Estimation Methods for Incomplete Insurance Data

Noah Gblonyah, Seth Okyere, Michael Throolin

29 April 2022

Introduction

The growth of the insurance industry is fueled by societal expectations for protection against a variety of risks associated with unfavorable random events that have a large economic impact. Insurance is a process that involves the payment of a premium in exchange for an equitable manner of offsetting the risk of a potential future loss. The basic idea is to set up a fund to which insured members contribute specified sums of premium for specific loss levels. When the random events that policyholders are protected against occur, this gives rise to claims which are then settled from the fund.

The Maximum Likelihood Estimation (MLE) technique is commonly used by insurance firms to estimate claim distribution parameters. Maximum likelihood techniques are typically applied to complete data sets where we have exact values for all of the data points; however, data is rarely perfect in real life. When modeling the underlying loss variable, insurance contracts have coverage adjustments that must be taken into account. Typically, loss control methods such as deductibles, policy limits, and coinsurance are implemented to reduce undesirable policyholder behavioral effects such as adverse selection.

Estimation by method of moments and maximum likelihood are often easy to do, but these estimators tend to perform poorly, mainly because they use a few features of the data rather than the entire set of observations. It is important to use as much information as possible when the population has a heavy tail as in insurance data. This paper discusses some considerations for properly handling truncated and censored data for modeling of insurance data. The methods to be discussed are Maximum Likelihood Estimations and the EM Algorithm.

Methodology

Maximum Likelihood Estimation

Insurance Contracts

Insurance contracts have coverage modifications that need to be considered when modeling the underlying loss variable. Usually, the coverage modifications such as deductibles, policy limits, and coinsurance are introduced as loss control strategies so that unfavorable policyholder behavioral effects can be minimized. There are also situations when certain features of the contract emerge naturally (e.g., the value of insured property in general insurance is a natural upper policy limit). Here we describe two common transformations of the loss variable along with the corresponding probability distribution/mass functions.

Truncation

Incomplete data in the form of truncation occurs when an observation is not recorded due to that observation being below or above a certain threshold. In practice these are referred to as left truncation and right

truncation respectively. A common example of left truncation that arises in practice when working with insurance data is with ordinary insurance deductibles.

Left Truncation

An observation is left truncated at d if it is not recorded when less than or equal to d and recorded at the observed value if the observation is greater than d. Let X be a random variable representing the size of loss and L be a random variable representing the recorded value. Mathematically, left truncation would be represented as:

$$L = \begin{cases} \text{not recorded} & : x \leq d \\ x & : x > d \end{cases}$$

The likelihood function is

$$f(x|x > d) = f(x)/S(d)$$

where $S(\cdot) = 1 - F(\cdot)$ There are a variety of ways that insurance deductibles operate, but with ordinary insurance deductibles there is no incentive for policyholders to report losses less than the deductible amount to the insurance company. In this situation, the insurer's data is left truncated for these losses which would be paid by the policyholder and not recorded in the insurer's data systems.

Censoring

Incomplete data in the form of censoring occurs when the observation is recorded at a fixed value if it is below or above a certain threshold. In practice this is referred to as left censoring and right censoring respectively. The right censoring situation arises frequently when working with policy limits in insurance data.

Right Censoring

An observation is right censored at u if it is recorded at its observed value if less than u and recorded at u if the observed value is greater than or equal to u. Let X be a random variable representing the size of loss and L be a random variable representing the recorded value. Mathematically, right censoring would be represented as:

$$L = \begin{cases} x & : x < u \\ u & : x \geq u \end{cases}$$

$$f(x|u) = F_X(u) + S_X(u)$$

It is common for an insurance policy to have a limit which is the maximum amount that the insurer will pay under the terms of the insurance agreement. In situations where the actual damages exceed the limits of the policy, the payment from the insurer will be limited to the policy limit and the loss will be considered right censored.

The derivation of the estimator for the parameter through the maximum likelihood approach is as follows:

- Likelihood function

$$
\begin{aligned}
L(\theta) &= \prod_{i=1}^{n} \frac{\frac{1}{\theta} e^{-x_i/\theta}}{e^{-d_i/\theta}} \cdot \prod_{i=n+1}^{n+c} \frac{e^{-x_i/\theta}}{e^{-d_i/\theta}} \\
&= \prod_{i=1}^{n} \frac{1}{\theta} e^{-(x_i-d_i)/\theta} \cdot \prod_{i=n+1}^{n+c} e^{-(x_i-d_i)/\theta} \\
&= \frac{1}{\theta^n} e^{-\sum_{i=1}^{n+c}(x_i-d_i)/\theta}
\end{aligned}
$$

- The log-likelihood and derivative

$$
\begin{aligned}
\ell(\theta) &= -n \ln \theta - \frac{\sum_{i=1}^{n+c}(x_i - d_i)}{\theta} \\
\ell'(\theta) &= -\frac{n}{\theta} + \frac{\sum_{i=1}^{n+c}(x_i - d_i)}{\theta^2} = 0
\end{aligned}
$$

- The estimator

$$
\hat{\theta}_{MLE} = \frac{\sum_{i=1}^{n+c}(x_i - d_i)}{n}
$$

Application

To demonstrate how the methods described in the previous section is used in the insurance industry, We are given the information in Table 1 about a group of policies and assume payments at the policy limit resulted from losses above the maximum covered loss.

