_{NA}(t) = \begin{cases} 0, & t < y_1 \\ \sum_{i=1}^{j-1} \frac{s_i}{r_i}, & y_{j-1} \le t < y_j, \qquad j = 2, 3, ..., k, \\ \sum_{i=1}^{k} \frac{s_i}{r_i}, & t \ge y_k. \end{cases}$$

We obtain its associated survival function $\hat{S}_{NA}(t)$ via

$$\hat{S}_{NA}(t) = e^{-\widehat{H}_{NA}(t)}, \qquad t \ge 0.$$

Example 12 *Using the melanoma data, we get*

j	y_j	r_j	s_j	$\hat{H}_{NA}(y_j)$	$\hat{S}_{NA}(y_j)$
1	10	18	1	$\frac{1}{18}$	0.9460
2	19	15	1	$\frac{1}{18} + \frac{1}{15}$	0.8850
3	30	13	1	$\frac{1}{18} + \frac{1}{15} + \frac{1}{13}$	0.8194
4	36	12	1	$\frac{1}{18} + \ldots + \frac{1}{12}$	0.7539
5	59	8	1	$\frac{1}{18} + \ldots + \frac{1}{8}$	0.6653
6	75	7	1	$\frac{1}{18} + \ldots + \frac{1}{7}$	0.5768
7	93	6	1	$\frac{1}{18} + \ldots + \frac{1}{6}$	0.4882
8	97	5	1	$\frac{1}{18} + \ldots + \frac{1}{5}$	0.3997
9	107	3	1	$\frac{1}{18} + \ldots + \frac{1}{3}$	0.2864

Before deriving an approximation for the variance of the NA estimator, it is useful to mention that this estimator too can be derived as a MLE. Indeed, if we assume T is a discrete rv on $y_1 < y_2 < ... < y_k$ then we have

$$H(t) = \sum_{j=1}^{m} \lambda_j$$

for $t \in [y_m, y_{m+1})$ and as we discussed in the case of the KM estimator, the MLE estimator for λ_j is precisely $\widehat{\lambda}_j = s_j/r_j$. Hence, we can write

$$\hat{H}_{NA}(t) = \sum_{j=1}^{m} \widehat{\lambda}_j$$

for $t \in [y_m, y_{m+1})$. We can then approximate the variance of $\hat{H}_{NA}(t)$ by conditioning on the r_j's and assuming the s_j's are independent (as we did for the KM estimator). Hence we get

$$\mathrm{Var}(\hat{H}_{NA}(y_m)) \approx \sum_{j=1}^{m} \frac{\lambda_j(1-\lambda_j)}{r_j}.$$

which is approximated by

$$\begin{aligned}\widehat{\mathrm{Var}}(\hat{H}_{NA}(y_m)) &= \sum_{j=1}^{m} \frac{\widehat{\lambda}_j(1-\widehat{\lambda}_j)}{r_j} \\ &= \sum_{j=1}^{m} \frac{\frac{s_j}{r_j}(1-\frac{s_j}{r_j})}{r_j} \\ &= \sum_{j=1}^{m} \frac{s_j(r_j-s_j)}{(r_j)^3} \\ &\approx \sum_{j=1}^{m} \frac{s_j}{(r_j)^2}.\end{aligned}$$

We can then construct an approximate confidence interval for $H(t)$ as follows: $(x_1, x_2) = \hat{H}_{NA}(t) \pm z_{1-\alpha/2}\sqrt{\sum_{j=1}^{m} \frac{s_j}{(r_j)^2}}$ for $t \in [y_m, y_{m+1})$. The resulting confidence interval for $S(t)$ is given by (e^{-x_2}, e^{-x_1}).

Finally, to compute the NA estimator using the survival package in R, you need to specify the type of estimator as being "fleming-harrington" as follows:

```
NAdsurv <- survfit(Sdata ~1, type = "fleming-harrington" , conf.type= "log" )
```

The name "Fleming-Harrington" refers to the estimator $e^{-\hat{H}_{NA}(t)}$ for $S(t)$ when $\hat{H}_{NA}(t)$ is the Nelson-Aalen estimator. It is important to note that by default, the standard error returned under the command `summary(NAdsurv)` is $\sqrt{\left(e^{-\hat{H}_{NA}(y_m)}\right)^2 \sum_{j=1}^{m} \frac{s_j}{(r_j)^2}}$. Also, the default for computing a CI for $S(t)$ is to use the method described above.

1.4 Cox proportional hazard model

The nonparametric approach discussed earlier may not work well if we believe there are a number of covariates (e.g., age, sex, income) that affect the survival function. Why? As the future lifetimes are not homogeneous, we would need to repeat the modeling approach for each combination of covariate values.

While we could choose a parametric family to represent the survival function and let the parameters be a function of the covariates, a more flexible approach is to model a (possibly parametric) baseline case and then let the covariates modify it accordingly. This is the idea behind the *Cox proportional hazard model*.

Definition 13 *A lifetime rv T follows a* Cox proportional hazard model *if its corresponding hazard function (i.e., force of mortality) for a life with covariates $\mathbf{z} = (z_1, \ldots, z_p)'$ can be written as*

$$\lambda(t; \mathbf{z}) = \lambda_0(t) \exp(\beta_1 z_1 + \ldots + \beta_p z_p), \tag{6}$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$ is a vector of coefficients and $\lambda_0(t)$ is the baseline hazard rate function.

Example 14 *Let $p = 2$, z_1 =age and*

$$z_2 = \begin{cases} 0, & \text{if female} \\ 1, & \text{if male} \end{cases}$$

Then,

$$\lambda(t, \mathbf{z}) = \lambda_0(t) e^{\beta_1 z_1 + \beta_2 z_2}.$$

For a male and a female of the same age x, we find

$$\lambda(t; (x, 1)) = \lambda_0(t) e^{\beta_1 x + \beta_2} = \lambda(t; (x, 0)) e^{\beta_2}.$$

This means that the gender effect is the same for all ages x, i.e., it doesn't depend on x as

$$\frac{\lambda(t; (x, 1))}{\lambda(t; (x, 0))} = e^{\beta_2}.$$

Remark 15 *For the Cox proportional hazard rate (6), the survival function for a life with covariates $\mathbf{z}$ is given by*

$$\begin{aligned} S(t; \mathbf{z}) &= e^{-\int_0^t \lambda_0(s) \exp(\beta_1 z_1 + \ldots + \beta_p z_p) ds} \\ &= \left(e^{-\int_0^t \lambda_0(s) ds} \right)^{\exp(\beta_1 z_1 + \ldots + \beta_p z_p)} \\ &= \left(e^{-H_0(t)} \right)^{\exp(\beta_1 z_1 + \ldots + \beta_p z_p)}, \end{aligned}$$

where $H_0(t) = \int_0^t \lambda_0(s) ds$ is the baseline cumulative hazard rate.

The term "proportional" in the Cox model comes from the fact that for two individuals represented respectively by the vectors $\mathbf{z}_a$ and $\mathbf{z}_b$, we have that

$$\frac{\lambda(t; \mathbf{z}_a)}{\lambda(t; \mathbf{z}_b)} = \frac{e^{\boldsymbol{\beta} \cdot \mathbf{z}_a}}{e^{\boldsymbol{\beta} \cdot \mathbf{z}_b}},$$

which is independent of t. This also means we can compare lives (i.e., make relative statements) without knowing the baseline hazard rate function λ_0.

Based on these observations, it should be clear that the part of this model that should first be estimated is the vector $\boldsymbol{\beta}$.

1.4.1 Estimation of $\boldsymbol{\beta}$ using the partial likelihood

The "proportional" property of the model allows us to estimate the vector $\boldsymbol{\beta}$ separately from the baseline hazard rate $\lambda_0(t)$. As before, we assume $y_1 < \ldots < y_k$ are the (distinct) times of death. For now, we also assume that $s_i = 1$ for $i = 1, \ldots, k$, i.e., there are no ties. We first need to introduce a bit more notation: we define $\mathcal{R}_j$ to be the set of individuals who are under observation at time y_j, so that r_j (what we called the risk set before) is the size of the set $\mathcal{R}_j$.

Rather than trying to construct the full likelihood, we instead work with a *partial likelihood* function, which focuses on the selection of who will die at each time y_j given who is in the risk set. So we refer to this likelihood as the partial likelihood at it ignores the timing of deaths (and censoring events).

More precisely, we first remind ourselves that the conditional probability density function of T at y_j given that $T \geq y_j$ is precisely the hazard rate function evaluated at y_j. Hence, for a death at time y_j with the set $\mathcal{R}_j$ of people at risk, the probability that it is a life with covariates $\mathbf{z}_j$ that ends at time y_j is given by

$$\frac{\lambda(t;\mathbf{z}_j)}{\sum_{l\in\mathcal{R}_j}\lambda(t;\mathbf{z}_l)} = \frac{e^{\boldsymbol{\beta}\cdot\mathbf{z}_j}}{\sum_{l\in\mathcal{R}_j}e^{\boldsymbol{\beta}\cdot\mathbf{z}_l}}$$

We build the partial likelihood by multiplying these probabilities, i.e.

$$L(\boldsymbol{\beta}) = \prod_{j=1}^{k}\frac{e^{\boldsymbol{\beta}\cdot\mathbf{z}_j}}{\sum_{l\in\mathcal{R}_j}e^{\boldsymbol{\beta}\cdot\mathbf{z}_l}}.$$

This is a function of only $\boldsymbol{\beta}$ (and not $\lambda_0(t)$); the idea is then to find the value of $\boldsymbol{\beta}$ that maximizes $L(\boldsymbol{\beta})$. Note also that the partial likelihood function $L(\boldsymbol{\beta})$ doesn't depend on the actual death times, but only on the ordering (who dies first, second, etc.).

Note: as done in the Loss Models textbook, it is convenient to use the notation $c_i = e^{\boldsymbol{\beta}\cdot\mathbf{z}_i}$ to shorten the description of the partial likelihood function.

Example 16 *(Example 17.8 on page 495 of LM, 3rd edition) Suppose that the size of a homeowner's fire insurance claim as a percentage of the house's value depends on the age of the house and the type of construction (wood or brick). We can develop a Cox proportional hazard model for this situation. Indeed, let z_1 be the age of the house (in years) amd $z_2 = 1$ if the construction is in wood or $z_2 = 0$ if it is in brick. Suppose we have the following data, where the payments are expressed as a percentage of the value of the house:*

z_1	z_2	*payment*
10	*0*	*70*
20	*0*	*22*
30	*0*	*90+*
40	*0*	*81*
50	*0*	*8*
10	*1*	*51*
20	*1*	*95+*
30	*1*	*55*
40	*1*	*85+*
50	*1*	*93*

The ordered uncensored values are 8, 22, 51, 55, 70, 81 and 93. Working out the first term of the partial likelihood function, we observe that the vector $\mathbf{z}_1$ of covariates associated with the observation 8 is $\mathbf{z}_1 = (50, 0)$ and the risk set $\mathcal{R}_1$ includes all 10 data points. Using the above notation for the c_i's, the first term of the partial likelihood function is

$$\frac{c_1}{c_1 + \ldots + c_{10}}.$$

The next term corresponding to the payment of 22 has $\mathbf{z}_2 = (20, 0)$ and risk set $\mathcal{R}_2$ that includes all the observations except the smallest payment of 8. It thus contributes a term

$$\frac{c_2}{c_2 + \ldots + c_{10}}.$$

Continuing this way for the remaining 5 observations, we get

$$L(\boldsymbol{\beta}) = \frac{c_1}{c_1 + \ldots + c_{10}} \frac{c_2}{c_2 + \ldots + c_{10}} \frac{c_3}{c_3 + \ldots + c_{10}} \frac{c_4}{c_4 + \ldots + c_{10}} \frac{c_5}{c_5 + \ldots + c_{10}} \frac{c_6}{c_6 + \ldots + c_{10}} \frac{c_9}{c_9 + c_{10}}$$

where

$$\begin{array}{llll} c_1 & = e^{50\beta_1} & c_6 & = e^{40\beta_1} \\ c_2 & = e^{20\beta_1} & c_7 & = e^{40\beta_1+\beta_2} \\ c_3 & = e^{10\beta_1+\beta_2} & c_8 & = e^{30\beta_1} \\ c_4 & = e^{30\beta_1+\beta_2} & c_9 & = e^{50\beta_1+\beta_2} \\ c_5 & = e^{10\beta_1} & c_{10} & = e^{20\beta_1+\beta_2}. \end{array}$$

Thus we see that the numerator of $L(\boldsymbol{\beta})$ is given by

$$e^{300\beta_1+5\beta_2}.$$

We could (although it is tedious) also find the denominator and then solve for β_1 and β_2. Note that this cannot be done by hand and requires numerical methods. In practice, we should handle problems like this by making use of built-in functions in R.

Example 17 *We want to use a Cox proportional hazards model for the survival of a certain type of bird (in years) based on whether they come from a large pet store ($z = 1$) or a specialized bird store ($z = 0$). We have the following data:*

spec. store	*3+*	*10*	*16*	*20+*
large store	*2*	*4+*	*12*	*13+*

We can estimate β by first getting our partial likelihood:

$$L(\beta) = \frac{e^\beta}{4e^\beta + 4} \frac{1}{2e^\beta + 3} \frac{e^\beta}{2e^\beta + 2} \frac{1}{2}.$$

Letting $x = e^\beta$, it follows that

$$g(x) = \frac{x^2}{16(x+1)^2(2x+3)}$$

then maximize

$$\ln g(x) = 2\ln x - 2\ln(x+1)\ln(2x+3) - \ln 16$$

by setting

$$\frac{\partial \ln g(x)}{\partial x} = \frac{2}{x} - \frac{2}{x+1} - \frac{2}{2x+3} = 0$$

which after some manipulations can be seen to be equivalent to $-x^2 + x + 3 = 0$, with positive solution $x = 0.5(1 + \sqrt{13}) = 2.3027756$ and thus $\beta = \ln x = 0.83411$. This means the hazard rate for a bird from the large pet store is about 2.3 times larger than the one for a bird from the specialized store.

1.4.2 Properties of $\hat{\boldsymbol{\beta}}$ as a maximum (partial) likelihood estimator

Interestingly, the estimator $\hat{\boldsymbol{\beta}}$ has properties that are similar to those of a pure MLE. That is, asymptotically $\hat{\boldsymbol{\beta}} \sim$ multivariate normal and is unbiased with an asymptotic variance matrix that can be estimated by the inverse of the observed information matrix $I(\hat{\boldsymbol{\beta}})$ whose (i,j)th entry is given by

$$I(\hat{\boldsymbol{\beta}})_{i,j} = -\mathbb{E}\left(\frac{\partial^2 \ln L(\boldsymbol{\beta})}{\partial\beta_i \partial\beta_j}\right), \qquad 1 \le i,j \le p$$

The efficient score function, given by

$$u(\boldsymbol{\beta}) = \left(\frac{\partial \ln L(\boldsymbol{\beta})}{\partial\beta_1}, \ldots, \frac{\partial \ln L(\boldsymbol{\beta})}{\partial\beta_p}\right)$$

is important for the estimation process. Indeed, solving $u(\boldsymbol{\beta}) = \mathbf{0}$ yields $\hat{\boldsymbol{\beta}}$.

The reason why it is useful to know these properties is that it allows us (among other things) to construct confidence intervals for the β_j's. If a confidence interval for β_j contains 0, then we conclude that the corresponding covariate z_j does not have much explanatory value.

Example 18 *In the bird example, we can estimate* $\mathrm{Var}(\hat{\beta})$ *using*

$$\left(\frac{-\partial^2 \ln L(\beta)}{\partial\beta^2}\right)^{-1}\Big|_{\beta=\hat{\beta}}$$

where

$$\begin{aligned}
\ln L(\beta) &= 2\ln e^{\beta} - 2\ln(e^{\beta}+1) - \ln(2e^{\beta}+3) - \ln 16 \\
\Rightarrow \frac{\partial \ln L(\beta)}{\partial\beta} &= 2\left(1 - \frac{1}{e^{-\beta}+1} - \frac{1}{2+3e^{-\beta}}\right) \\
\Rightarrow \frac{\partial^2 \ln L(\beta)}{\partial\beta^2} &= -2e^{-\beta}\left(\frac{1}{(1+e^{-\beta})^2} + \frac{3}{(2+3e^{-\beta})^2}\right) = -0.66106
\end{aligned}$$

when evaluated at $\beta = \hat{\beta}$*. Thus we have that*

$$\mathrm{Var}(\hat{\beta}) \approx -1/0.66106 = 1.5127 \text{ and } \hat{\sigma}_{\hat{\beta}} = 1.2299.$$

Therefore, our CI for β *is* $[0.83411 \pm 1.96 \times 1.2299] = [-1.5765, 3.2448]$*. The CI for* e^{β} *can be obtained by exponentiating the end points of the CI for* β*, thus getting* $[e^{-1.5765}, e^{3.2448}] = [0.2067, 25.65]$.

1.4.3 Cox proportional hazard models in R (skip)

The survival package in R includes functions to perform estimation for the Cox proportional hazard model. Below is an example that shows how to use R to handle the above example (from LM). The function `coxph` replaces the function `survfit` that we were using for Kaplan-Meier and Nelson-Aalen. One feature to explain is that we have the option to specify which covariates we want to use: see below the `~(z1+z2)` versus `~z2` or `~z1`.

```
library(survival)
covdata<-list(time=c(70,22,90,81,8,51,95,55,85,93),
            status=c(1,1,0,1,1,1,0,1,0,1),
            z1=c(10,20,30,40,50,10,20,30,40,50),
            z2=c(0,0,0,0,0,1,1,1,1,1))
fitcox<-coxph(Surv(time,status)~(z1 + z2),covdata)
print(summary(fitcox))

fitcoxjust2<-coxph(Surv(time,status)~z2,covdata)
print(summary(fitcoxjust2))

fitcoxjust1<-coxph(Surv(time,status)~z1,covdata)
print(summary(fitcoxjust1))
```

The function `summary` returns a lot of information on $\hat{\boldsymbol{\beta}}$, including results from statistical tests informing us about the validity of the null hypothesis that $\boldsymbol{\beta} = \mathbf{0}$. More on this later.

1.4.4 Some hypothesis tests for the Cox proportional hazards model

More precisely, if we want to test the hypothesis $H_0 : \beta_i = 0$ against $H_a : \beta_i \neq 0$, then we can use the above fact that $\hat{\beta}_i$ is asymptotically normal and perform:

1. **Z-test**: here we use the fact that $\frac{\hat{\beta}_i}{\sigma_{\hat{\beta}_i}}$ is approximately $N(0,1)$ under H_0; then we can compute $z = \frac{\hat{\beta}_i}{\sigma_{\hat{\beta}_i}}$ and the corresponding p-value $p = \mathbb{P}(|N(0,1)| > z)$ and if p is close to 0, we reject H_0.
2. **Likelihood ratio test**: here we compute $\ell_0 = \ln L(\beta_0)$ where β_0 is the value of the parameters $\beta_1, \ldots, \beta_p$ under H_0 and $\ell_a = \ln L(\hat{\beta})$ is the log-likelihood evaluated at the value of our estimated parameters. Then we know that $2(\ell_a - \ell_0)$ is a chi-square with v degrees of freedom, where v is the number of estimated parameters in H_a minus the number of estimated parameters in H_0. Here again, we can compute $p = \mathbb{P}(\chi^2(v) > y)$ where y is the value taken by the test statistic $2(\ell_a - \ell_0)$ and reject H_0 if p is too small.

We may also be interested in the *factor effect*, i.e., we want a CI on e^{β_i} that represents by how much the hazard rate increases when z_i goes from 0 to 1. A CI for e^{β_i} can be simply obtained by exponentiating the endpoints of the CI for β_i.