Table 1: Insurance Claims payments with deductibles and policy limits.

Claim.Payment	Deductible	Policy.Limit
30	0	80
50	10	100
80	10	100
120	20	150
150	30	150

We also will assume that the losses follow the exponential distribution. The aim is to determine the likelihood function and the maximum likelihood estimate of the mean losses assuming that the losses are independent.

Table 2: Insurance data with maximum covered losses and their respective likelihood funtions.

Loss.Payment	Deductible	Maximum.Covered	likelihood.function
30	0	80	f(30)
60	10	110	f(60)/S(10)
90	10	110	f(90)/S(10)
140	20	170	f(140)/S(20)
180+	30	180	S(180)/S(30)

$$L(\theta|x) = f(30)\frac{f(60)}{S(10)}\frac{f(90)}{S(10)}\frac{f(140)}{S(20)}\frac{S(180)}{S(30)}$$

$$L(\theta|x) = \frac{1}{\theta}e^{-\frac{30}{\theta}}\frac{\frac{1}{\theta}e^{-\frac{60}{\theta}}}{e^{-\frac{10}{\theta}}}\frac{\frac{1}{\theta}e^{-\frac{90}{\theta}}}{e^{-\frac{10}{\theta}}}\frac{\frac{1}{\theta}e^{-\frac{140}{\theta}}}{e^{-\frac{20}{\theta}}}\frac{e^{-\frac{180}{\theta}}}{e^{-\frac{30}{\theta}}}$$

$$L(\theta|x) = \frac{1}{\theta}e^{-\frac{30}{\theta}}\frac{1}{\theta}e^{-\frac{50}{\theta}}\frac{1}{\theta}e^{-\frac{80}{\theta}}\frac{1}{\theta}e^{-\frac{120}{\theta}}e^{-\frac{150}{\theta}}$$

$$L(\theta|x) = \frac{1}{\theta^4}e^{-\frac{430}{\theta}}$$

$$logL(\theta|x) = \ell(\theta|x) = -4ln(\theta) - \frac{430}{\theta}$$

$$\frac{d}{d\theta}\ell(\theta|x) = -\frac{4}{\theta} + \frac{430}{\theta^2} = 0$$
$$-4\theta + 430 = 0$$
$$4\theta = 430$$
$$\theta = \frac{430}{4} = 107.5$$

Now we check if θ is the maximum.

$$\frac{d^2}{d\theta^2}\ell(\theta|x) = \frac{4}{\theta^2} - \frac{860}{\theta^3}$$
$$\frac{d^2}{d\theta^2}\ell(\theta|x) = \frac{4}{107.5^2} - \frac{860}{107.5^3} = -0.000346 < 0$$

Hence, $\hat{\theta}_{MLE} = 107.5$

Alternatively

$\hat{\theta}_{MLE}$ for an exponential distribution is given by

$$\hat{\theta}_{MLE} = \frac{\sum_{i=1}^{n+c}(x_i - d_i)}{n}$$
$$\hat{\theta}_{MLE} = \frac{30 + 50 + 80 + 120 + 150}{4} = 107.5$$

Where;