Example 19 *For the bird example, we can perform the Z-test and obtain*

$$z = \frac{\hat{\beta}}{\hat{\sigma}_{\hat{\beta}}} = \frac{0.83411}{1.2299} = 0.6782.$$

so that $\mathbb{P}(|N(0,1)| > z) = 0.4978$ *which is not small so we do not reject* H_0.

We can also perform the likelihood ratio test, which gives us $\ell_a = -5.5228$ *(*$\ln L(\beta)$ *evaluated at* $\beta = \hat{\beta} = 0.83411$*) and* $\ell_0 = \ln L(0) = -5.7683$ *so* $y = 0.49111$ *and* $P(\chi^2(1) > 0.49111) = 0.4834$ *so again we do not reject* H_0.

1.4.5 Handling ties (skip)

If we have ties for some of the observations, then we need to modify the above approach. So assume we have $s_i > 1$ observations for which the time of death is y_i, and let $\S_i$ be the set of size s_i giving the indices v such that $\mathbf{z}_v$ is a vector of covariates corresponding to one of the s_i observations with a time of death equal to y_i. Then one possible approach attributed to Breslow is to replace the term

$$\frac{e^{\boldsymbol{\beta}\cdot\mathbf{z}_i}}{\sum_{l\in\mathcal{R}_i} e^{\boldsymbol{\beta}\cdot\mathbf{z}_l}}$$

in $L(\boldsymbol{\beta})$ by

$$\frac{\prod_{v\in\S_i} e^{\beta\mathbf{z}_v}}{\left(\sum_{l\in\mathcal{R}_i} e^{\boldsymbol{\beta}\cdot\mathbf{z}_l}\right)^{s_i}}. \tag{7}$$

Note that the risk set $\mathcal{R}_i$ remains the same for all observations tied at y_i.

Another approach, which is the default one used in R, is due to Efron and replaces the denominator in (7) by

$$\prod_{r=0}^{s_i-1}\left(\sum_{l\in\mathcal{R}_i-\mathcal{S}_i} e^{\boldsymbol{\beta}\cdot\mathbf{z}_l} + \frac{s_i-r}{s_i}\sum_{v\in\mathcal{S}_i} e^{\boldsymbol{\beta}\cdot\mathbf{z}_v}\right).$$

1.4.6 Estimating the baseline hazard function

Next, we discuss one possible approach to estimate the baseline hazard function $\{\lambda_0(t), t\geq 0\}$. We do so by estimating the cumulative baseline hazard rate

$$H_0(t) = \int_0^t \lambda_0(s)\,\mathrm{d}s, \qquad t\geq 0,$$

instead (rather than estimating $\lambda_0(t)$ directly).

For this purpose, let S_i be the number of observed data points at time y_i (whose realization is denoted by s_i). Given that $\mathcal{R}_i$ individuals are at risk at time y_i, the conditional mean of S_i can be rewritten as

$$\mathrm{E}(S_i|\mathcal{R}_i) = \sum_{l\in\mathcal{R}_i}\mathrm{E}(I_l),$$

where

$$I_l = \begin{cases} 1, & \text{if the } l\text{-th individual at risk at time } y_i \text{ dies,}\\ 0, & \text{otherwise.}\end{cases}$$

It follows that

$$\begin{aligned}\mathrm{E}(I_l) &= H_0(\mathrm{d}y_i)e^{\beta\cdot\mathbf{z}_l}(1) + \left(1 - H_0(\mathrm{d}y_i)e^{\beta\cdot\mathbf{z}_l}\right)(0)\\ &= H_0(\mathrm{d}y_i)e^{\beta\cdot\mathbf{z}_l}.\end{aligned}$$

Hence,

$$\mathrm{E}(S_i|\mathcal{R}_i) = \sum_{l\in\mathcal{R}_i} H_0(\mathrm{d}y_i)e^{\beta\cdot\mathbf{z}_l} = H_0(\mathrm{d}y_i)\sum_{l\in\mathcal{R}_i} e^{\beta\cdot\mathbf{z}_l}.$$

Now, to find an estimator for H_0, we replace $\mathrm{E}(S_i|\mathcal{R}_i)$ by the realization of the rv S_i (namely, s_i) and $\boldsymbol{\beta}$ by its MLE $\widehat{\boldsymbol{\beta}}$, i.e.

$$\widehat{H}_0(\mathrm{d}y_i) = \frac{s_i}{\sum_{l\in\mathcal{R}_i} e^{\hat{\beta}\cdot\mathbf{z}_l}}.$$

When there are no ties in the observed times (which is the assumption we will make in this course), $s_i = 1$ which leads to

$$\widehat{H}_0(\mathrm{d}y_i) = \frac{1}{\sum_{l\in\mathcal{R}_i} e^{\hat{\beta}\cdot\mathbf{z}_l}}.$$

Note that $\widehat{H}_0$ has jumps at the observed times y_i of size $\frac{1}{\sum_{l\in\mathcal{R}_i} e^{\hat{\beta}\cdot\mathbf{z}_l}}$ and is otherwise constant.

In conclusion,

$$\widehat{H}_0(t) = \sum_{\{i:y_i\le t\}} \frac{1}{\sum_{l\in\mathcal{R}_i} e^{\hat{\beta}\cdot\mathbf{z}_l}},$$

for $t \ge 0$, which is known as the Breslow estimator or the generalized Nelson-Aalen estimator of $H_0(t)$ (note that if $\hat{\beta}$ is 0, then $\widehat{H}_0(t)$ reverts to the NA estimator).

Example 20 *For the bird example, we get:*

i	y_i	z_{d_i}	$\widehat{H}_0(t)$ *for* $t \in [y_i, y_{i+1})$
1	*2*	*1*	$1/(4+4\exp(\hat{\beta})) = 0.07569$
2	*10*	*0*	$0.07569 + 1/(3+2\exp(\hat{\beta})) = 0.2072$
3	*12*	*1*	$0.2072 + 1/(2+2\exp(\hat{\beta})) = 0.3586$
4	*16*	*0*	$0.3586 + 1/2 = 0.8586$

1.4.7 Using R to get the baseline hazard rate (skip)

Using the bird example, we can get the survival function $\hat{S}_0(t) = e^{-\hat{H}_0(t)}$ corresponding to the baseline hazard rate using R by using the following code:

```
library(survival)
coxdata<-list(times=c(2,3,4,10,12,13,16,20),
            status=c(1,0,0,1,1,0,1,0),
            z=c(1,0,1,0,1,1,0,0))
fitcox<-coxph(Surv(times,status)~z,coxdata)
print(summary(fitcox))
birdfit<-survfit(fitcox,newdata=list(z=0))
summary(birdfit)
```

You can check that the estimate for $S_0(t)$ returned by R at $t = y_1 = 2$ is indeed given by $\exp(-0.07569) = 0.927$ etc.

1.5 Multiple-state models

1.5.1 Introduction

Here, we consider the multiple-state model introduced in ACTSC 622. We recall that the multiple-state model is a flexible mathematical framework used to model the evolution of the status of a policyholder or a group of policyholders over time.

Let $Y(t)$ denote the status of a policyholder or a group of policyholders at time $t \geq 0$. For the multiple-state Markov model, we assume that $\{Y(t)\}_{t\geq 0}$ is a non-homogeneous continuous-time Markov process on the finite state space $\{0, 1, ..., n\}$ with time-t infinitesimal generator

$$G(t) = \begin{bmatrix} \mu_t^{00} & \mu_t^{01} & \mu_t^{02} & \cdots & \mu_t^{0n} \\ \mu_t^{10} & \mu_t^{11} & \mu_t^{12} & \cdots & \mu_t^{1n} \\ \mu_t^{20} & \mu_t^{21} & \mu_t^{22} & \cdots & \mu_t^{2n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mu_t^{n0} & \mu_t^{n1} & \mu_t^{n2} & \cdots & \mu_t^{nn} \end{bmatrix},$$

where $\mu_t^{ij} \geq 0$ is the time-t instantaneous rate of transition from state i to state j $(i \neq j)$ and

$$\mu_t^{ii} = -\sum_{\substack{j=0 \\ j\neq i}}^{n} \mu_t^{ij},$$

is minus the rate of exit from state i at time t. For this model, we recall the following properties:

1. As $\{Y(t), t \geq 0\}$ is a Markov process, the future evolution of the status only depends on the current status and is irrelevant of the past history.

2. For the Markov process $\{Y(t), t \geq 0\}$, we assume that

$$\lim_{h\to 0} \frac{\mathbb{P}(2 \text{ or more transitions of } Y \text{ within any time period } h)}{h} = 0,$$

   which is equivalent to state that $\mathbb{P}(2 \text{ or more transitions within a time period } h)$ is a $o(h)$ function.

3. Letting

$${}_tp_x^{\overline{ii}} = \mathbb{P}(Y(x+s) = i \text{ for } \forall s \in [0,t] \,|Y(x) = i),$$

   we know that

$${}_tp_x^{\overline{ii}} = \exp\left(\int_0^t \mu_{x+s}^{ii} \mathrm{d}s\right).$$

To carry the estimation, we further make the following assumption:

Assumption: for x a positive integer and $t \in [0, 1)$, let $\mu_{x+t}^{jk} = \mu_x^{jk}$ for all $j \neq k$.

In other words, we assume the transition rates are constant in each year of age. As such, the objective is to carry out the estimation of the non-diagonal elements of the infinitesimal generator $G(t)$ for each year of

age. More precisely, for each year of age x the estimation of μ_x^{jk} (for all j,k for which μ_x^{jk} is not trivially 0) needs to be carried out.

We focus on observations between ages x and $x+1$ from a group of N independent lives in order to estimate the above rates. Namely, for each life we need

1. V_i^j: time spent by the ith life in state j over the year of age x to $x+1$
2. N_i^{jk}: number of transitions from state j to state k $(j \neq k)$ over the year of age x to $x+1$

We aggregate each life's contributon by defining $V^j = \sum_{i=1}^N V_i^j$ and $N^{jk} = \sum_{i=1}^N N_i^{jk}$. Also, we use the corresponding small letters to denote the observed values.

For the ith life, the contribution to the likelihood is

$$\left(\prod_{j=0}^{n} e^{\mu_x^{jj} V_i^j}\right)\left(\prod_{j=0}^{n}\prod_{\substack{k=0\\k\neq j}}^{n}\left(\mu_x^{jk}\right)^{N_i^{jk}}\right)$$

Given that the N lives are independent, the total likelihood L is given by

$$\begin{aligned} L &= \prod_{i=1}^{N}\left\{\left(\prod_{j=0}^{n} e^{\mu_x^{jj} V_i^j}\right)\left(\prod_{j=0}^{n}\prod_{\substack{k=0\\k\neq j}}^{n}\left(\mu_x^{jk}\right)^{N_i^{jk}}\right)\right\} \\ &= \left(\prod_{j=0}^{n} e^{\mu_x^{jj} V^j}\right)\left(\prod_{j=0}^{n}\prod_{\substack{k=0\\k\neq j}}^{n}\left(\mu_x^{jk}\right)^{N^{jk}}\right) \end{aligned}$$

To find e.g., the MLE of μ_x^{01}, we differentiate L with respect to μ_x^{01} and set the derivative at 0. We find

$$-V^0\left(\mu_x^{01}\right)^{N^{01}} + N^{01}\left(\mu_x^{01}\right)^{N^{01}-1} = 0,$$

or equivalently, the MLE must be the solution of

$$\mu_x^{01} = \frac{N^{01}}{V^0}.$$

More generally, we find that the MLE of μ_x^{jk} $(j \neq k)$ is

$$\widehat{\mu}_x^{jk} = \frac{N^{jk}}{V^j}.$$

Example 21 *For the alive/dead model (state 0 - alive, state 1 - dead), the only parameter to estimate is μ_x^{01} whose MLE is*

$$\widehat{\mu}_x^{01} = \frac{N^{01}}{V^0},$$

corresponding to the ratio of the number of deaths in the year of age x to $x+1$ over the total exposure of the N lives in the year of age x to $x+1$.

Before we discuss additional properties of the MLEs, we first review the multivariate version of the maximum likelihood theorem.

Theorem 22 *Suppose $\hat{\boldsymbol{\theta}}$ satisfies $\frac{\partial \ln L}{\partial \theta_j} = 0$ for each j and $\boldsymbol{\theta}_0$ is the true value of $\boldsymbol{\theta}$. Then $\hat{\boldsymbol{\theta}}$ is (i) asymptotically unbiased, i.e., $\mathrm{E}(\hat{\boldsymbol{\theta}}) \to \boldsymbol{\theta}_0$; (ii) asymptotically efficient, i.e., $\mathrm{Var}(\hat{\boldsymbol{\theta}}) \to I(\boldsymbol{\theta}_0)^{-1}$, where $I(\boldsymbol{\theta}_0)$ has entry (i,j) given by*

$$-\mathbb{E}\left[\frac{\partial^2 \ln L}{\partial \theta_i \partial \theta_j}\right]$$

(Fisher information function); (iii) asymptotically normal.

In our case,

$$L = \left(\prod_{j=0}^{n} e^{\mu_x^{jj} V^j}\right)\left(\prod_{j=0}^{n}\prod_{\substack{k=0\\k\neq j}}^{n}\left(\mu_x^{jk}\right)^{N^{jk}}\right)$$

and

$$\ln L = \sum_{j=0}^{n} \mu_x^{jj} V^j + \left(\sum_{j=0}^{n}\sum_{\substack{k=0\\k\neq j}}^{n} N^{jk} \ln \mu_x^{jk}\right).$$

We have

$$\frac{\partial^2 \ln L}{\partial \left(\mu_x^{jk}\right)^2} = \frac{-N^{jk}}{\left(\mu_x^{jk}\right)^2}, \qquad j \neq k,$$

while

$$\frac{\partial^2 \ln L}{\partial \mu_x^{jk} \partial \mu_x^{lm}} = 0,$$

when μ^{jk} and μ^{lm} are two distinct μ. It follows that

$$-\mathbb{E}\left[\frac{\partial^2 \ln L}{\partial \left(\mu_x^{jk}\right)^2}\right] = \frac{\mathbb{E}\left[N^{jk}\right]}{\left(\mu_x^{jk}\right)^2}.$$

Hence based on the Maximum Likelihood Theorem, we can say that the asymptotic distribution of $\widehat{\mu}_x^{jk}$ is normal with mean μ_x^{jk} and variance $\frac{\left(\mu_x^{jk}\right)^2}{\mathbb{E}\left[N^{jk}\right]}$. Also, the μ_x^{jk}'s are asymptotically independent. Note that in practice, $Var\left(\widehat{\mu}_x^{jk}\right) = \frac{\left(\mu_x^{jk}\right)^2}{\mathbb{E}\left[N^{jk}\right]}$ is rather approximated by

$$\widehat{Var}\left(\widehat{\mu}_x^{jk}\right) = \frac{\left(\widehat{\mu}_x^{jk}\right)^2}{N^{jk}} = \frac{\left(\frac{N^{jk}}{V^j}\right)^2}{N^{jk}} = \frac{N^{jk}}{\left(V^j\right)^2}.$$

Example 23 *Consider the 2-state alive-dead multiple state model where state 0 is the alive state and state 1 is the dead state. Suppose 450 individuals of age 50 have been observed over a period of (at most) one year with the following characteristics:*

- *325 persons were observed from age 50 until they died or reached 51 (whichever occurred first); of those, 1 died after 3 months, 1 after 5 months and 2 after 10 months.*
- *87 persons were observed starting at age 50 and 2 months until they died or reached 51; of those 1 died at age 50 and 7 months.*
- *The rest of the group was observed from age 50 until they died or reached 50 and 9 months; of those 1 died at 50 and 1 month and 1 died at 50 and 7 months.*

Compute the MLE for μ_{50}^{01}. Determine a 95% confidence interval for μ_{50}^{01}.

***Solution**: Here, N^{01} stands for the number of observed deaths in the sample with realization*

$$n^{01} = 7.$$

Also, V^0 is the total exposure (total time spent by the group of policyholders in state 0) with realization

$$\begin{aligned} v^0 &= \left(321 + \frac{3}{12} + \frac{5}{12} + 2 \cdot \frac{10}{12}\right) + \left(86 \cdot \frac{10}{12} + \frac{5}{12}\right) + \left((450 - 87 - 325 - 2) \cdot \frac{9}{12} + \frac{1}{12} + \frac{7}{12}\right) \\ &= 423.08. \end{aligned}$$

It follows that

$$\widehat{\mu}_{50} = \frac{7}{423.08} = 0.016545.$$

The variance of $\widehat{\mu}_{50}$ is approximated by

$$\widehat{Var}\left(\widehat{\mu}_{50}\right) = \frac{n^{01}}{\left(v^0\right)^2} = \frac{7}{\left(423.08\right)^2} = 3.9107 \times 10^{-5}.$$

A 95% confidence interval for μ_{50}^{01} is

$$\begin{aligned} \widehat{\mu}_{50} \pm 1.96\sqrt{\frac{7}{\left(423.08\right)^2}} &= 0.016545 \pm 1.96\sqrt{\frac{7}{\left(423.08\right)^2}} \\ &= \left(0.004288, 0.028802\right). \end{aligned}$$

1.6 Poisson model of mortality (skip)

Rather than modeling each life individually according to the alive/dead model, an alternative is to model the number of deaths D as a Poisson rv. The Poisson distribution can approximate the binomial distribution fairly well when n is very large and p is very small. So if we think of our experiment as observing N people who each have the small probability of dying $p = 1 - e^{-\mu}$ over the interval $[x, x+1)$ (note that this is not exactly our experimental set up since we may observe them for less than one year), then the Poisson approximation makes sense.

To explain how it works, we first introduce the notation

$$E_x = v,$$

which is the realized value of the total exposure time rv V^0 from a group of N policyholders over the year of age x and $x+1$. It is important to point out that we take this as a known quantity here.

The Poisson model then assumes that $D \sim Poisson(\mu E_x)$, and therefore we have that

$$\mathbb{P}(D = d) = \frac{(\mu E_x)^d e^{-\mu E_x}}{d!}, \qquad d = 0, 1, \ldots.$$

Why is this an approximation? For one thing, with this model we have $P(D > N) > 0$, which is obviously wrong. Note that this probability should be very small though, if N is large and μ is small.

With the Poisson model, the only data observed consists in the number d of deaths. Hence, the likelihood L is given by

$$L = \frac{(\mu E_x)^d e^{-\mu E_x}}{d!}.$$

Therefore, the MLE for μ is obtained as follows:

$$\ln L = d \ln(\mu E_x) - \mu E_x - \ln d!$$

and thus

$$\frac{\partial \ln L}{\partial \mu} = \frac{d}{\mu} - E_x = 0,$$

which implies that the MLE estimator is

$$\tilde{\mu} = \frac{D}{E_x}.$$

Note the difference with the maximum likelihood estimator we derived in the alive/dead model, which was instead given by $\tilde{\mu} = N^{01}/V^0$. So here V^0 is replaced by its observed value E_x (while N^{01} and D both stand for the number of deaths among the N policyholders). This implies that we can directly compute the expectation and variance of this estimator. Indeed, we get that $\mathbb{E}[\tilde{\mu}] = \mathbb{E}[D]/E_x = \mu E_x/E_x = \mu$, and $\mathrm{Var}(\tilde{\mu}) = \mu E_x/(E_x)^2 = \mu/E_x$. As in the alive/dead model, in practice we would estimate this variance by $d/(E_x)^2$.

1.7 Binomial model of mortality

Rather than modeling each life individually according to the alive/dead model, an alternative is to model the number of deaths D as a binomial rv. Suppose we observe N identical, independent lives aged x for exactly one year, and record the number d who die. This means d is a sample value observed from the rv D. If we assume that each life dies with probability q_x and survives with probability $1 - q_x$, then $D \sim Binomial(N, q_x)$ and $\hat{q}_x = \frac{d}{N}$ is the MLE for q_x, with corresponding estimator $\tilde{q}_x = \frac{D}{N}$ such that $\mathbb{E}[\tilde{q}_x] = q_x$ and $Var(\tilde{q}_x) = \frac{q_x(1-q_x)}{N}$. This is a simple version of the binomial model of mortality. This approach is often used in textbooks. What are the problems with it?

1. lives may not all be observed over the same interval of age

2. there might be other decrements than death.

Hence it is more realistic to introduce the times a_i (entry) and b_i (exit) as before. Our goal is then to get an estimator for q_x.

Let

$$D_i = \text{indicator for death of the } i\text{th life}, \qquad i = 1, \ldots, N.$$

Then

$$\mathbb{P}(D_i = d_i) = \left({}_{b_i-a_i}q_{x+a_i}\right)^{d_i} \left(1 - {}_{b_i-a_i}q_{x+a_i}\right)^{1-d_i}, \qquad d_i = 0, 1.$$

Let

$$\mathbf{q} = \left({}_{b_1-a_1}q_{x+a_1}, \ldots, {}_{b_N-a_N}q_{x+a_N}\right) \qquad \text{(parameters)}$$
$$\mathbf{d} = (d_1, \ldots, d_N) \qquad \text{(observed data)}.$$

Then we can write the total likelihood as

$$L = \prod_{i=1}^{N} \left({}_{b_i-a_i}q_{x+a_i}\right)^{d_i} \left(1 - {}_{b_i-a_i}q_{x+a_i}\right)^{1-d_i}$$

and our goal is to find the N values in the vector $\mathbf{q}$ that will maximize L. This is a difficult problem because we need to maximize L as a function of possibly N variables (less than N if some pairs (a_i, b_i) are equal). In order to make the optimization problem easier and also get smoother estimates, we will need additional assumptions. Two different assumptions are commonly used to reduce L to a function of one parameter (which is q_x). They are:

1. Uniform distribution of deaths (UDD)
2. Constant force of mortality (CFM)

Next, we will look into the UDD case by assuming that ${}_tq_x = t \cdot q_x$ for $0 \le t \le 1$. The CFM case can be handled similarly.

Hence,

$${}_{b_i-a_i}q_{x+a_i} = \frac{(b_i - a_i)q_x}{1 - a_i q_x}$$

Therefore, under UDD we have

$$L = \prod_{i=1}^{N} \left(\frac{(b_i - a_i)q_x}{1 - a_i q_x}\right)^{d_i} \left(1 - \frac{(b_i - a_i)q_x}{1 - a_i q_x}\right)^{1-d_i}$$

which is a function of only one variable: q_x. We can find the value $\hat{q}_x$ of q_x that maximizes L numerically using R or Matlab. We can then evaluate $(-\partial^2 \ln L/\partial q_x^2)^{-1}$ at $\hat{q}_x$ to estimate the variance of our estimator.

1.8 Graduation (skip)

1.8.1 Introduction

Up to now, a few mortality/survival models were presented and a special emphasis was given on the parameter estimation of the related models. In this section, we consider the resulting estimation of mortality rates (or probability of deaths) as a function of x, where x runs over all possible (integer) ages in our life table, and highlight a few challenges that may arise.

The models/methods we have discussed earlier give us estimates $\hat{\mu}_x$ (or $\hat{q}_x$) called *crude estimates* for each x. These crude estimates do not typically exhibit the level of regularity/smoothness required for mortality rates due to the random fluctuations in the sampling process (which is more severe when the exposure to risk at a given age x is low). As discussed in ACTSC 612/622, we typically expect mortality rates to be smooth with respect to age (and generally, increasing with age) because the aging process that causes the mortality rates to vary is assumed to be smooth and gradual. In fact, the premium rates calculated from these mortality rates are also expected to behave smoothly as a function of x.

Graduation is a process that takes crude rates as input and produces a set of smooth rates $\dot{\mu}_x$ or $\dot{q}_x$ (called *graduated rates*) that mimic the trends of the original crude rates closely enough. This does not remove the non-random systematic error resulting from poor data collection or inadequate statistical models/methods, so in order for this process to be successful we need to make sure these issues are properly addressed.

There are two main steps in the graduation process.

1. Construction of graduated rates (curve-fitting problem)
2. Test graduated rates for (i) smoothness and (ii) acceptable fit to original crude rates. Note that these two features can be contradictive (too much smoothness may imply bad fit and vice-versa)

We will start by discussing the second of the above steps.

1.8.2 Testing smoothness

In mathematics, the smoothness of functions is often measured by looking at partial derivatives up to a certain order and make sure they exist. They are also various concepts of variation that can be used. In practice, smoothness of mortality rates is often assessed by looking at the third differences of the graduated estimates, and making sure these differences are small compared to the rates themselves. Another aspect that may be considered is to look at the successive first, second and third-order differences and make sure they successively decrease. That is, we look at

$$\begin{aligned}\Delta\dot{q}_x &= \dot{q}_x - \dot{q}_{x-1}\\ \Delta^2\dot{q}_x &= \Delta\dot{q}_x - \Delta\dot{q}_{x-1}\\ \Delta^3\dot{q}_x &= \Delta^2\dot{q}_x - \Delta^2\dot{q}_{x-1}.\end{aligned}$$

This is not a formal test though, and as such we don't have a way to conclude whether or not the smoothness assumption should be rejected. It can however be useful to choose between two possible sets of graduated rates (i.e., could choose the one with the smallest third-order differences).

Example 24 *Here is an example of third order differences calculation (rates extracted from SSA Period Life Table 2009). We see that the third order differences are about 1000 times smaller than the original rates.*

x	$\dot{q}_x$	$\Delta\dot{q}_x$	$\Delta^2\dot{q}_x$	$\Delta^3\dot{q}_x$
50	*0.005347*			
51	*0.005838*	*0.000491*		
52	*0.006337*	*0.000499*	*0.000008*	
53	*0.006837*	*0.000500*	*0.000001*	*-0.000007*
54	*0.007347*	*0.000510*	*0.000010*	*0.000009*

1.8.3 Testing adherence to data

Next, we want to test whether or not a set of graduated rates $\dot{\mu}_x/\dot{q}_x$ is close enough to the crude rates. This is typically done by considering the death count rv's D_x and formulating the assumption (null hypothesis) as one in which D_x follows either a Poisson or binomial model, with parameters based on the graduated rates. More precisely, under the null hypothesis, for the Poisson model we view D_x ∼Poisson with mean $\dot{\mu}_{x+1/2}E_x$ under the null hypothesis, while for the binomial model we assume D_x is binomial with parameters (E_x, q_x) (if E_x is not an integer, one can round this number to the nearest integer). Also recall that implicit in the Poisson or binomial model, there is an assumption that the lives contributing to the exposed to risk are independent and identically distributed. This assumption has to be checked carefully in situations that threaten its validity, such as when ages are grouped together (e.g., 100–105) to increase the exposed-to-risk, or when the data has not sufficiently been divided into homogeneous groups (males vs females, smokers vs non-smokers).

Most statistical tests focus on the deviations between the actual death count rv D_x and its expectation under the chosen model, denoted by $\dot{D}_x$. That is, we look at $D_x - \dot{D}_x$, where

$$\dot{D}_x = \begin{cases} E_x\dot{q}_x, & \text{for the binomial model} \\ E_x\dot{\mu}_{x+1/2}, & \text{for the Poisson model} \end{cases}$$

The standardized deviations (or residuals) are

$$Z_x = \frac{D_x - E_x\dot{q}_x}{\sqrt{E_x\dot{q}_x\,(1-\dot{q}_x)}},$$

for the binomial model and

$$Z_x = \frac{D_x - E_x\dot{\mu}_{x+1/2}}{\sqrt{E_x\dot{\mu}_{x+1/2}}},$$

for the Poisson model. When E_x is sufficiently large, we can safely assume that under the assumption that D_x follows the chosen distribution (Poisson/binomial) with parameters given by the graduated rates (i.e., D_x ∼Binomial$(E_x, \dot{q}_x)$ or D_x ∼Poisson$(E_x\dot{\mu}_{x+1/2})$, then Z_x has a $N(0,1)$ distribution.

In addition, we assume that the Z_x's are independent. Note that this assumption is not entirely correct because (as we'll see later) the graduation is often performed so that, e.g., $\sum_x Z_x = 0$, which contradicts the assumption of independence.

Next, we will describe various tests that can be used to verify adherence to data.

Chi-square test To test the overall fit of the graduated rates, we can simply compute

$$W = \sum_x (Z_x)^2 . \tag{8}$$

The test statistic W has a chi-square distribution with L degrees of freedom (where L is the number of ages x entering the sum (8)) under the null hypothesis, because under H_0, each $Z_x \sim N(0,1)$ and therefore $Z_x^2 \sim \chi^2(1)$. However, note that if the observed values d_x for D_x used to compute W have also been used to estimate the parameters of the graduated rates, then the number of degrees of freedom will be $L-$number of parameters.

In this test, we wish to detect situations where the fit is not so good, as indicated by larger than usual deviations. So it is a one-sided test, in which we reject H_0 if $P(\chi^2(L-\text{number of parameters}) > w)$ is too small (less than say 5%-10%), where w is the observed value for W.

One should also be careful and watch out for cases where the nb of observations for a certain age x is too small. As a rule of thumb, it is suggested to make sure $\dot{D}_x \geq 5$, and if for certain ages it is not the case, then ages should be grouped together into a larger class.

Example 25 *(From T. Konstantapoulos on Survival models at Heriot Watt)*

In this example, the logistic regression model

$$\ln\left(\frac{q_x}{1-q_x}\right) = a_0 + a_1 x + a_2 x^2 \tag{9}$$

is fitted to a group of 16 crude rates for ages 12, 17,...,87, obtained using the data in the following table (ignore the two last columns for now):

x	E_x	d_x	$\dot{D}_x$	Z_x
12	*8119*	*14*	*18.78*	*-1.105*
17	*7750*	*20*	*19.68*	*0.0714*
22	*6525*	*22*	*18.84*	*0.7302*
27	*5998*	*23*	*20.37*	*0.5830*
32	*5586*	*26*	*23.11*	*0.6019*
37	*5245*	*28*	*27.36*	*0.1220*
42	*4659*	*32*	*31.72*	*0.0502*
47	*4222*	*37*	*38.81*	*-0.2915*
52	*3660*	*44*	*46.97*	*-0.4364*
57	*3012*	*54*	*55.77*	*-0.2391*
62	*2500*	*68*	*68.91*	*-0.1112*
67	*2113*	*87*	*89.25*	*-0.2433*
72	*1469*	*100*	*97.44*	*0.2679*
77	*883*	*95*	*93.56*	*0.1573*
82	*418*	*70*	*71.03*	*-0.1340*
87	*181*	*49*	*48.51*	*0.0822*

The parameters are estimated to be $\hat{a}_0 = -6.148874$, $a_1 = -0.001511$ *and* $a_2 = 0.000697$. *From (9) we can see that*

$$q_x = \frac{1}{1 + e^{-(a_0 + a_1 x + a_2 x^2)}}. \tag{10}$$

By replacing a_0, a_1, a_2 *by their estimated values, we get the graduated rates*

$$\dot{q}_x = \frac{1}{1 + e^{-(\hat{a}_0 + \hat{a}_1 x + \hat{a}_2 x^2)}}.$$

So for instance, for $x = 12$ *we get*

$$\dot{D}_{12} = E_{12}\dot{q}_{12} = 8119 \times 0.00231539 = 18.78.$$

Once we have all the $\dot{D}_x$*'s then we can compute the standardized deviations*

$$Z_x = \frac{d_x - E_x \dot{q}_x}{\sqrt{E_x \dot{q}_x (1 - \dot{q}_x)}}$$

which are found in the last column of the table.

The number of parameters to be used in the chi-square test is $16 - 3 = 13$, and we compute the test statistic to be $\sum_x Z_x^2 = 3.0044$. We have that the p-value is $P(\chi^2_{13} > 3.004) = 0.9979$, so we do not reject H_0.

It is clear that this test by itself is not enough to test adherence to data. For instance, it cannot detect cases where $\dot{q}_x$ systematically underestimates or overestimates q_x, or even when the underestimation or overestimation appears in a non-random way (e.g., rates underestimated before age x_0 and overestimated after x_0). Some of the upcoming tests will be able to detect such issues.

Standardized Deviations Test Here we test the assumption that the Z_x's are iid $N(0,1)$ and see how well they fit the normal distribution. To do so, we divide the real axis $(-\infty, \infty)$ into p sub-intervals, and compare the number of Z_x's that fall in each with its expected number. The sub-intervals are usually chosen to either have equal length or probability. Denote the jth interval by I_j.

More precisely, let

$$M_j = \text{nb of } Z_x \text{ that fall in } j\text{th interval } I_j$$
$$m_j = \frac{M_j - \mathrm{E}(M_j)}{\sqrt{\mathrm{E}(M_j)}}.$$

where $\mathrm{E}(M_j) = L\theta_j$, where L is the number of Z_x's we have (i.e., the number of ages x in the table) and $\theta_j = \mathbb{P}(N(0,1) \in I_j)$. Note that the vector $(M_1, \ldots, M_p)$ has a multinomial distribution with parameters $(\theta_1, \ldots, \theta_p)$. It is exactly the type of frequency vector used to form a Pearson test statistic of the form

$$\sum_{j=1}^{p} (m_j)^2, \tag{11}$$

which under H_0 can be proven to have a chi-square distribution with $p - 1$ degrees of freedom (we lose one degree because we force $\sum M_j = \sum \mathrm{E}(M_j) \Rightarrow \sum m_j = 0$). This is a one-sided test, where our concern is when

the m_j's are too big. So if we find that $\mathbb{P}(\chi^2(p-1) > y_0)$, where y_0 is the value taken by (11) is small, we reject H_0.

Using the above data, if we use the 4 intervals $I_1 = (-\infty, -1]$, $I_2 = (-1, 0]$, $I_3 = (0, 1]$, and $I_4 = (1, \infty)$. So we have $\theta_1 = \theta_4 = 0.159$ and $\theta_2 = \theta_3 = 0.5 - 0.159 = 0.341$. Then we get $M_1 = 1, M_2 = 6, M_3 = 9, M_4 = 0$. The corresponding test statistic is

$$\frac{(1 - 16 \times 0.159)^2}{16 \times 0.159} + \ldots + \frac{(0 - 16 \times 0.159)^2}{16 \times 0.159} = 5.87.$$

We compare this against a chi-square with 3 degrees of freedom and find $p = P(\chi^2(3) > 5.87) = 0.1181$, which is a bit small and suggests the fit may not be so good.

1.8.4 Methods of Graduation

We will now discuss two possible approaches to perform a graduation. That is, to decide which function of x will be used to represent the behavior of either q_x or μ_x based on the estimated crude rates.

Graduation by reference to a standard table This method can be used when we have a mortality study with a relatively small amount of data, and suspect lives under consideration are similar to those whose experience forms the basis of a related standard table. The standard table (based on a larger number of lives) is there to impose a basic structure on our new graduation. The level of the new rates can be different from those in the standard table but should be related to them in a simple way.

More precisely, assume the rates in the standard table are denoted by q^s_x or μ^s_x, and denote the new graduated rates by $\dot{q}_x$ or $\dot{\mu}_x$. Examples of functions that can be used to relate the two sets of rates are:

1. $\dot{q}_x = aq^s_x$
2. $\dot{q}_x = aq^s_x + b$
3. $\dot{q}_x = (ax + b)q^s_x$
4. $\dot{q}_x = q^s_{x+k}$,which is known as age rating with $k = \pm 1, \pm 2$ (used for mortality improvements)
5. $\dot{q}_x = aq^{s_1}_x + bq^{s_2}_x$ which consists in mixing two standard tables.

We now explain the four steps of this type of graduation method:

1. Choose the standard table: this should be guided by the lives under consideration, looking at factors such as range of ages, period of time, selection effect, type of insurance, geographical location
2. Determine the type of relation between crude rates and standard table: certain tricks to help there are (i) plot $\hat{q}_x$ against q^s_x and see if the relation looks linear; (ii) plot $\hat{q}_x/q^s_x$ against x; (iii) transform the data, e.g., plot $\log\left(\frac{1-\hat{q}_x}{1-q^s_x}\right)$ against x.

3. Once the relation between $\hat{q}_x$ and q_x^s is determined, then the parameters of this relation need to be estimated. This is often done using weighted least-squares regression, where, for example if we think $\dot{q}_x = aq_x^s + b$, then we would find a and b that minimize

$$\sum_x w_x(\hat{q}_x - aq_x^s - b)^2$$

where the weights w_x can be used to give more weights to certain ages where the confidence in the data and/or the relation is higher, e.g., use $w_x = E_x$ or $E_x/\hat{q}_x$.

4. Perform statistical tests as seen before (to check adherence to data).

Advantages of graduation by reference to a standard table

- Can be used when we don't have too much data
- The information from the more extensive study that led to the standard table is used
- With a simple relation between new rates and standard rates, the graduated rates are guaranteed to be smooth since the standard table will be smooth by definition

Disdvantages of graduation by reference to a standard table

- Not always clear which standard table to use
- Adherence to data might be hard to get
- Features from the standard table that are not relevant for current data will show up in graduated rates
- Should not be used if we have enough data to do a graduation from scratch (next method)

Example 26 *Suppose we have rates from age 80 to 99 from the Scottish female population and want to perform a graduation by using the corresponding ages from a broader UK graduation. Here is a plot of the ratios $\hat{q}_x/q_x^s$ against x*

From this plot it looks like a relation of the form $\dot{q}_x = aq_x^s$ would be reasonable. Suppose we want to estimate a by using weighted least squares-regression with weights $w_x = E_x/q_x$. Then we are looking for a that minimizes

$$S(a) = \sum_{x=80}^{99} \frac{E_x}{\hat{q}_x}(\hat{q}_x - aq_x^s)^2.$$

We compute

$$\frac{d}{da}S(a) = -2\sum_{x=80}^{99} q_x^s \frac{E_x}{\hat{q}_x}(\hat{q}_x - aq_x^s)$$

and by setting it to 0 we get

$$a = \frac{\sum_{x=80}^{99} E_x q_x^s}{\sum_{x=80}^{99} \frac{E_x}{\hat{q}_x}(q_x^s)^2}.$$

With the summary data $\sum_{x=80}^{99} E_x q_x^s = 3224.806$ and $\sum_{x=80}^{99}(q_x^s)^2 \frac{E_x}{\hat{q}_x} = 2698.292$, we obtain $\hat{a} = 1.195$. (see lec17.xlsx on LEARN).

Below we plot the crude rates (series 1) and the graduated rates $1.195q_x^s$ (series 5).

Graduation by mathematical formula This is the preferred method when we have a large amount of data. Since we are fitting a simple function of x through the crude rates, it automatically gets the desired smoothness, but the adherence to data might be problematic. To help with that, choosing functions with more parameters or fitting different functions to different age ranges can help (note that the latter also amount to increase the number of parameters). For instance, Heligman and Pollard (1980) suggested the formula

$$\frac{q_x}{1-q_x} = A^{(x+B)^c} + De^{-E(\ln x - \ln F)^2} + GH^x$$

in which the first component models infant mortality (rapidly decreasing exponential term), the second term represents the accident hump (includes maternal mortality), while the third term is a Gompertz term that accounts for general mortality due to the natural aging process.

For the force of mortality, we start by reviewing two well-known parametric models, namely

- Gompertz: $\mu_x = Bc^x$
- Makeham: $\mu_x = A + Bc^x$

A generalized version of this is the Perks formula

$$\mu_x = \frac{A + Bc^x}{1 + Dc^x}.$$

In the Perks model (contrary to the Gompertz and Makeham models), when $D > 0$, the force of mortality tends to a constant when x gets large. Recent modeling of mortality trends at very old ages seems to support this feature.

Another way to get this kind of behavior is to use a model of the form

$$logit(q_x) = \ln\left(\frac{q_x}{1-q_x}\right) = a_0 + a_1 x + a_2 x^2 + \dots$$

An advantage of working with this type of transformation is that there is no restriction on the range allowed for $\text{logit}(q_x)$, while when working directly with q_x we are forced to stay in $[0, 1]$.

Another example is the Kannisto model used to model the mortality rates for $x \geq 95$ in the 2005–2007 life tables produced by Statistics Canada. It is based on the formula

$$logit(\mu_x) = \alpha e^{\beta x}$$

or equivalently

$$\mu_x = \frac{\alpha e^{\beta x}}{1 + \alpha e^{\beta x}}.$$

Most of the above models can fit into what is called the *Gompertz-Makeham family*. Let the vector of parameters defining this type of model be denoted by

$$\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \dots, \alpha_{r+s}),$$

where r and s are two non-negative integers with $r + s \geq 1$. We say that the *Gompertz-Makeham formula of type* (r, s) is given by

$$GM^{r,s}_{\boldsymbol{\alpha}}(x) = \sum_{i=1}^{r} \alpha_i \; x^{i-1} + \exp\left(\sum_{i=1}^{s} \alpha_{r+i} x^{i-1}\right).$$

For example, the Gompertz model translates to $\mu_x = GM^{0,2}_{\boldsymbol{\alpha}}(x)$ because

$$\mu_x = e^{\alpha_1+\alpha_2 x} = e^{\alpha_1}(e^{\alpha_2})^x = Bc^x,$$

where $B = e^{\alpha_1}$ and $c = e^{\alpha_2}$. Similarly, the Makeham formula amount to use $\mu_x = GM^{1,2}_{\boldsymbol{\alpha}}(x)$.

We may also use the GM family to model $q_x/(1-q_x)$ (as was done in the paper by Heligman and Pollard) or $logit(q_x) = \log(q_x/(1-q_x))$. For example, we can use

$$\frac{q_x}{1-q_x} = GM^{r,s}_{\boldsymbol{\alpha}}(x)$$

which is equivalent to

$$q_x = \frac{GM^{r,s}_{\boldsymbol{\alpha}}(x)}{1+GM^{r,s}_{\boldsymbol{\alpha}}(x)}.$$

For instance, the UK table for assured lives 1967-1970 was graduated using

$$\frac{q_x}{1-q_x} = A + Hx + Bc^x$$

which is a $GM^{2,2}_{\boldsymbol{\alpha}}(x)$ function, and is therefore equivalent to

$$q_x = \frac{GM^{2,2}_{\boldsymbol{\alpha}}(x)}{1+GM^{2,2}_{\boldsymbol{\alpha}}(x)}.$$

The GM family covers a wide variety of models, and assuming we choose to work with this family, the first thing to do is choose r and s, i.e., how many parameters will be used. This can be done by plotting the crude rates and comparing with different (low order) members of the GM family. Once r and s are chosen, the next step is to estimate the parameters. While there are a few options to do this (including weighted least-squares), here we choose to focus on maximum likelihood estimation.

2 Insurance Pricing using GLMs

A comprehensive coverage of *generalized linear models (GLMs)* was carried out in ACTSC 623. In this module, we focus on some actuarial science applications of these statistical models. In particular, we show how GLMs can be used to model the frequency and severity components of the aggregate loss/payment model for an insurance portfolio (ACTSC 625).

2.1 Introduction

In a typical non-life insurance policy, the insurer agrees to compensate a policyholder against incurred losses over a given coverage period (usually, one year) in return for a premium (usually, paid up front). These losses may be related to damages sustained to a property, bodily injury to policyholder(s) or a third party, or others.

To determine this premium, the insurer usually makes modelling assumptions on the frequency and severity of claims to be made by the policyholder (or a group of policyholders) over the course of the coverage period. As such, in a non-life insurance, the usual context consists in:

- **Response variable** y: number of claims **(frequency)** or amount paid per claim **(severity)**
- **Explanatory variables** x: usually a rather large set of explanatory variables, either related to the policyholders, insured object, geographic region, etc.

A few observations:

- the response variables are not known ahead of time; predicting them is critical for the determination of a fair premium to take on the risk;
- the explanatory variables may (or may not) have predictive power to explain past observed response variables;
- a theoretical methodology is needed to determine which explanatory variables are statistically meaningful to predict the response variable(s) and ultimately set the premium.

In a competitive market, the insurer should charge a fair premium on a policy which should be in line with the expected total claim amount for the policy (or a group of policies). Obviously, it is expected that loading factors or margins will be added to the fair premium to appropriately compensate the insurer for accepting the risk. The introduction of these loading factors or margins goes beyond the scope of this course.

We focus our attention on the **fair premium calculation** which will vary between policies depending on some **rating factors** (explanatory variables/predictors) that need to be identified. These rating factors will be used to quantify the risk associated to a given policy.

Remark 27 *Throughout this module, we assume that rating factors are treated as categorical variables (e.g., age of driver is divided into groups). In other words, we partition the population (i.e., entire group of policyholders) into a finite set n of tariff cells. A tariff cell is a group of "fairly homogeneous" policyholders with the*

same rating factors. Each tariff cell gets its own premium. The idea is that a tariff cell contains policyholders with a very similar "risk level" and as such, should pay the same premium. The process that determines these premiums is called a tariff analysis*.*

Remark 28 *In light of the above remark, it is assumed that the insurer will include a quantitative predictor into the tariff analysis by first turning it into a categorical variable (i.e., by creating a number of "non-overlapping" categories for the possible outcomes of this predictor). As an example, one can consider a predictor* KILOMETERS *which is the reported number of kilometers expected to be put on the vehicle over the next year. Such a predictor may be incorporated into the tariff analysis by turning it into the following categorical predictor (which we shall call* MILEAGE*):*

$$\text{MILEAGE}=\begin{cases} 1, & 0 \leq \text{KILOMETERS} < 8,000 \\ 2, & 8,000 \leq \text{KILOMETERS} < 16,000 \\ 3, & 16,000 \leq \text{KILOMETERS} < 25,000 \\ 4, & \text{KILOMETERS} \geq 25,000 \end{cases}$$

2.2 Basic Modelling Assumptions

In the following tariff analysis, the following assumptions will be made:

- Assumption 1: All policies are mutually independent
- Assumption 2: Time independence (for a policy over disjoint intervals)
- Assumption 3: For any 2 policies in a tariff cell, their total claim amounts over any period of time of a given duration have the same probability distribution.

In reality, any of the above assumptions can be violated. For instance, a common violation of Assumption 1 arises in catastrophic insurance when a single event can trigger claims for a number of insurance policies. Assumption 2 may not hold if it is expected that the claim experience in a given period may affect the claim experience in a later period (positively or adversely for a car driver, for instance).

Remark 29 *Note that under Assumption 3, we are justified to charge the same premium for each policy within the tariff cell. If we rather believe that Assumption 3 does not hold, then it is expected that the tariff model would be refined (if possible - by e.g., adding more tariff factors/cells) to reduce the non-homogeneity factor to a more acceptable level. For now, we assume that Assumption 3 holds.*

In general, if any of these assumptions are severily violated, it would be non-advisable to go ahead with the analysis of this chapter.

2.3 Context and Definitions

To perform the tariff analysis, we assume that the insurer possesses information on the claim history of all policies within each tariff cell at an aggregate level. In other words, this information will be aggregated among all policies within a tariff cell (rather than being at the policy level).

Example 30 *Consider the* MOPED *insurance dataset from the Swedish company Wasa (1999) which can be viewed in R by typing the command:*

```
print(moped <- read.table("http://www.karlin.mff.cuni.cz/~pesta/NMFM402/moped.txt", header=T))
```

As we can see, the Wasa tariff is based on three rating factors:

- *Age of vehicle (two categories: at most 1 year or otherwise)*
- *Vehicle class (two categories: weight* ≥ 60 *kgs and* > 2 *gears; all others)*
- *Geographic zone: divided into 7 zones*

The tariff analysis contains a total of 28 tariff cells ($2 \times 2 \times 7$)*. Each tariff cell (row) provides information on the claim history for the group of policies in a tariff cell (at the aggregate level):*

- ***duration****: amount of time all policies in a given tariff cell were in force;*
- ***nb of claims****: total number of claims made by the group of policies in the tariff cell;*
- ***claim frequency:*** $\frac{nb\ of\ claims}{duration}$ *(measure of the average number of claims per duration unit)*
- ***claim severity****: average claim size in the tariff cell*
- ***pure premium****: Claim frequency* $\times$ *Claim severity (measure of the average claim amount per duration unit)*
- ***actual premium:*** *premium charged by the insurer per duration unit (contains expense loading, cost of capital charges, etc. so not directly comparable to pure premium).*

We may zoom on one of the tariff cells, say cell [1,2,5], which reports

Duration	*nb of claims*	*Claim frequency*	*Claim severity*	*Pure premium*	*Actual premium*
114.1	*2*	$\frac{2}{114.1} = 1.7528 \times 10^{-2}$	11131	$11131 \times \frac{2}{114.1} \approx 195$	594

It appears that the actual premium of 594 *is significantly greater than the pure premium of* 195*. The reverse conclusion holds for tariff cell [1,1,7] where the actual premium is* 396 *vs a pure premium of* 1829*.*

As in ACTSC 625, we propose to separately handle the frequency and severity component of the modelling exercise.

2.4 Loss Models

2.4.1 Claim frequency

Our goal is to consider possible models for the key frequency ratio Y_i for tariff cell i which will be of the form

$$Y_i = \frac{X_i^{w_i}}{w_i},$$

where

- w_i stands for the duration in tariff cell i (among all policies)
- $X_i^{w_i}$ stands for the total number of claims made in tariff cell i (among all policies over the duration w_i).

Here, w_i is assumed to be given (deterministic). For each tariff cell i, we are provided with the observed outcome for the rv Y_i, namely y_i.

Remark 31 *Consider tariff cell [1,2,5] of the* MOPED *insurance dataset. For this tariff cell, we have* $w_i = 114.1$ *and*

$$y_i = \frac{2}{114.1}.$$

In this sub-section, we only consider the classical relative Poisson model for Y_i. Other models (including those with overdispersion) can also be considered. The relative Poisson model naturally arises in the context of loss modelling. Indeed, a simple but common modelling assumption for the claim arrival process is that claims for any policy in tariff cell i arrive according to a Poisson process with claim arrival rate μ_i (per unit time). For a tariff cell i with duration w_i (duration of the group of policies in tariff cell i), it follows from Assumptions 1 to 3 that the total number of claims $X_i^{w_i}$ for the entire tariff cell i follows a Poisson distribution with mean $w_i\mu_i$. Consequently, the frequency key ratio Y_i has support on the multiples of $\frac{1}{w_i}$ (i.e. $Y_i \in \left\{0, \frac{1}{w_i}, \frac{2}{w_i}, ...\right\}$) with probability mass function

$$\begin{aligned}\mathbb{P}\left(Y_i = \frac{k}{w_i}\right) &= \mathbb{P}\left(X_i^{w_i} = k\right) \\ &= \frac{\left(w_i\mu_i\right)^k e^{-w_i\mu_i}}{k!},\end{aligned} \tag{12}$$

for $k = 0, 1, ...$ Alternatively,

$$\mathbb{P}\left(Y_i = y_i\right) = \frac{\left(w_i\mu_i\right)^{w_i y_i} e^{-w_i\mu_i}}{k!}, \tag{13}$$

for $y_i \in \left\{0, \frac{1}{w_i}, \frac{2}{w_i}, ...\right\}$.

The distribution of Y_i is called the **relative Poisson distribution** (which is Poisson if $w_i = 1$).

2.4.2 Claim severity

We now shift our attention to the severity component of the loss modelling exercise. The goal is to consider possible models for the key severity ratio Y_i for tariff cell i which will again be of the form

$$Y_i = \frac{X_i^{w_i}}{w_i},$$

where

- w_i now stands for the total number of claims made in tariff cell i;
- $X_i^{w_i}$ now stands for the total amount of claims in tariff cell i.

Once again, w_i is assumed to be given (deterministic). Note that w_i is a nonnegative integer in this context. For each tariff cell i, we are provided with the observed outcome for the rv Y_i, namely y_i.

Remark 32 *Consider tariff cell [1,2,5] of the* MOPED *insurance dataset. For this tariff cell, we have* $w_i = 2$ *and*

$$y_i = 11131 = \frac{11131 * 2}{2}.$$

To model the claim severity, we should use a distribution defined on $\mathbb{R}^+$ that is skewed to the right (i.e, with a heavy tail for large claim severities) for conservatism. A common choice to begin the analysis is the gamma distribution.

Consider that all claims in tariff cell i are gamma distributed with mean μ_i and variance ξ_i, independently of one another. Then,

- $X_i^{w_i}$ is gamma distributed with mean $w_i\mu_i$ and variance $w_i\xi_i$ (iid sum of gamma rv's is also gamma);
- Y_i is gamma distributed with mean μ_i and variance $\frac{\xi_i}{w_i}$.

It follows that the density of Y_i is given by

$$\begin{aligned} f_{Y_i}(y) &= \frac{\left(\frac{\mu_i}{\xi_i}w_i\right)^{\frac{w_i(\mu_i)^2}{\xi_i}} y^{\frac{w_i(\mu_i)^2}{\xi_i}-1} e^{-\frac{w_i\mu_i}{\xi}y}}{\Gamma\left(\frac{w_i(\mu_i)^2}{\xi_i}\right)} \\ &= \frac{\left(\frac{\mu_i}{\xi_i}w_i\right)^{\frac{w_i(\mu_i)^2}{\xi_i}} e^{-\frac{1}{\frac{\xi_i}{(\mu_i)^2 w_i}}y\frac{1}{\mu_i}}}{\Gamma\left(\frac{w_i(\mu_i)^2}{\xi_i}\right)} y^{\frac{w_i(\mu_i)^2}{\xi_i}-1}. \end{aligned} \tag{14}$$

For both the relative Poisson frequency model and the gamma severity model, the rating factors (explanatory variables) will enter the model through the mean $\mu_i = \mathbb{E}[Y_i]$. We formalize this connection via the theory on GLMs in the next subsection.

2.5 Basics of pricing with GLMs

The goal is to explain how the key frequency/severity ratio Y_i varies according to different rating factors. We heavily rely on the theory of GLMs to do so. GLMs can be viewed as generalizations of multiple linear regression models. GLMs allow for more modelling flexibility and have a rich theory in statistics (estimate standard deviation, build CI, do model selection) which will be helpful in the present context.

In what follows, we consider GLMs for the key ratio Y_i where Y_i is an **exponential dispersion model (EDM)** and the so-called **link function** will be used to relate how the key ratio behaves in relation to the rating factors.

Next, we discuss the exponential dispersion models (EDMs) and show that the relative Poisson frequency model of Section 2.4.1 and the gamma severity model of Section 2.4.2 are special cases. We later discuss the role of the link function in GLMs.

2.5.1 Exponential dispersion models

Let y_i be the observed key ratio for the i-th tariff cell where Y_i is the corresponding rv. We assume the corresponding exposure/weight to be w_i. We say that Y_i is an EDM if its density (or probability mass function) is of the form

$$f_{Y_i}(y_i, \theta_i, \phi) = \exp\left\{ \frac{y_i\theta_i - b(\theta_i)}{\frac{\phi}{w_i}} + c(y_i, \phi, w_i) \right\}, \tag{15}$$

where θ_i is the so-called canonical parameter (which is allowed to depend on the tariff cell i), $\phi > 0$ is the dispersion parameter (which is assumed common for all tariff cells) and $w_i > 0$. Also, b is a twice continuously differentiable function with invertible first derivative and the parameter θ_i takes values in an open set. As we will see, the function c in (15) is not important in the application of EDMs in a GLM context.

For the model (15), its moment generating function exists and is given by

$$\mathbb{E}\left[e^{tY_i}\right] = e^{\Psi(t)},$$

with

$$\Psi(t) = \frac{b\left(\theta_i + \frac{t\phi}{w_i}\right) - b(\theta_i)}{\frac{\phi}{w_i}},$$

for at least some t in a neighborhood of 0 (i.e., for $|t| < \delta$). Using properties of mgfs, we can show that

$$\mathbb{E}[Y_i] = b'(\theta_i),$$

and

$$Var(Y_i) = \frac{\phi b''(\theta_i)}{w_i}. \tag{16}$$

(see Q1 in the Module 2 - Recommended problems). Letting $\mu_i := \mathbb{E}[Y_i]$, it follows that

$$\theta_i = \left(b'\right)^{-1}(\mu_i),$$

which allows to rewrite (16) as

$$Var(Y_i) = \frac{\phi v(\mu_i)}{w_i},$$

where v is known as the **variance function** defined as $v(\mu_i) = b''\left((b')^{-1}(\mu_i)\right)$.

Remark 33 *Eq. (16) explains why ϕ is called the dispersion parameter. The variance of Y_i is proportional to the value of this dispersion parameter.*

In the next two examples, we show that the relative Poisson frequency model of Section 2.4.1 and the gamma severity model of Section 2.4.2 are special cases of the EDM (15).

Example 34 *(Relative Poisson model) From the pmf (13) for the relative Poisson model, we find*

$$\begin{aligned}\mathbb{P}(Y_i = y_i) &= (\mu_i)^{w_i y_i} e^{-w_i \mu_i} \frac{(w_i)^{w_i y_i}}{(w_i y_i)!} \\ &= e^{w_i y_i \ln(\mu_i) - w_i \mu_i} \frac{(w_i)^{w_i y_i}}{(w_i y_i)!} \\ &= e^{\frac{y_i \ln(\mu_i) - \mu_i}{\frac{1}{w_i}}} \frac{(w_i)^{w_i y_i}}{(w_i y_i)!}. \end{aligned} \tag{17}$$

We conclude that (17) is of the form (15) with $\theta_i = \ln \mu_i$, $b(\theta_i) = e^{\theta_i}$ and $\phi = 1$. Also, the variance function is $v(\mu_i) = b''\left((b')^{-1}(\mu_i)\right) = \mu_i$, a result which is not surprising as for a Poisson rv, the variance is equal to its mean.

Example 35 *(Gamma model) For (14) to be of the form (15), we shall assume that the ratio $\frac{\xi_i}{(\mu_i)^2}$ does not depend on i. Hence, letting $\phi = \frac{\xi_i}{(\mu_i)^2}$ (which is fixed for all tariff cells), (14) becomes*

$$\begin{aligned} f_{Y_i}(y) &= \frac{\left(\frac{w_i}{\phi \mu_i}\right)^{\frac{w_i}{\phi}} e^{-\frac{1}{\frac{\phi}{w_i}} y \frac{1}{\mu_i}}}{\Gamma\left(\frac{w_i}{\phi}\right)} y^{\frac{w_i}{\phi} - 1} \\ &= \left(\frac{1}{\mu_i}\right)^{\frac{w_i}{\phi}} e^{-\frac{1}{\frac{\phi}{w_i}} y \frac{1}{\mu_i}} \frac{\left(\frac{w_i}{\phi}\right)^{\frac{w_i}{\phi}} y^{\frac{w_i}{\phi} - 1}}{\Gamma\left(\frac{w_i}{\phi}\right)}. \end{aligned}$$

Further letting $\theta_i = -\frac{1}{\mu_i}$, we obtain

$$\begin{aligned} f_{Y_i}(y) &= \frac{(-\theta_i)^{\frac{w_i}{\phi}} e^{\frac{1}{\frac{\phi}{w_i}} y \theta_i}}{\Gamma\left(\frac{w_i}{\phi}\right)} \left(\frac{w_i}{\phi}\right)^{\frac{w_i}{\phi}} y^{\frac{w_i}{\phi} - 1} \\ &= \frac{e^{\frac{1}{\frac{\phi}{w_i}} (y\theta_i + \ln(-\theta_i))}}{\Gamma\left(\frac{w_i}{\phi}\right)} \left(\frac{w_i}{\phi}\right)^{\frac{w_i}{\phi}} y^{\frac{w_i}{\phi} - 1}. \end{aligned}$$

Thus, $b(\theta_i) = -\ln(-\theta_i)$. Also, the variance function v is $v(\mu_i) = b''\left((b')^{-1}(\mu_i)\right) = (\mu_i)^2$.

As both the relative Poisson model and the gamma model are special cases of EDMs, the later analysis will simply assume that the key ratio Y_i is an EDM member with density (or pmf) given by (15).

2.5.2 Link function

In GLMs, the link function provides much needed flexibility to model how the response (here, the key ratio Y_i) relates to the rating factors. In other words, the link function is the technical tool used to incorporate the rating factors into an EDM model to yield a fully defined GLM. This will be done by imposing a relationship between the mean response $\mathbb{E}[Y_i] = \mu_i$ and the set of rating factors.

Let's consider two simple examples before formally introducing the link function.

Example 36 *Consider a tariff model with only one rating factor taking possible values* $\{a, b, c\}$*. Given that the rating factor is a categorical variable, one category will be the base category for the rating factor (say, category a) and 2 binary variables will be created to indicate whether the category is b or not, and whether the category is c or not. As such, we define*

$$\eta_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2},$$

where $x_{i1} = 1$ *if tariff cell i has category b for its only rating factor (or otherwise 0) and* $x_{i2} = 1$ *if tariff cell i has category c for its rating factor (or otherwise 0). It follows that*

Tariff cell i	*Rating factor*	η_i
1	(a)	β_0
2	(b)	$\beta_0 + \beta_1$
3	(c)	$\beta_0 + \beta_2$

Example 37 *Consider now a tariff model with two rating factors:*

- *Rating factor 1 takes on two possible values* $\{a, b\}$*;*
- *Rating factor 2 takes on three possible values* $\{c, d, e\}$*.*

As for the previous example, one shall create 1 binary variable for Rating factor 1 (assuming that the base category for the first rating factor is a), and 2 binary variables for Rating factor 2 (assuming that the base category for the second rating factor is c). Hence, we define

$$\eta_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3},$$

where

- $x_{i1} = 1$ *if tariff cell i has category b for its first rating factor (or otherwise 0);*
- $x_{i2} = 1$ *if tariff cell i has category d for its second rating factor (or otherwise 0)*
- $x_{i3} = 1$ *if tariff cell i has category e for its second rating factor (or otherwise 0)*

It follows that

Tariff cell i	*Rating factors*	η_i
1	(a, c)	β_0
2	(a, d)	$\beta_0 + \beta_2$
3	(a, e)	$\beta_0 + \beta_3$
4	(b, c)	$\beta_0 + \beta_1$
5	(b, d)	$\beta_0 + \beta_1 + \beta_2$
6	(b, e)	$\beta_0 + \beta_1 + \beta_3$

Recall that each rating factor is assumed to be a categorical variable in the present tariff analysis. As such, for a rating factor with say p categories, we shall create $(p-1)$ binary variables, each of which takes the value 1 if tariff cell i is in a given category and 0, otherwise. One of the categories is assumed to be the base category which explains why we create only $(p-1)$ binary variables. We repeat this process for each rating factor.

Remark 38 ***But how to determine the base category for each rating factor?*** *This process is an arbitrary one. It is a common practice to designate the tariff cell with the largest duration as the* ***base tariff cell****. As such, we designate as the base category for a rating factor to be the category of this rating factor for the base tariff cell.*

For a tariff model with a total of r binary variables, let $\mathbf{x}_i = (x_{i1}, ..., x_{ir})$ be the set of binary variables for tariff cell i, and define

$$\eta_i = \beta_0 + \sum_{j=1}^{r} x_{ij}\beta_j. \tag{18}$$

Remark 39 *Note that each η_i is unique to its tariff cell (i.e., $\eta_i \neq \eta_j$ for $i \neq j$). For instance, for the base tariff cell, all x_{ij}'s are 0 and $\eta_i = \beta_0$. This is not the case for the other tariff cells.*

For simplicity, we rewrite (18) as

$$\eta_i = \sum_{j=0}^{r} x_{ij}\beta_j,$$

where $x_{i0} = 1$ for all i. The link function g is what ties up the mean response $\mu_i = \mathbb{E}[Y_i]$ to η_i via the relationship

$$g(\mu_i) = \eta_i,$$

where g is assumed to be monotone and differentiable.

A few examples:

- Linear models arise by choosing $g(\mu) = \mu$ (i.e., g is the *linear* link function) and

$$\mu_i = \eta_i$$

- Multiplicative models arise by choosing $g(\mu) = \ln \mu$ (i.e., g is the *logarithmic* link function) and

$$\mu_i = e^{\eta_i}$$

- For a binary response (i.e., $Y_i \in \{0, 1\}$ with $\mu_i \in [0, 1]$), it may be best to use the *logit* link function

$$g(\mu) = \ln \frac{\mu}{1-\mu},$$

For tariff cell i, this implies that

$$\mu_i = \frac{e^{\eta_i}}{1 + e^{\eta_i}} \in [0, 1].$$

This topic will be covered more thoroughly in the logistic regression module of this course.

For practical considerations, we consider multiplicative models for the pricing applications of GLMs (i.e., the link function is the logarithmic link function $g(\mu) = \ln \mu$). They are widely used in practice (and ensures that $\mu_i = e^{\eta_i}$ is always positive - as it should be!).

Remark 40 *For a multiplicative tariff model, the pure premium for tariff cell i is given by*

$$\mu_i = e^{\eta_i} = e^{\beta_0 + \sum_{j=1}^{r} x_{ij}\beta_j}.$$

For the base tariff cell, all x_{ij} are set to be 0 which implies that

$$\mu_i = e^{\beta_0}$$

For the other tariff cells, not all x_{ij} are 0 and hence,

$$\begin{aligned}
\mu_i &= e^{\beta_0 + \sum_{j=1}^{r} x_{ij}\beta_j} \\
&= e^{\beta_0} \prod_{j=1}^{r} e^{x_{ij}\beta_j} \\
&= \underbrace{e^{\beta_0}}_{\text{premium for the tariff cell}} \cdot \left\{ \prod_{j=1}^{r} \underbrace{\left(e^{\beta_j}\right)}_{\text{Relativity related to binary variable } x_{ij}}^{x_{ij}} \right\}
\end{aligned}$$

We refer to the terms e^{β_j} $(j = 1, 2, ..., r)$ as relativities as they are multipliers added to the pure premium of the base tariff cell to determine the pure premium for all other tariff cells in the model.

2.6 Parameter estimation

For the GLM applications in non-life insurance, we assume that the key ratios Y_i $(i = 1, ..., n)$ are from the EDM class with $Var(Y_i) = \frac{\phi v(\mu_i)}{w_i}$. Also, the mean $\mu_i = \mathbb{E}[Y_i]$ satisfies $g(\mu_i) = \eta_i = \sum_{j=0}^{r} x_{ij}\beta_j$ where g is the link function.

For a given choice of $v()$ and $g()$, we are left with estimating the parameters $\overrightarrow{\beta} = (\beta_0, \beta_1, ..., \beta_r)$ and ϕ (if necessary) of the corresponding model. We will do so using MLE.

Step 1: Under the EDM model for the key ratios $\{Y_i\}_{i=1}^{n}$ as well as their independence, the log-likelihood l is given by

$$\begin{aligned}
l(\mathbf{y}; \boldsymbol{\theta}, \phi) &= \sum_{i=1}^{n} \ln f_{Y_i}(y_i, \theta_i, \phi) \\
&= \frac{1}{\phi} \sum_{i=1}^{n} w_i \left(y_i\theta_i - b(\theta_i)\right) + \sum_{i=1}^{n} c(y_i, \phi, w_i)
\end{aligned}$$

The maximization of the log-likelihood l with respect to $\boldsymbol{\theta} = (\theta_1, ..., \theta_n)$ does not involve ϕ.

Step 2: Use the chain rule and known relation between θ_i and $(\beta_0, \beta_1, ..., \beta_r)$ to write the log-likelihood l as a function of $(\beta_0, \beta_1, ..., \beta_r)$, i.e.

$$\theta_i = \left(b'\right)^{-1}(\mu_i),$$

and

$$g(\mu_i) = \eta_i = \sum_{j=0}^{r} x_{ij}\beta_j.$$

As such,

$$\frac{\partial}{\partial \beta_j} l(\mathbf{y}; \boldsymbol{\theta}, \phi) = \sum_{i=1}^{n} \left\{ \frac{\partial}{\partial \theta_i} l(\mathbf{y}; \boldsymbol{\theta}, \phi) \right\} \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_j},$$

where

$$\mu_i = b'(\theta_i) \Rightarrow \frac{\partial \mu_i}{\partial \theta_i} = b''(\theta_i) \Rightarrow \frac{\partial \theta_i}{\partial \mu_i} = \frac{1}{b''(\theta_i)},$$

and

$$g(\mu_i) = \eta_i \Rightarrow g'(\mu_i) \frac{\partial \mu_i}{\partial \eta_i} = 1 \Rightarrow \frac{\partial \mu_i}{\partial \eta_i} = \frac{1}{g'(\mu_i)}.$$

It follows that

$$\begin{aligned} \frac{\partial}{\partial \beta_j} l(\mathbf{y}; \boldsymbol{\theta}, \phi) &= \frac{1}{\phi} \sum_{i=1}^{n} w_i \left(y_i - b'(\theta_i) \right) \frac{1}{b''(\theta_i)} \frac{1}{g'(\mu_i)} x_{ij} \\ &= \frac{1}{\phi} \sum_{i=1}^{n} w_i \frac{y_i - \mu_i}{v(\mu_i) g'(\mu_i)} x_{ij}, \end{aligned} \tag{19}$$

for $j = 0, 1, ..., r$ where

$$\mu_i = g^{-1}\left(\sum_{j=0}^{r} x_{ij}\beta_j \right).$$

We obtain the MLE of $(\beta_0, \beta_1, ..., \beta_r)$, which we denote by $\left(\widehat{\beta}_0, \widehat{\beta}_1, ..., \widehat{\beta}_r\right)$, by setting all equations in (19) equal to 0. These $(r+1)$ equations are known as *ML equations* which are given by

$$\sum_{i=1}^{n} w_i \frac{y_i - \mu_i}{v(\mu_i) g'(\mu_i)} x_{ij} = 0. \tag{20}$$

In general, the ML equations must be solved numerically as no general closed-form solution exists. This will be done in R using the function GLM.

Remark 41 *For the satured model (i.e., if $r+1 = n$, the situation where the number of β's in the model is the same as the number of tariff cells), we can show that in this case the ML equations admit a closed-form expression where*

$$\widehat{\mu}_i = y_i,$$

for all $i = 1, 2, ..., n$. This is in general not interesting as the corresponding GLM overfits the data. However, we will see that this observation will be used to define the concept of deviance in the next sub-section to measure the overall fit of a particular GLM model.

Example 42 *Under the relative Poisson frequency model with a multiplicative tariff structure, we have $v(\mu) = \mu$ and $g(\mu) = \ln \mu$. Therefore, the $(r+1)$ ML equations become*

$$\sum_{i=1}^{n} w_i (y_i - \mu_i) x_{ij} = 0, \qquad j = 0, 1, ..., r,$$

where

$$\mu_i = \exp\{\eta_i\} = \exp\left\{\sum_{j=0}^{r} x_{ij}\beta_j\right\}, \qquad i = 1, 2, ..., n.$$

Example 43 *Under the gamma severity model with a multiplicative tariff structure, we have* $v(\mu) = \mu^2$ *and* $g(\mu) = \ln \mu$*. Therefore, the* $(r+1)$ *ML equations become*

$$\sum_{i=1}^{n} w_i \frac{(y_i - \mu_i)}{\mu_i} x_{ij} = 0, \qquad j = 0, 1, ..., r,$$

where

$$\mu_i = \exp\left\{\sum_{j=0}^{r} x_{ij}\beta_j\right\}, \qquad i = 1, 2, ..., n.$$

Example - moped insurance

Please refer to the *R* code.

2.7 GLM Model Building

Once the estimation of the GLM model is complete, the tariff analysis is far from done. A number of goals/objectives need to be considered to come up with a good and parsimonious tariff model. This may include:

- which rating factors (predictors) shall be included in the pricing model (frequency & severity)? Which ones are statistically meaningful?
- shall we group some categories within a rating factor or should we keep them all separate?
- how confident are we with the estimated betas/relativities, which are point estimates (i.e., are there enough observations in all cells)?
- how well the assumed model fits the data?

To answer the above questions (as well as many others), we call upon the comprehensive theory on GLMs to build confidence intervals for the betas/relativities, and perform general statistical (hypothesis) tests to assess the overall model fit and determine whether (or not) a given rating factor shall be included in the model.

2.7.1 Deviance

Let $l(\widehat{\mu}_1, ..., \widehat{\mu}_n; \boldsymbol{\theta}, \phi)$ to be the log-likelihood of the fitted GLM model (with a total of $(r+1)$ estimated β, namely $\widehat{\beta}_0, \widehat{\beta}_1, ..., \widehat{\beta}_r$) as a function of the estimated means. We pointed out earlier that for the saturated

model, the log-likelihood of the resulting model is $l(y_1, ..., y_n; \boldsymbol{\theta}, \phi)$. The concept of deviance makes use of the fact that the saturated model has a perfect fit to the data to use it as a benchmark to measure how well the fitted GLM model performs.

- **Scaled deviance:** likelihood ratio test with

  $$H_0: \text{ fitted GLM model}$$
  $$H_a: \text{ saturated model}$$

  leading to the statistic

  $$D^* := D^*(\mathbf{y}, \widehat{\boldsymbol{\mu}}) = 2\{l(y_1, ..., y_n; \boldsymbol{\theta}, \phi) - l(\widehat{\mu}_1, ..., \widehat{\mu}_n; \boldsymbol{\theta}, \phi)\}$$

  where the parameter ϕ is assumed to be the same for both models (usually estimated from the fitted model). We recall that

  $$l(\mathbf{y}; \boldsymbol{\theta}, \phi) = \frac{1}{\phi} \sum_{i=1}^{n} w_i (y_i \theta_i - b(\theta_i)) + \text{constant},$$

  where

  $$\theta_i = \left(b'\right)^{-1}(\mu_i),$$

  with $\mu_i = g^{-1}\left(\sum_{j=0}^{r} x_{ij}\beta_j\right)$. Hence,

  $$D^* = \frac{2}{\phi} \sum_{i=1}^{n} w_i \left\{ \left[y_i \left(b'\right)^{-1}(y_i) - b\left(\left(b'\right)^{-1}(y_i)\right)\right] - \left[y_i \left(b'\right)^{-1}(\widehat{\mu}_i) - b\left(\left(b'\right)^{-1}(\widehat{\mu}_i)\right)\right] \right\},$$

  with $\widehat{\mu}_i = g^{-1}\left(\sum_{j=0}^{r} x_{ij}\widehat{\beta}_j\right)$.

  The term *scaled* comes from the fact that D^* depends on ϕ (due to its presence in the denominator).

- **Unscaled deviance**:

  $$D = \phi D^*$$

  In this case, D does not depend on ϕ. (*Note:* This is by default the deviance statistic provided in R with the GLM function).

Under general conditions, D^* is approximately chi-squared distributed with $(n - r - 1)$ degrees of freedom. Large values of D^* imply that the simplified model may not be adequate (in comparison to the saturated model) while small values of D^* advocates in favor of the simplified model over the saturated one.

More precisely, if the likelihood ratio test is performed at a $(1 - \alpha)\,\%$ confidence level, we reject the null hypothesis (i.e., the fitted model) if $D^* > c_{1-\alpha}$ where $c_{1-\alpha}$ is the critical value of a chi-square rv χ^2_{n-r} with $(n - r - 1)$ degrees of freedom, i.e.

$$\mathbb{P}\left(\chi^2_{n-r-1} > c_{1-\alpha}\right) = \alpha.$$

Otherwise, the fitted model is accepted.

Example 44 *For the relative Poisson frequency model, we have $\theta_i = \ln \mu_i$, $b(\theta_i) = e^{\theta_i}$ and $\phi = 1$. Thus, $b'(\theta_i) = e^{\theta_i}$. Given that $\phi = 1$, there is no difference between the scaled and unscaled version of the deviance, i.e.*

$$D^* = D = 2\sum_{i=1}^{n} w_i \{[y_i \ln y_i - y_i] - [y_i \ln \widehat{\mu}_i - \widehat{\mu}_i]\}.$$

Example 45 *For the gamma severity model, we have* $b(\theta_i) = -\ln(-\theta_i)$ *which yields* $(b')^{-1}(\mu_i) = -\frac{1}{\mu_i}$. *As such, the unscaled deviance is*

$$D = 2\sum_{i=1}^{n} w_i \left\{ \left(\frac{y_i}{\widehat{\mu}_i} - 1 \right) - \ln\left(\frac{y_i}{\widehat{\mu}_i} \right) \right\}.$$

2.7.2 Pearson's chi-square test

For GLMs, the Pearson's goodness-of-fit test is generalized to

$$X^2 = \sum_{i=1}^{n} \frac{(y_i - \widehat{\mu}_i)^2}{Var(Y_i)} = \frac{1}{\phi} \sum_{i=1}^{n} w_i \frac{(y_i - \widehat{\mu}_i)^2}{v(\widehat{\mu}_i)},$$

which is a scaled statistic. As for the deviance statistic, the Pearson's goodness-of-fit test also has an unscaled statistic given by ϕX^2.

From statistical theory, we know that X^2 is approximately chi-squared distributed with $(n - r - 1)$ degrees of freedom (where $(r + 1)$ is the number of estimated β parameters). Large values of X^2 imply that the simplified model may not be adequate (in comparison to the saturated model) while small values of X^2 advocates in favor of the simplified model over the saturated one.

More precisely, if the Pearson's chi-square test is performed at a $(1 - \alpha)\,\%$ confidence level, we reject the null hypothesis (i.e., the fitted model) if $X^2 > c_{1-\alpha}$ where $c_{1-\alpha}$ is the critical value of a chi-square rv χ^2_{n-r-1} with $(n - r - 1)$ degrees of freedom, i.e.

$$\mathbb{P}\left(\chi^2_{n-r-1} > c_{1-\alpha}\right) = \alpha.$$

Otherwise, the fitted model is accepted.

Example 46 *For the relative Poisson frequency model,* X^2 *becomes*

$$X^2 = \sum_{i=1}^{n} w_i \frac{(y_i - \widehat{\mu}_i)^2}{\widehat{\mu}_i}.$$

When $w_i = 1$ *for all* i*, this reduces to the standard Pearson's goodness-of-fit test statistic*

$$X^2 = \sum_{i=1}^{n} \frac{(y_i - \widehat{\mu}_i)^2}{\widehat{\mu}_i}.$$

Example 47 *For the gamma severity model, we have*

$$X^2 = \sum_{i=1}^{n} \frac{(y_i - \widehat{\mu}_i)^2}{Var(Y_i)} = \frac{1}{\phi} \sum_{i=1}^{n} w_i \frac{(y_i - \widehat{\mu}_i)^2}{(\widehat{\mu}_i)^2} = \frac{1}{\phi} \sum_{i=1}^{n} w_i \left(\frac{y_i}{\widehat{\mu}_i} - 1 \right)^2.$$

2.7.3 Estimation of ϕ

There are several approaches to estimate ϕ, each with their pros and cons.

- **Method #1**: Moment matching technique with Pearson's test statistic: given that

$$\mathbb{E}\left[X^2\right] = n - r - 1,$$

we can set

$$X^2 = n - r - 1.$$

This leads to the following estimator for ϕ:

$$\widehat{\phi}_X = \frac{1}{n-r-1} \sum_{i=1}^{n} w_i \frac{(y_i - \widehat{\mu}_i)^2}{v\left(\widehat{\mu}_i\right)}$$

- **Method #2**: Moment matching technique with the scaled deviance: given that

$$\mathbb{E}\left[D^*\right] = n - r - 1,$$

we can set

$$D^* = n - r - 1.$$

This leads to

$$\widehat{\phi}_D = \frac{D}{n-r-1}.$$

- **Method #3**: Find the MLE of ϕ. (Practical consideration: find MLE for β's first as they don't depend on ϕ, then find MLE for ϕ by plugging MLE for β's).

Numerical studies have shown that Method 1 is the preferred estimation approach. The GLM function in R calculates the dispersion parameter ϕ using Method 1.

2.7.4 Testing hierarchical models

To determine whether or not to include a particular rating factor in a GLM, we may want to perform some statistical (hypothesis) tests. This can be done with the likelihood ratio test (LRT) as long as we compare two models for which one is a special case of the other. For instance, we may compare two models for which all else being equal, one model includes a particular rating factor while the other does not.

Mathematically speaking, for two models M_s and M_t such that $M_s \subset M_t$, we design the hypothesis test

H_0: data comes from M_s
H_a: data comes from the more complicated M_t.

Then, the LRT consists in computing the test statistic

$$D^*\left(\mathbf{y}, \widehat{\mu}^{(s)}\right) - D^*\left(\mathbf{y}, \widehat{\mu}^{(t)}\right) \geq 0,$$

where $\widehat{\mu}^{(s)}$ and $\widehat{\mu}^{(t)}$ are estimated means under models M_s and M_t, respectively. Note that the test requires M_s and M_t to be from the same EDM family (e.g., Poisson, gamma,...). Also, as both $D^*\left(\mathbf{y}, \widehat{\mu}^{(s)}\right)$ and $D^*\left(\mathbf{y}, \widehat{\mu}^{(t)}\right)$ depend on ϕ, we shall first estimate ϕ using one of the three above methods in the more complex model M_t.

From statistical theory, we know that $D^*\left(\mathbf{y}, \widehat{\mu}^{(s)}\right) - D^*\left(\mathbf{y}, \widehat{\mu}^{(t)}\right)$ approximately follows a chi-square distribution with degree of freedom equals to the difference of the number of fitted parameters in M_t and M_s.

Hence, the likelihood ratio test at a confidence level of $(1-\alpha)\,\%$ leads to the acceptance of H_0 if the p-value of the test is larger than α (otherwise, we reject H_0 and adopt H_a instead). The p-value of the test is given by

$$\mathbb{P}\left(\chi_d^2 > D^*\left(\mathbf{y}, \widehat{\mu}^{(s)}\right) - D^*\left(\mathbf{y}, \widehat{\mu}^{(t)}\right)\right),$$

where d is the difference of the number of fitted parameters in M_t and M_s.

2.7.5 Confidence intervals based on Fisher's information

In this section, the goal is to build confidence intervals for the relativities. This is important to quantify the level of confidence in the point estimate for the relativities, as well as make decisions as to include (or not) a particular rating factor or merge (or not) certain categories within a rating factor.

As a reminder, we recall that CIs based on MLE require the Fisher information matrix I, where

$$[I]_{jk} = -\mathbb{E}\left[\frac{\partial^2 l}{\partial\beta_j \partial\beta_k}\right], \qquad j,k = 0,1,...,r.$$

Plugging in the log-likelihood function, it can be shown that

$$I = X^T D X,$$

where X is our $n \times (r+1)$ design matrix (i.e., $[X]_{ij} = x_{ij}$), and D is a $n \times n$ diagonal matrix with i-th element

$$d_i = \frac{w_i}{\phi v\left(\mu_i\right)\left(g'\left(\mu_i\right)\right)^2}, \qquad i = 1,...,n.$$

We plug-in the estimates $\widehat{\mu}_i$ and $\widehat{\phi}$ for μ_i and ϕ (respectively) to get a numerical estimate of I.

Theoretically, we are justified to proceed in this fashion as asymptotically, we have $\widehat{\beta} \sim N\left(\beta, I^{-1}\right)$. As such, a confidence interval for β_j is given by

$$(a_j, b_j) = \widehat{\beta}_j \pm z_{1-\frac{\alpha}{2}}\sqrt{c_{jj}},$$

for $j = 0,1,...,r$ where c_{jj} is the j-th diagonal term of I^{-1} and $z_{1-\frac{\alpha}{2}}$ is the $1-\frac{\alpha}{2}$ quantile of a normal rv with mean 0 and variance 1. Note that this is approximately a $100\,(1-\alpha)\,\%$ confidence interval for β_j.

Remark 48 *Note that in R, confidence intervals for $\widehat{\beta}$ use the normal distribution when the parameter ϕ is not estimated. However, when ϕ is estimated, the normal distribution is replaced by a student $t-$distribution (with a degree of freedom equals to $n-r-1$) which has heavier tail than a normal distribution. As such, given that ϕ is estimated from the data, we penalize the significance of each predictor in the model by widening their confidence intervals.*

Hence, for the relativity $\gamma_j = e^{\beta_j}$,

$$\left(e^{a_j}, e^{b_j}\right)$$

is approximately a $100(1-\alpha)\,\%$ confidence interval for γ_j. We note that this CI will not be symmetric around $\widehat{\gamma}_j = e^{\widehat{\beta}_j}$.

Similarly, to build CIs for $\mu_i = e^{\eta_i}$ with $\eta_i = \sum_{j=0}^{r} x_{ij}\beta_j$, we know that

$$\widehat{\eta}_i = \sum_{j=0}^{r} x_{ij}\widehat{\beta}_j$$

is normally distributed with mean η_i and variance

$$(x_{i0}, x_{i1}, ..., x_{ir})\, I^{-1}\, (x_{i0}, x_{i1}, ..., x_{ir})^T .$$

Suppose (c_i, d_i) is the approximate CI for η_i, then

$$\left(e^{c_i}, e^{d_i}\right)$$

is approximately a $100\,(1-\alpha)\,\%$ confidence interval for μ_i.

2.7.6 Confidence intervals for pure premium relativities (skip)

When two separate GLM models are developed for the claim frequency and claim severity, we propose to estimate the pure premium relativity by simply multiplying the relativities from each model. But how do we construct CI? This is not a trivial task as the relativities are not independent. This is the case because, among other things, the weights in the severity model are the number of claims which is an (deterministic) input in the GLM frequency model.

Let γ^F and γ^S be the relativities for the frequency and severity model (respectively) for a given set of rating factors. Define

$$\gamma^P = \gamma^F \cdot \gamma^S.$$

Let

$$\beta^P = \ln \gamma^P = \ln \gamma^F + \ln \gamma^S = \beta^F + \beta^S.$$

We therefore estimate β^P by $\widehat{\beta}^P = \widehat{\beta}^F + \widehat{\beta}^S$. As for its variance,

$$\begin{aligned} Var\left(\widehat{\beta}^P\right) &= Var\left(\widehat{\beta}^F + \widehat{\beta}^S\right) \\ &= Var\left(\widehat{\beta}^F\right) + Var\left(\widehat{\beta}^S\right) + 2Cov\left(\widehat{\beta}^F + \widehat{\beta}^S\right). \end{aligned}$$

Given the complexity in calculating the covariance terms, we shall assume the covariance to be 0 and estimate $Var\left(\widehat{\beta}^P\right)$ by

$$\widehat{Var}\left(\widehat{\beta}^P\right) = \widehat{Var}\left(\widehat{\beta}^F\right) + \widehat{Var}\left(\widehat{\beta}^S\right).$$

2.8 Residuals (skip)

As for the ordinary regression, the analysis of residuals can be useful to check the GLM fit and its underlying assumptions. Two types of residuals will be considered:

1. **Pearson residuals:** For a given GLM model, we define the Pearson residuals as

$$r_{P_i} = \frac{y_i - \widehat{\mu}_i}{\sqrt{\frac{v(\widehat{\mu}_i)}{w_i}}}.$$

We note that

$$R_{P_i} := \frac{Y_i - \mu_i}{\sqrt{\frac{v(\mu_i)}{w_i}}}$$

has mean 0 and variance ϕ. Also, the GLM goodness-of-fit test X^2 can be written as

$$X^2 = \frac{1}{\phi}\sum_{i=1}^{n}\left(r_{P_i}\right)^2.$$

2. **Deviance residuals**: For a given GLM model, we define the deviance residuals as

$$r_{D_i} = \sqrt{w_i d\left(y_i, \widehat{\mu}_i\right)} sign\left(y_i - \widehat{\mu}_i\right),$$

where we recall

$$d\left(y_i, \widehat{\mu}_i\right) = \left[y_i h\left(y_i\right) - b\left(h\left(y_i\right)\right)\right] - \left[y_i h\left(\widehat{\mu}_i\right) - b\left(h\left(\widehat{\mu}_i\right)\right)\right].$$

Note that the scaled deviance can be expressed as

$$D^* = \frac{1}{\phi}\sum_{i=1}^{n}\left(r_{D_i}\right)^2.$$

A few remarks are now appropriate:

- given that both types of residuals have variance scaled by ϕ, their unscaled versions are

$$\frac{r_{P_i}}{\sqrt{\phi}}$$

and

$$\frac{r_{D_i}}{\sqrt{\phi}}$$

- additionally, given that the mean μ_i is estimated by $\widehat{\mu}_i$, an additional correction term can be applied

$$\frac{r_{P_i}}{\sqrt{\phi\left(1 - h_i\right)}}$$

and

$$\frac{r_{D_i}}{\sqrt{\phi\left(1 - h_i\right)}}$$

where h_i is the i-th element of the hat matrix

$$H = D^{\frac{1}{2}} X\left(X^T D X\right)^{-1} X D^{\frac{1}{2}}.$$

As a reminder, we recall that D is the diagonal matrix with i-th element

$$d_i = \frac{w_i}{\phi v\left(\mu_i\right)\left(g'\left(\mu_i\right)\right)^2}, \qquad i = 1, ..., n.$$

The deviance D^* and Pearson's goodness-of-fit test X^2 are used to assess the overall fit to the data. The analysis of residuals pushes this analysis a step further by possibly helping to find outliers or particular patterns in the fitted model (poor fitting in some cases). For the outliers, one can plot the unscaled residuals (which have 0 mean and variance of 1) against their index i, and check for outliers. As for patterns, it may be good to plot the residuals against $\widehat{\mu}_i$ to check the appropriateness of the variance function in the model. If it shows that the residuals tend to have more variability for large or small $\widehat{\mu}_i$, this may indicate that the variance function has been incorrectly specified.

2.9 Alternative models (skip)

2.9.1 Frequency models - overdispersion

In many insurance contexts, empirical studies have shown that claim counts within a tariff cell exhibit more variability than what is typically assumed in the traditional Poisson setup. We recall that in the (relative) Poisson model

$$Var\left(Y_i\right) := \frac{\phi v\left(\mu_i\right)}{w_i} = \frac{\mu_i}{w_i},$$

i.e., $\phi = 1$ and the variance function v is $v\left(\mu\right) = \mu$. In other words, even though risks have been classified as homogeneous within a rating cell, in practice a certain level of heterogeneity remains (for instance, due to the effect of explanatory variables of lesser importance not included in the model or random variations among the assumed homogeneous risks).

The goal is therefore to find ways to model overdispersion for claim frequency data. As before, the rv's $X_1^{w_1}, ..., X_n^{w_n}$ will be assumed to be independent. However, rather than directly modelling $X_i^{w_i}$ as a Poisson rv with mean $w_i\mu_i$, we consider a mixed Poisson model for $X_i^{w_i}$. More specifically, we define $X_i^{w_i}$ conditionally on a risk parameter Λ_i as follows:

- let $X_i^{w_i}\left|\Lambda_i = \lambda_i\right.$ be a Poisson rv with mean $w_i\mu_i\lambda_i$;
- the mixing rv Λ_i has a gamma distribution with mean 1 and variance $\frac{1}{v}$ (where $v > 0$).

We silently assume that the rv's $\Lambda_1, ..., \Lambda_n$ are independent (otherwise, the assumption that the X_i's are independent is violated).

Remark 49 *In the above setup, if $v \to \infty$, we have $X_i^{w_i}$ is a Poisson rv with mean $w_i\mu_i$. We therefore recover the (relative) Poisson model for Y_i.*

Under the above assumptions, it follows that

$$\begin{aligned}
\mathbb{E}\left[X_i^{w_i}\right] &= \mathbb{E}\left[\mathbb{E}\left[X_i^{wi}\left|\Lambda_i\right.\right]\right] \\
&= \mathbb{E}\left[w_i\mu_i\Lambda_i\right] \\
&= w_i\mu_i\mathbb{E}\left[\Lambda_i\right] \\
&= w_i\mu_i,
\end{aligned}$$

and

$$\begin{aligned} Var\left(X_i^{wi}\right) &= \mathbb{E}\left[Var\left(X_i^{w_i}\left|\Lambda_i\right.\right)\right] + Var\left(\mathbb{E}\left[X_i^{w_i}\left|\Lambda_i\right.\right]\right) \\ &= \mathbb{E}\left[w_i\mu_i\Lambda_i\right] + Var\left(w_i\mu_i\Lambda_i\right) \\ &= w_i\mu_i\mathbb{E}\left[\Lambda_i\right] + \left(w_i\mu_i\right)^2 Var\left(\Lambda_i\right) \\ &= w_i\mu_i + \left(w_i\mu_i\right)^2\frac{1}{v}. \end{aligned}$$

More generally, the pmf of $X_i^{w_i}$ is

$$\begin{aligned} \mathbb{P}\left(X_i^{w_i}=x\right) &= \int_0^\infty e^{-w_i\mu_i\lambda}\frac{\left(w_i\mu_i\lambda\right)^x}{x!}\frac{v^v\lambda^{v-1}e^{-v\lambda}}{\Gamma\left(v\right)}\mathrm{d}\lambda \\ &= \frac{\left(w_i\mu_i\right)^x}{x!}\frac{v^v}{\Gamma\left(v\right)}\int_0^\infty \lambda^{x+v-1}e^{-\left(w_i\mu_i+v\right)\lambda}\mathrm{d}\lambda \\ &= \frac{\left(w_i\mu_i\right)^x}{x!}\frac{v^v}{\Gamma\left(v\right)}\frac{\Gamma\left(x+v\right)}{\left(w_i\mu_i+v\right)^{x+v}} \\ &= \frac{\Gamma\left(x+v\right)}{x!\Gamma\left(v\right)}\left(\frac{v}{w_i\mu_i+v}\right)^v\left(\frac{w_i\mu_i}{w_i\mu_i+v}\right)^x, \end{aligned}$$

which is a negative binomial model. For a fixed $v>0$, we note that the pmf of $X_i^{w_i}$ can be rewritten as

$$\mathbb{P}\left(X_i^{w_i}=x\right) = \exp\left\{x\ln\left(\frac{w_i\mu_i}{w_i\mu_i+v}\right) + v\ln\frac{v}{w_i\mu_i+v} + \ln\left(\frac{\Gamma\left(x+v\right)}{x!\Gamma\left(v\right)}\right)\right\},$$

which is of the EDM form (15) with $\theta_i=\ln\left(\frac{w_i\mu_i}{w_i\mu_i+v}\right)$, $b\left(\theta_i\right)=-v\ln\left(1-\exp(\theta_i)\right)$, $\phi=1$ and $w_i=1$ for all i.

The build-in function GLM.NB under the library MASS in R can be used to run a negative binomial glm model for $X_i^{w_i}$. The parameter v is found through an iterative procedure.

2.10 Miscellanea (skip)

Some additional practical issues to consider in the selection of the tariff model.

1. **Model selection**: typically, there is a large number of explanatory variables that can be included in the tariff model. New technology has facilitated the collection and the availability of data. More explanatory variables are included in the model, better is the fit. However, our goal should be to find parsimonious models that have good enough fit (rather than overfitting the data as the resulting model may not be robust). Also, the inclusion of rating factors in the chosen tariff model should make sense from a practical standpoint.

2. **Interaction**: for the same reasons mentioned in (1), it is suggested to not include too many interaction terms, but in certain cases it may be justified. For instance, in car insurance, gender and age is a good example. This is done by segmenting the portfolio into young male drivers, young female drivers, older male drivers and older female drivers. Young male drivers have on average worst claim experience than the other three groups. As such, considering a multiplicative effect for age and another for gender may not be appropriate as young female drivers will be subjected to the multiplicative effect associated to their "young" age group label (which is not justified from the claim experience).

3 Classification methods

In this module, we consider methods for predicting a qualitative (categorical) response, a process known as *classification*. The classification process consists in classifying observations into one of a finite number of categories (or bins). We propose to use the information about a set of predictors X (which may either be continuous or categorical) to make this prediction about the qualitative response Y.

For instance, let the response Y be either 0 or 1 (where we typically refer to "0" as a negative response and "1" as a positive response). Define $\widehat{Y}$ to be the classification prediction for a given set of predictor(s) X. Then, the objective is to minimize the classification error

	$Y = 0$	$Y = 1$
$\widehat{Y} = 0$	correct decision	**Type II error** (false negative)
$\widehat{Y} = 1$	**Type I error** (false positive)	correct decision

In other words, a good classifier is one for which a vast majority of the observations are on the diagonal of this table.

Classification methods often predict the *probability* associated to each outcome of a categorical variable. These probabilities form the basis for making the classification. It is quite common to classify an observation into the category which generates the largest probability. We will look at a number of classification methods (also known as *classifiers*) in this module:

- logistic regression
- linear/quadratic discriminant analysis (LDA/QDA)
- K-nearest neighbors (KNN)

Before further diving into the theory of these classifiers, a brief (and certainly not exhaustive) list of classification problems is first presented:

- *Default risk:* determine if a credit card holder will default on payment (possible explanatory variables: credit card balance, student or not, income)
- *Car insurance claim:* determine if an insurance policy will generate a claim or not over the coverage period (possible explanatory variables: driver's age, zone, car type, value, claim experience)
- *Insurance marketing:* determine if a given individual will purchase or not an extended coverage protection (possible explanatory variables: salary, occupation)
- *Health insurance:* determine if a policyholder will develop or not a specific disease (possible explanatory variables: age, occupation, general health condition)
- *Fraud detection:* determine whether a submitted claim should be investigated or not (possible explanatory variables: claim size, type of claims)

But why not using a regression-type model (e.g., glm of Module 2) to classify a qualitative variable? The answer is simple: the lack of ordering (and its arbitrariness) in categorical variables makes it unsuitable in that case. Also, the prediction for a set of predictors X will usually not be an integer in which case it is not clear how to handle the classification exercise (even in the binary response case).

3.1 Logistic regression

Although this classifier can be used for categorical variables with more than two possible outcomes, we limit the presentation to binary responses in this course (i.e., $Y \in \{0,1\}$ without loss of generality).

We first consider the logistic model with $r = 1$ predictor. Let $p(x) = \mathbb{P}(Y = 1 \,|X = x)$ which we model using a logistic function, i.e.

$$p(x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}} \in [0,1].$$

Note that $p(x)$ is not linear in the predictor x, but rather has an S-shaped form.

Furthermore, we point out that

- odds ratio:
$$\frac{p(x)}{1-p(x)} = e^{\beta_0+\beta_1 x},$$
- log-odds or logit link function:
$$\ln \frac{p(x)}{1-p(x)} = \beta_0 + \beta_1 x.$$

We can thus interpret a change in the predictor x by one unit to result in a change in the log-odds by β_1 (or a change in the odds by a factor of e^{β_1}). Additionally, we note that

$$\frac{\mathrm{d}}{\mathrm{d}x} p(x) = \beta_1 p(x)(1-p(x)),$$

(see Q1 - Module 3 Recommended problems) for which the derivative of $p(x)$ has the same sign as β_1. As such,

- $\beta_1 < 0$ implies that $p(x)$ is decreasing in x;
- $\beta_1 > 0$ implies that $p(x)$ is increasing in x.

Remark 50 *The logistic regression model is a special case of the GLM setup discussed in Module 2 where the response Y is a Bernoulli distributed rv (binomial family in R) and the link function is a logit function. Indeed, let Y_i be a Bernoulli rv with mean p_i, its pmf is given by*

$$p_{Y_i}(y_i) = (p_i)^{y_i}(1-p_i)^{1-y_i}, \qquad y_i \in \{0,1\}.$$

Simple algebraic manipulations lead to

$$\begin{aligned} p_{Y_i}(y_i) &= (1-p_i)\left(\frac{p_i}{1-p_i}\right)^{y_i} \\ &= \exp\left\{y_i \ln\left(\frac{p_i}{1-p_i}\right) + \ln(1-p_i)\right\}, \end{aligned}$$

for $y_i \in \{0,1\}$ which is of the form (15), i.e.

$$f_{Y_i}(y_i, \theta_i, \phi) = \exp\left\{ \frac{y_i\theta_i - b(\theta_i)}{\frac{\phi}{w_i}} + c(y_i, \phi, w_i) \right\},$$

with $\theta_i = \ln\left(\frac{p_i}{1-p_i}\right)$, $b(\theta_i) = \ln\left(1+e^{\theta_i}\right)$, $\phi = 1$ and $w_i = 1$ for all i. The mean response $\mathbb{E}[Y_i] = p_i$ is related to $\eta_i = \beta_0 + \beta_1 x_{i1}$ through

$$\ln\left(\frac{p_i}{1-p_i}\right) = \eta_i,$$

i.e. the link function $g(p_i) = \ln\left(\frac{p_i}{1-p_i}\right)$ is the so-called logit *function.*

To estimate the parameters β_0 and β_1 of the logistic regression model, we propose to rely on the MLE theory developed in Section 2.6. In this case, the likelihood function is given by

$$\begin{aligned} L(\beta_0, \beta_1) &= \prod_{\text{all } i} p(x_i)^{y_i} (1 - p(x_i))^{1-y_i} \\ &= \left\{ \prod_{i:y_i=1} p(x_i) \right\} \left\{ \prod_{j:y_j=0} (1 - p(x_j)) \right\}, \end{aligned}$$

and we denote their maximum likelihood estimators by $\widehat{\beta}_0$ and $\widehat{\beta}_1$. In general, no explicit solution exists. Once again, one can use the glm function (with binomial family) in R to numerically find the MLE of β_0 and β_1. Note that in R, the logit function is the default link function for the binomial family.

Example 51 *Consider the Default dataset. The response variable is a* `default on payment` *indicator (yes/no) with the following possible predictors:* `balance`, `income` *and* `student`. *First, we consider a logistic regression model to predict default with only balance as a predictor. Thus, let $p(x) = \mathbb{P}(default = yes \,|\, balance = x)$. Using the R code*

```
def.fit=glm(default ~ balance, data=Default,family=binomial),
```

we obtain $\widehat{\beta}_0 = -10.6513$ and $\widehat{\beta}_1 = 0.0055$. If we test $H_0 : \beta_1 = 0$ vs $H_a : \beta_1 \neq 0$, the z-statistic $\widehat{\beta}_1 / SE(\widehat{\beta}_1)$ produces a value of 24.95 with a p-value of < 0.0001. We therefore reject H_0: $\beta_1 = 0$. To make a prediction, we estimate the probability of default by

$$\widehat{p}(x) = \frac{e^{\widehat{\beta}_0 + \widehat{\beta}_1 x}}{1 + e^{\widehat{\beta}_0 + \widehat{\beta}_1 x}},$$

which less than 1% for a balance of \$1000 and 58.6% for a balance of \$2000.

If we now replace the predictor `balance` *by* `student` *(where $X = 1$ if student and $X = 0$ otherwise), we get $\widehat{\beta}_0 = -3.5041$ and $\widehat{\beta}_1 = 0.4049$. The predictor* `student` *seems meaningful as the p-value for the hypothesis test $H_0 : \beta_1 = 0$ vs $H_a : \beta_1 \neq 0$ is < 0.0004. It is interesting to note that the ML estimators can be found explicitly in this simple case. (See Q2 in the Module 3 - Recommended problems).*

We now turn our attention to the logistic regression with a finite positive number r of predictors. The idea remains the same. Let

$$p(\mathbf{x}) = \mathbb{P}(Y = 1 \,|\, X_1 = x_1, ..., X_r = x_r),$$

which we model using a logistic function, i.e.

$$p(\mathbf{x}) = \frac{e^{\beta_0+\sum_{j=1}^r \beta_j x_j}}{1+e^{\beta_0+\sum_{j=1}^r \beta_j x_j}} \in [0,1].$$

It follows that

- odds ratio:
$$\frac{p(\mathbf{x})}{1-p(\mathbf{x})} = e^{\beta_0+\sum_{j=1}^r \beta_j x_j},$$
- log-odds or logit link function:
$$\ln \frac{p(\mathbf{x})}{1-p(\mathbf{x})} = \beta_0 + \sum_{j=1}^r \beta_j x_j.$$

Once again, we shall interpret a change in the predictor x_j (all else being equal) by one unit to result in a change in the log-odds by β_j.

Example 52 *For the Default dataset, a logistic regression model is used to predict the default on payment indicator (yes/no) using all three predictors:* `balance`, `income` *and* `student`. *The following R code*

```
def.fit=glm(default ~ balance+income+student, data=Default,family=binomial)
```

yields the following estimates:

$$1.005200 in 4.559400 in$$

We highlight the following observations based on the above results:

- *The results strongly indicate that the predictor* `income` *should be dropped out of the model. Indeed, the hypothesis test* $H_0 : \beta_2 = 0$ *vs* $H_a : \beta_2 \neq 0$ *has a* z *value of* 0.3698 *with a* p*-value of* 0.7115;
- *The predictor* `student` *is statistically significant to predict default on payment with* $\widehat{\beta}_3 < 0$. *This implies that all else being equal, a student has a smaller probability of default than a non-student. This seems in contradiction with an earlier observation that students are more likely to default than non-students. This phenomenom can be explained by the interaction between the predictors* `balance` *and* `student`. *Indeed, students tend to carry higher balances (than non-students) and we know from before that a high balance increases the likelihood of default. From the logistic model, we know that for the same balance, a student is less likely to default than a non-student.*

We can assess the accuracy/performance of the logistic regression classifier using different metrics. We represent these metrics in the context of a binary response Y through the so-called **confusion matrix**:

Predicted\True	$Y=0$	$Y=1$
$\widehat{Y}=0$	k_{00}	k_{01}
$\widehat{Y}=1$	k_{10}	k_{11}

For convenience, let $k = k_{00}+k_{10}+k_{01}+k_{11}$. Note that $k_{ij} = \sum_{l=1}^n 1_{\{\widehat{Y}_l=i, Y_l=j\}}$.

- **error rate**: % of incorrectly classified observations

$$\frac{k_{01} + k_{10}}{k}$$

- **sensitivity rate**: % of "$Y = 1$" correctly identified

$$\frac{k_{11}}{k_{01} + k_{11}}$$

  $\Rightarrow$ *Remark:* **sensitivity rate = 1 − Type II error**

- **specificity rate**: % of "$Y = 0$" correctly identified

$$\frac{k_{00}}{k_{00} + k_{10}}$$

  $\Rightarrow$ *Remark:* **specificity rate = 1 − Type I error**

Example 53 *For the* `default` *data, we use the logistic regression classifier to predict default on payment using the predictors* `balance` *and* `student`*. We obtain the following confusion matrix:*

Predicted\ True	*No*	*Yes*
No	9628	228
Yes	39	105

The error rate is $\frac{228+39}{10000} = 2.67\%$ *with a sensitivity ratio of* $\frac{105}{333} = 31.5\%$ *and a specificity ratio of* $\frac{9628}{9667} = 99.6\%$*. The logistic regression classifier has a low error rate. However, it should be noted that the trivial classifier which would assign all observations to the No default category has confusion matrix*

Predicted\ True	*No*	*Yes*
No	9667	333
Yes	0	0

resulting in an error rate of 3.33%*, only a bit higher than the logistic regression error rate. By further digging into the other performance metrics, we can identify situations where the logistic regression classifier performs well (and not so well). Indeed, the logistic regression sensitivity rate sits at* 31.5% *which is relatively low (i.e., the logistic regression classifier seems to do a poor job to identify defaulting customers) while the specificity rate is* 99.6% *(i.e., the logistic regression classifier identifies the non-defaulting customers quite well). The latter explains why the error rate is quite low as most customers in the data set do not default and the specificity rate is high.*

3.2 Linear Discriminant Analysis (LDA)

The linear discriminant analysis can be viewed as an alternative classifier to the logistic regression.

- **logistic regression**: we model the conditional distribution of the response Y as a function of the predictor(s) X (via a logistic function);

- **linear discriminant analysis:** we rather model the predictor(s) X separately in each of the response classes (i.e., all outcomes of Y) and use Bayes theorem to find the resulting model for $Y\,|X$.

We will now explain the LDA classifier in more detail.

Remark 54 *The LDA approach is known to be generally more stable than logistic regression approach (especially in cases where the response classes of Y are well separated). In addition, the LDA model tends to be more popularly used when the response Y is not binary (i.e., Y has more than 2 possible outcomes).*

For illustrative purposes, let $Y \in \{1, 2, ..., K\}$ for K a finite and positive integer. Define

- $\pi_k = \mathbb{P}(Y = k)$ to be the *prior* probability that a randomly chosen observation comes from the k-th class $(k = 1, 2, ..., K)$;
- $f_k(x)$ to be the density (probability mass function) of the continuous (discrete) predictor X at x given the observation comes from the k-th class (i.e., $Y = k$).

From Bayes theorem, it follows that

$$p_k(x) := \mathbb{P}(Y = k\,|X = x) = \frac{\pi_k f_k(x)}{\sum_{j=1}^{K} \pi_j f_j(x)}, \qquad k \in \{1, 2, ..., K\},$$

which is the *posterior* probability that the response variable Y is k given $X = x$. Hence, when $X = x$, the LDA method classifies an observation in the category that maximizes $p_k(x)$ across all categories k.

As a first illustration, consider a LDA model with $r = 1$ predictor. We assume that the distribution of $X\,|Y = k$ is normal with mean μ_k and variance σ^2. (**Note**: In R, the `lda` function is coded so that all predictors are normally distributed). It follows that the posterior probability is given by

$$p_k(x) = \frac{\pi_k\, e^{-\frac{(x-\mu_k)^2}{2\sigma^2}}}{\sum_{j=1}^{K} \pi_j\, e^{-\frac{(x-\mu_j)^2}{2\sigma^2}}}, \qquad k \in \{1, ..., K\}.$$

The LDA classifier assigns an observation to the category k that maximizes $p_k(x)$. As all p_k's have the same denominator, the maximization exercise reduces to maximizing their numerators $\pi_k\, e^{-\frac{(x-\mu_k)^2}{2\sigma^2}}$ or equivalently

$$\begin{aligned}\ln\left(\pi_k\, e^{-\frac{(x-\mu_k)^2}{2\sigma^2}}\right) &= \ln \pi_k - \frac{(x-\mu_k)^2}{2\sigma^2} \\ &= \ln \pi_k - \frac{x^2 - 2x\mu_k + (\mu_k)^2}{2\sigma^2}.\end{aligned}$$

Given that the x^2 term is constant over all k, we shall maximize (in k for a given x)

$$\delta_k(x) := \ln \pi_k + \frac{2x\mu_k - (\mu_k)^2}{2\sigma^2},$$

which is a linear function in x.

Remark 55 *When $K = 2$, an observation with predictor $X = x$ is assigned to $\widehat{Y} = 1$ if*

$$\delta_1(x) - \delta_2(x) = \ln \frac{\pi_1}{\pi_2} + \left(x - \frac{\mu_1 + \mu_2}{2} \right) \frac{(\mu_1 - \mu_2)}{\sigma^2} > 0$$

or $\widehat{Y} = 2$ otherwise.

To use the LDA model, we shall estimate the parameters μ_k $(k = 1, ...K)$, π_k $(k = 1, ...K)$ and σ^2. Naturally, we can use the following estimates:

$$\begin{aligned}
\widehat{\mu}_k &= \frac{1}{n_k} \sum_{i:y_i = k} x_i \\
\widehat{\pi}_k &= \frac{n_k}{n} \\
\sigma^2 &= \frac{1}{n - K} \sum_{k=1}^{K} \sum_{i:y_i = k} (x_i - \widehat{\mu}_k)^2 \\
&= \sum_{k=1}^{K} \omega_k \widehat{\sigma}_k^2 \qquad (\text{with } \omega_k = \frac{n_k - 1}{n - K})
\end{aligned}$$

where $n_k = \#$ of observations of Y in the k-th category (with $n = \sum_{k=1}^{K} n_k$) and

$$\widehat{\sigma}_k^2 = \frac{\sum_{i:y_i = k} (x_i - \widehat{\mu}_k)^2}{n_k - 1}.$$

This is the default estimation procedure being carried out in the `lda` function in R.

In summary, the *LDA classifier* states that when $X = x$, we assign an observation to the k-th category that maximizes

$$\widehat{\delta}_k(x) = \ln \widehat{\pi}_k - \frac{2x\widehat{\mu}_k - (\widehat{\mu}_k)^2}{2\widehat{\sigma}^2}.$$

Now, for a LDA model with more than 1 predictor, the idea follows along the same lines. For illustrative purposes, we shall assume that $X = (X_1, ..., X_r)$ is multivariate Gaussian (i.e., multivariate normal) with

$$f_k(x) = \frac{1}{(2\pi)^{\frac{r}{2}} \det(\Sigma)^{\frac{1}{2}}} e^{-\frac{1}{2}(x - \mu_k)^T \Sigma^{-1} (x - \mu_k)}.$$

It follows that the posterior probability is given by

$$p_k(x) = \frac{\pi_k \ e^{-\frac{1}{2}(x - \mu_k)^T \Sigma^{-1} (x - \mu_k)}}{\sum_{j=1}^{K} \pi_j \ e^{-\frac{1}{2}(x - \mu_j)^T \Sigma^{-1} (x - \mu_j)}}, \qquad k \in \{1, ..., K\}.$$

If we follow Bayes classifier rule, we shall choose k that maximizes the numerator $\pi_k \ e^{-\frac{1}{2}(x - \mu_k)^T \Sigma^{-1} (x - \mu_k)}$ which is equivalent to maximize

$$\ln \pi_k - \frac{1}{2} (x - \mu_k)^T \Sigma^{-1} (x - \mu_k)$$

Leaving out the constant term, this is equivalent to maximize

$$\delta_x(x) = \ln \pi_k + x^T \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k,$$

which is once again linear in x.

Remark 56 *When $K = 2$, an observation with predictors $X = x$ is assigned to $\widehat{Y} = 1$ if*

$$\delta_1(x) - \delta_2(x) > 0$$

or $\widehat{Y} = 2$ otherwise. Note that

$$\delta_1(x) - \delta_2(x) = \ln\frac{\pi_1}{\pi_2} + x^T\Sigma^{-1}(\mu_1 - \mu_2) - \frac{\mu_1 + \mu_2}{2}\Sigma^{-1}(\mu_1 - \mu_2),$$

which is of the form

$$\delta_1(x) - \delta_2(x) = \alpha + x^T\beta,$$

where

$$\alpha = \ln\frac{\pi_1}{\pi_2} - \frac{\mu_1 + \mu_2}{2}\Sigma^{-1}(\mu_1 - \mu_2),$$

and

$$\beta = \Sigma^{-1}(\mu_1 - \mu_2).$$

We point out that for the function lda in R, the coefficients of linear discriminants (scaling) are not β but rather

$$\frac{\beta}{\sqrt{(\mu_1 - \mu_2)^T\Sigma^{-1}(\mu_1 - \mu_2)}}.$$

To use the LDA model, we shall estimate the unknown parameters $(\mu_1, ..., \mu_K)$, $(\pi_1, ..., \pi_K)$ and Σ using a similar approach as the one presented earlier.

Example 57 *For the* `Default` *data, we use the LDA classifier to predict* `default on payment` *using the predictors* `balance` *and* `student`*. We obtain the following confusion matrix:*

Predicted\ True	*No*	*Yes*
No	9644	252
Yes	23	81

The LDA error rate is low at $\frac{252+23}{10000} = 2.75\%$ with a sensitivity rate of $\frac{81}{252+81} = 24.3\%$ and a specificity rate of $\frac{9644}{9667} = 99.8\%$. Overall, the results are very much in line with those of the logistic regression classifier. The only notable difference is that the logistic regression seems to perform better to identify defaulting customers in comparison to the LDA approach (as the logistic regression sensitivity rate is 31.5% vs 24.3% for the LDA classifier).

Remark 58 *For the* `Default` *data, the financial consequence of mis-classifying defaulting customers ($Y = 1$) may be far greater than mis-classifying non-defaulting customers ($Y = 0$). As such, the company may want to increase the sensitivity ratio (at the detriment of potentially increasing the error rate and/or lowering the specificity ratio). In this binary response example, both the LDA classifier and the logistic regression classifier assigns an observation to the default category whenever*

$$\mathbb{P}(Y = 1 \,|X = x) > 0.5.$$

The 0.5 threshold could be reduced to make it easier to classify observations into the default category. Once again, this comes at the expense of increasing the error test and/or decreasing the sensitivity ratio. The following is the confusion matrix for a LDA model with a threshold of 0.2:

Predicted\ True	*No*	*Yes*
No	9432	138
Yes	235	195

which has an error rate of 3.73% with a sensitivity ratio of $\frac{195}{333} = 58.6\%$ *and a specificity ratio of* $\frac{9432}{9667} = 97.6\%$.

The ROC curve is a visual representation of the performance of a classifier. The ROC curve is a plot of the true positive rate (i.e., sensitivity ratio) against false positive rate (i.e., $1-$ specificity ratio) for different classifying thresholds. The ROC curve goes from $(0, 0)$ for a classifying threshold of 1 to $(1, 1)$ for a threshold of 0. The overall performance of a classifier is quantified by the area under the ROC curve which would ideally be close to 1. The larger is the area under the ROC curve, the better is the classifier. This can be justified by our goal to maximize the sensitivity ratio (true positive) while minimizing 1 - specificity ratio (false positive).

Example 59 *For the Default data, a threshold of 0.5 yields the point* $(0.002, 0.243)$ *on the ROC curve while the threshold of 0.2 generates the point* $(0.024, 0.586)$ *on the ROC curve. The entire ROC curve can be obtained by varying the threshold level of the LDA model. As we will se,, the built-in function* `roc` *in R is very helpful to obtain this curve. The area under the ROC curve is 0.95 (which is close to 1) which implies that the LDA classifier is performing well.*

3.3 Quadratic Discriminant Analysis (QDA)

The QDA classifier is an extension of the LDA classifier. Recall that for the LDA classifier, the conditional distribution of $X\,|Y = k$ is assumed to be normal with a common covariance matrix Σ for all k. For the QDA classifier, we rather assume that this covariance matrix can now depend on the response type, i.e. $X\,|Y = k$ has a multivariate normal distribution with mean μ_k and covariance matrix Σ_k. It is not difficult to show that the discriminant function becomes

$$\begin{aligned}\delta_k(x) &= -\frac{1}{2}(x-\mu_k)^T \Sigma_k^{-1}(x-\mu_k) - \frac{1}{2}\ln|\Sigma_k| + \ln\pi_k \\ &= -\frac{1}{2}x^T\Sigma_k^{-1}x + x^T\Sigma_k^{-1}\mu_k - \frac{1}{2}\mu_k^T\Sigma_k^{-1}\mu_k - \frac{1}{2}\ln|\Sigma_k| + \ln\pi_k,\end{aligned}$$

which we shall assign to the category with the largest $\delta_k(x)$. Note that $\delta_k(x)$ has a quadratic term in x (which explains the reference to quadratic discriminant analysis).

The QDA classifier is obtained by plugging estimates to Σ_k, μ_k and π_k into $\delta_k(x)$ and assigning an observation $X = x$ to the class for which this quantity is the largest. The main differences between LDA and QDA are listed below:

- QDA requires an additional $(K-1)\frac{p(p+1)}{2}$ parameters to be estimated (as the covariance matrix Σ_k differs for each k)
- LDA is much less flexible than QDA
- LDA classifier has substantially lower variance but high bias (especially if the LDA assumptions are significantly violated)
- Rule of thumb: use LDA if n is small (reducing variance is crucial), use QDA for very large n or if the assumption that $\Sigma_k = \Sigma$ for all k is significantly violated.

Example 60 *Consider the Default data and perform the QDA to predict default using the predictors* `balance` *and* `student`*. We obtain the following confusion matrix:*

Predicted\True	*No*	*Yes*
No	9637	244
Yes	30	89

The QDA error rate is $\frac{244+30}{10000} = 2.74\%$ *which is essentially the same as the LDA error rate. However, the sensitivity ratio is a bit higher for QDA at* $\frac{89}{333} = 26.7\%$ *(vs* 24.3% *for LDA) while the QDA sensitivity ratio is* $\frac{9637}{9667} = 99.7\%$ *(vs 99.8% for the LDA sensitivity ratio).*

3.4 K-nearest neighbors (KNN)

The KNN model is a non-parametric classifier that defines the conditional probability of the response Y given a set of predictor(s) $X = x_0$ by the empirical distribution obtained from the K nearest neighbors to the predictor set $X = x_0$. We will use the Euclidean norm to measure the distance between a point, say x, and the predictor set x_0. For a vector $x_0 = (x_{01}, ..., x_{0r})$ and a point $x = (x_1, ..., x_r)$, the Euclidean norm is defined as

$$d(x, x_0) = \|x - x_0\| = \sqrt{(x_1 - x_{01})^2 + (x_2 - x_{02})^2 + ... + (x_r - x_{0r})^2}.$$

Let $\mathcal{N}_0$ be the set of the K nearest neighbors of x_0 from the training data (i.e., choose the K points from the training set with the smallest Euclidean distance d to the predictor set x_0). We therefore approximate

$$\mathbb{P}(Y = j \,|X = x_0)$$

by

$$\widehat{\mathbb{P}}(Y = j \,|X = x_0) = \frac{1}{K} \sum_{i \in \mathcal{N}_0} 1_{\{y_i = j\}}. \tag{21}$$

The KNN classifier assigns an observation to the class with the largest probability in (21).

Remark 61 *With the KNN approach, it is necessary to split the set of observations into a training set (used to find the closest neighbors) and a test set (over which predictions will be made). Indeed, if an observation with predictor* x_0 *for which we want to predict the outcome* y_0 *can be used to make the prediction, the KNN method has an unfair advantage. As an extreme case, if* $K = 1$ *the predicted and true outcome would always match.*

A few additional points to make on the KNN classifier:

- no assumption is made on the shape of the decision boundary (can be expected to better perform than LDA and QDA when decision boundary is highly non-linear);
- a clear disadvantage of the KNN classifier is that it provides no information on which covariates are important to predict the response;

Practical considerations:

- the choice of K greatly affects the performance of the KNN classifier. We usually choose a relatively small odd number ($K = 3, 5, 7$) and see how the classifier performs with the test set;
- the treatment of categorical predictors to assess the "proximity" criterion (i.e., Euclidean distance) is somewhat subjective. One can turn a categorical predictor with d categories into $(d-1)$ dummy variables. Otherwise, if the categories of the predictor have some meaning scale-wise, we can simply use the categorical predictor as is. This should be specified and may be tailored to the problem at hand;
- It is important to scale independent variables so that the contribution to the distance from each predictor is somewhat of an equal importance and is not significantly impacted by the unit of measurement (with the help of the function `scale` in R); we will apply this rule to all predictor variables in the model;
- one can use the `sample` function in R to randomly select a sample of $n_0 < n$ observations to define the training set of data (i.e., sample(1:n, n_0)).

4 Tree-based methods

Decision tree-based methods can be applied to both regression- (Module 2) and classification-type (Module 3) problems. These methods involve segmenting the predictor space (the multi-dimensional space of the predictor set X) into a number of simple regions. To make a prediction, we typically use the mean (regression) or mode (classification) of all training observations in the region to which the observation belongs. These methods are simple and useful for interpretation, but often lacks in predictive accuracy. To remedy to this last point, we later introduce the concept of *bagging*, *random forests* and *boosting* in the context of decision-tree models.

Why using the terminology "tree" to refer to these prediction methods? This is because of the way the predictor space is stratified into regions which is done according to a set of splitting rules that can be summarized into a tree.

Example 62 *In this module, we consider the data set* `usautoBI` *in the* `CASdatasets` *package in R. The response variable is* `LOSS` *which measures the claimant's total economic loss (in thousand USD); the predictor set contains the following variables:*

- `ATTORNEY`: whether the claimant is represented by an attorney: 1 is yes; 0 is no;
- `CLMSEX`: Claimant's gender: M for male and F for female;
- `MARITAL`: Claimant's marital status : 1 if married, 2 if single, 3 if widowed, and 4 if divorced/separated;
- `CLMINSUR`: Whether or not the driver of the claimant's vehicle was uninsured: 1 if yes, 2 if no, and 3 if not applicable;
- `SEATBELT`: Whether or not the claimant was wearing a seatbelt/child restraint: 1 if yes, 2 if no, and 3 if not applicable;
- `CLMAGE`: Claimant's age.

A possible tree can be of the form:

$$5.694800 in 12.165300 in$$

A tree should be read from top to bottom. As such, the first split in the tree results from the predictor `ATTORNEY`*. Claimants not represented by an attorney (i.e.,* `ATTORNEY`*=0) go to the left of the tree, while claimants represented by an attorney (i.e.,* `ATTORNEY`*=1) go to the right. For claimants represented by an attorney, a second split divides the group into those wearing a seatbelt at the time of the accident (left) or not (right). The other splits in the tree can be interpreted similarly.*

Here are some useful terminologies related to a tree-based model:

- **internal node**: node that has child nodes
- **terminal nodes or leaves**: node with no child (i.e., end points of a tree)
- **branches**: segments of the tree that connect the nodes

- size of tree = number of terminal nodes

Example 63 *For the above tree, we have a total of 5 leaves (or terminal nodes), 4 internal nodes, and 8 branches.*

Steps to build a tree: roughly speaking, it is a two-step process:

1. Divide the predictor space (i.e., the space of all possible values for X) into J distinct and non-overlapping regions $R_1, R_2, ..., R_J$
2. For every observation that falls into a given region R_j, we make the (same) prediction given by the mean of the response values over the training observations in R_j (for classification problems, the mean prediction is replaced by the majority rule, i.e. the most represented response values over the training observations in R_j).

The delicate task lies in Step 1 which requires to partition the predictor space X. We first present possible ways to proceed in the context of a quantitative response, and later consider adjustments to be made if the response variable is categorical.

4.1 Regression tree

In general, a tree is constructed by dividing the predictor space X into high-dimensional rectangles or boxes (which is the convention applied in R via the use of the function `tree` in the library `tree`) to make the task more manageable:

- quantitative predictors are divided into a criterion of the form $X_i \leq s$ or $X_i > s$ for some i at each step;
- categorical predictors are divided into two non-empty sets of categories.

The ultimate goal consists in finding regions $R_1, R_2, ..., R_J$ that minimizes the residual sum of squares

$$RSS = \sum_{j=1}^{J} \sum_{i \in R_j} \left(y_i - \widehat{y}_{R_j} \right)^2,$$

where $\widehat{y}_{R_j}$ is the mean response over R_j. However, it is computationally infeasible to minimize RSS over all possible partitions of X into J boxes. A possible remedy is to use the so-called *recursive binary splitting*.

Recursive binary splitting

The recursive binary splitting is a top-down and greedy approach that consists in introducing the best possible split each step of the way starting from the top of the tree. For the first split, start from the entire predictor space X and split it according to whether $X_j < s$ or $X_j \geq s$ (where s is a threshold value). Let

$$R_1(j, s) = \{X \,|X_j < s\},$$

and

$$R_2(j,s) = \{X \,|X_j \geq s\}.$$

Our objective is to seek the pair (j,s) such that

$$\sum_{i:x_i \in R_1(j,s)} \left(y_i - \widehat{y}_{R_1(j,s)}\right)^2 + \sum_{i:x_i \in R_2(j,s)} \left(y_i - \widehat{y}_{R_2(j,s)}\right)^2$$

is minimized. At the next iteration, we repeat this process by breaking either region R_1 or R_2 into two regions. This iterative process is repeated until either (1) the desired number of regions is reached; (2) there are fewer than a given number of observations in each region; (3) the reduction in RSS in introducing another region does not exceed a certain threshold.

A few remarks:

- The function `tree` in the R library `tree` applies this procedure;
- Variables higher up in the tree are deemed "more important" than variables further down the tree to predict the response variable.

Example 64 *For the data set* `usautoBI` *in the* `CASdatasets` *in R, the above tree was obtained by recursive binary splitting via the R function* `tree`*. Given that* `ATTORNEY` *is the first predictor to make its way into the model, this is the most significant variable to predict the response* `LOSS`*. This is followed by the predictor* `SEATBELT`*.*

Note that the recursive binary splitting is likely to overfit the data (especially, if the resulting tree has a lot of leaves in relation to the number of points in the data set). The test MSE may not be as good as the training MSE, as the resulting tree is too complex and overfits the training data set. We may want to reduce the complexity of the tree by sacrifying a little bit of bias to reduce the variance and gain in interpretation.

Tree pruning

To overcome the overfit issue, a better strategy may be to grow a big tree and prune it back (e.g., cut it down in size) in order to obtain a subtree.

A first simple pruning technique consists in starting from a big tree (say T_0) and prune it back by choosing the best tree of a specified size. By reducing the tree size, you can find out how the big tree was iteratively constructed. This can be done in R by specifying the argument `best` in the function `prune.tree`.

Equivalently, one can penalize the tree in relation to its size by introducing a tuning parameter $\alpha \geq 0$ and finding the subtree $T \subset T_0$ (where T_0 is the initial big tree) which minimizes

$$\left\{ \sum_{m=1}^{|T|} \sum_{i:x_i \in R_m} \left(y_i - \widehat{y}_{R_m}\right)^2 \right\} + \alpha |T|. \tag{22}$$

In (22), the penalty term $\alpha |T|$ is added to the RSS to penalize trees for their complexity (where $|T|$ stands for the tree size). Larger is the tuning parameter α, heavier is the penalty assigned to trees with a large number of leaves. As such, the chosen tree size is a non-increasing function of α. This can be done in R by specifying the argument `k` (which stands for the argument α in (22)) in the function `prune.tree`.

But how do we optimally select the best subtree of the big tree T_0 (i.e., how to optimally choose the argument `best` or `k` in `prune.tree`)? It may be ideal to choose the subtree with the smallest test MSE, which is estimated using cross-validation. This method is known as the *cost complexity pruning* or *weakest link pruning*.

Cost complexity pruning

The cost complexity pruning is a cross-validation tree procedure which can be summarized as follows:

- consider a sequence of non-negative tuning parameters $\{\alpha_l\}_{l=0}^{L}$ with $\alpha_0 = 0$ and $\alpha_l < \alpha_j$ for $l < j$
- for a given α_l,
  - using the entire data set, find the subtree $T \subset T_0$ (where T_0 is the initial big tree) such that

$$\left\{ \sum_{m=1}^{|T|} \sum_{i:x_i \in R_m} (y_i - \widehat{y}_{R_m})^2 \right\} + \alpha_l |T| \tag{23}$$

    is minimized. Recall that $|T|$ represents the size of the tree T and $R_1, ..., R_{|T|}$ are regions corresponding to the terminal nodes.
  - denote the optimal tree by T^{α_l}
- For $\alpha_0 = 0$, T_0 minimizes (23)
- As α increases, we penalize more and more for the size of the tree so the optimal tree should decrease in size, i.e.

$$|T^{\alpha_0}| \geq |T^{\alpha_1}| \geq ... \geq |T^{\alpha_L}|$$

- To choose the final tree, use $K-$fold cross-validation to choose the optimal α_l (and correspondingly the optimal tree T^{α_l}). That is, we divide the training data into K subsets. For each $k = 1, ..., K$, keep the structure of the tree T^{α} intact, re-evaluate the predictions in each leave when the $k-$th subset of the training data is taken out and calculate the mean squared prediction error on the left-out $k-$th subset. For a given $\alpha \in \{\alpha_l\}_{l=0}^{L}$, average out the results of the mean square prediction errors and pick the α's which provides the minimal value (which is denoted by α^*). The selected tree is T^{α^*}.

Once again, we can count on the built-in function `cv.tree` in the R library `tree` to perform this task. This function has the added benefit of looking at all the subtrees of the big tree T_0 by looking at a wide array of tuning parameters $\{\alpha_l\}_{l=0}^{L}$ (which we won't need to specify/input when we call the function `cv.tree`). See the Module 4 - R code for more details.

4.2 Classification tree

To build a tree for classification purposes, the approach remains very much the same. The only difference lies in the fact that the response variable is categorical rather than quantitative. For the regression tree, an observation in a given region was assigned the mean value of all training responses in the region. For a categorical response, we shall assign an observation to the category which is the most heavily represented in the region.

One can therefore use the recursive binary splitting, but replace the decision criterion based on RSS by the classification error rate which, for each leave, corresponds to the fraction of observations not in the most represented class. For region (leave) R_m, the classification error is given by

$$E_m = 1 - \max_k \widehat{p}_{m,k},$$

where $\widehat{p}_{m,k}$ is the proportion of the training observations in region R_m that are from category k.

Note that it is possible for a classification tree to have 2 terminal nodes (with the same parent node) predicting the same response. Why splitting in this case? Because it can increase the so-called *node purity*. If an observation falls into one node (rather than the other), a purer node will imply more confidence in the prediction.

We now list the main advantages and disadvantages of (regression or classification) trees to end this subsection. Decision trees

- can handle non-linear relation between the response and the predictors;
- are very easy to explain to a non-technical audience (i.e., visual display, logic behind the model);
- some people think they better mimic human decision-making;
- can be displayed graphically and are easy to interpret;
- can handle categorical predictors without the need for dummy variables (simpler for a non-technical user);
- don't have the same level of prediction accuracy as other methods;
- often lacks robustness (i.e., small changes in training data lead to big changes in the tree structure).

4.3 Bagging, Random Forest and Boosting

These methods can be used to address some of the issues alluded above (in particular, the last two bullet points).

4.3.1 Bagging

This method is also known as bootstrap aggregation. The idea is to reduce the variance of the decision tree (i.e., variability of the outcome of the tree) by bootstrapping the sample and create a sample of decision trees via a two-step procedure:

1. For $b = 1, ..., B$, we sample n observations (y_i, x_i) with replacement to create B training sets and consequently, B trees (which are generally not pruned as the reduction in variance will come from Step 2). In other words, the trees are built using the recursive binary splitting tree procedure.

2. For an observation x, we predict

$$\widehat{f}_{\text{bag}}(x) = \frac{1}{B}\sum_{b=1}^{B}\widehat{f}^{b}(x),$$

where $\widehat{f}^{b}(x)$ is the prediction made from the b-th iterative step of the bootstrap procedure, i.e. the b-th bootstrapped tree (For classification, rather than an average, we can simply use the majority rule - assign observation to the most represented outcome among the B bootstrapped trees).

We can use this approach to perform "Out-of-bag" (OOB) error estimation. Indeed, when bootstrapping, we leave out approximately 1/3 of the sample. For a given (y_i, x_i), we can measure error by using prediction done by bootstrapped trees that did not use (y_i, x_i) to construct the tree. More precisely, we measure

$$\widehat{f}_{\text{bag}}^{\text{OOB}}(x_i) = \frac{\sum_{b=1}^{B}\widehat{f}^{b}(x_i)\,1_{\{b\text{th tree does not use }(y_i,x_i)\}}}{\sum_{b=1}^{B}1_{\{b\text{th tree does not use }(y_i,x_i)\}}}.$$

We can get the overall OBB MSE (regression)

$$\sum_{i=1}^{n}\left(y_i - \widehat{f}_{\text{bag}}^{\text{OOB}}(x_i)\right)^2.$$

Classification problems follow along the same lines but by replacing averages with the majority rule. For an observation (y_i, x_i), the OOB classification prediction is the most commonly occurring prediction among all bootstrapped trees that did not the observation (y_i, x_i) in its construction phase.

Generally speaking, bagging improves prediction accuracy at the expense of interpretability (as the final outcome is an average over a large number of trees). For instance, it is not as easy to communicate how the response relates to the predictors in the model. This is also true to determine the importance of each predictor in the model. A possible way to quantify this importance is to utilize the concept of RSS (for bagging regression trees) and classification error (for bagging classification trees). In the case of bagging regression trees, we can record the total amount that the RSS is decreased due to splits over a given predictor, averaged over all B bootstrapped trees. A large (small) value indicates an important (not-so-important) predictor.

4.3.2 Random forests

Random forests impose a small tweak to the bagging procedure. Indeed, as in bagging, we build a number of decision trees on bootstrapped training samples. But when building these decision trees, each time a split in a tree is considered, a random sample of m predictors are selected among the most comprehensive set of p predictors ($m \leq p$). We must choose one of the m predictors to perform the split. As a rule of thumb, we choose $m \approx \sqrt{p}$.

What is the rationale behind this procedure? If we don't do that, it is likely that the collection of bootstrapped trees will look very similar and hence, the reduction in variance of the prediction may be minimal. Note that if $m = p$, the random forest procedure reverts to bagging.

Please note that we will make use of the function `randomForest` in the R library `randomForest` to run this tree procedure. The same is true for the bagging procedure as bagging can be viewed as a special case of the random forest procedure.

4.3.3 Boosting (skip)

Like bagging, the boosting approach can be used in conjuncture with many other statistical learning methods including regression and classification problems. The approach consists in creating a large number of trees to make prediction. However, unlike bagging, the trees are not independent but built sequentially on a given training sample (no bootstrapping involved). Each subsequent tree construction depends on information from previously grown trees. Each tree is built on a modified version of the original data set. The approach is designed to learn slowly using the following approach.

The boosting procedure has three tuning parameters: the number of trees B, the shrinkage parameter λ and the number of splits in each tree d (also known as the interactive depth). The procedure goes as follows:

- Initialize the algorithm by setting $\widehat{f}(x) = 0$ and $r_i = y_i$ (where r_i stands for the i-th residual) for all i in the training set
- For $b = 1, ..., B$, repeat the following steps:
  - Fit a tree with d splits (i.e., $d+1$ terminal nodes) to the training set (r_i, x_i) for all i. For a predictor x_i, the tree generates a prediction of $\widehat{f}^b(x_i)$.
  - Update $\widehat{f}$ to
$$\widehat{f}(x) \leftarrow \widehat{f}(x) + \lambda \widehat{f}^b(x)$$
  - Update the residual to
$$r_i \leftarrow r_i - \lambda \widehat{f}^b(x_i)$$
- Output
$$\widehat{f}(x) = \sum_{b=1}^{B} \lambda \widehat{f}^b(x)$$

Some remarks on the tuning parameters:

- B too large may cause overfitting
- $l = 0.01$ or 0.001 are typical. λ is a parameter of the speed of learning. Very small λ may require larger B to achieve good performance.
- $d = 1$ normally works well. Each generated tree consists of a single split (2 terminal nodes)