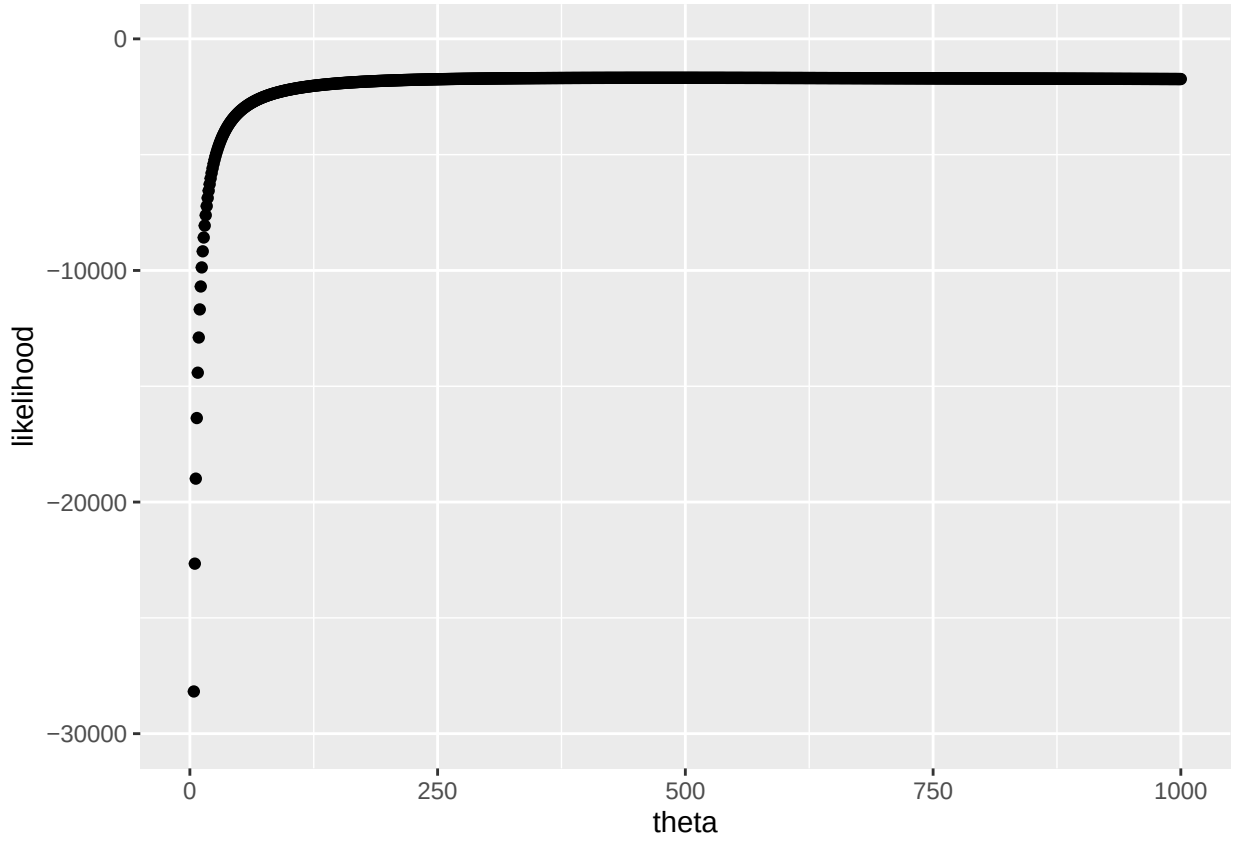
n = number of uncensored data c = number of censored data points x_i = observed loss value, or censoring point, or censored data d_i = truncated point

```
insurance_payment <- read_excel("insurance.payment.xlsx")

exponential.mle <- function(payment,max.covered){
  theta.mle <- sum(payment)/ length(payment[payment < max.covered])
  return(theta.mle)
}

exponential.likelihood <- function(payment, theta){
  n <- length(payment)
```

```
  likelihood <- 0
  for(i in 1:length(theta)){
  power <- -sum(payment)/theta[i]
  likelihood[i] <- -n*log(theta[i])+power
  }
  data <- tibble(theta = theta, likelihood = likelihood)
  plot <- data %>%
    ggplot(aes(x = theta, y = likelihood)) +
    geom_point()+
    xlim(0,1000) +
    ylim(-30000,0)
 return(plot)
}
payment <- insurance_payment$payment
max.covered <- 1100
theta <- 1:7000
exponential.mle(payment ,  max.covered)
```

```
## [1] 525.5
```

```
exponential.likelihood(payment, theta)
```



The EM Algorithm

The EM Algorithm has its roots in work done in the 1950s but really came into statistical prominence after the seminal work of Dempster, Laird, and Rubin, which detailed the underlying structure of the algorithm and illustrated its use in a wide variety of applications. (Casella and Berger 2002).

Another common tool used for getting a maximum-likelihood estimation of censored data is called the EM Algorithm. Here, the 'E' canonically stands for the 'Expectation' step and 'M' represents the 'Maximization' step. Hence the EM Algorithm takes the expectation of the log-likelihood function, then maximizes that quantity. It repeats that process until the parameter converges to a specified value.

Formally, if we let $\theta^{(p)}$ represent the p^{th} iteration of the algorithm to estimate the parameter θ. these two steps can be written out as follows:

> *Expectation (E-Step):* Compute $Q\left(\theta^{(p)}|\theta^{(p-1)}\right) = E\left[\log\left(f(\mathbf{x}|\theta^{(p)}\right)|\mathbf{y}, \theta^{(p-1)}\right]$ where $\mathbf{x}$ represents the complete data and $\mathbf{y}$ represents the censored, or incomplete data.
>
> *Maximization (M-Step):* Maximize $Q\left(\theta^{(p)}|\theta^{(p-1)}\right)$

The EM algorithm can often lead to functions that are tricky to evaluate. However in special cases, such as the exponential family family case, the algorithm becomes much easier to evaluate. Specifically a function $f(\mathbf{x}|\theta)$ is an exponential family if it can be written as $f(\mathbf{x}|\theta) = h(\mathbf{x})c(\theta)\exp\left(\sum_{i=1}^{k} w_i(\theta)t_i(\mathbf{x})\right)$. It has been shown that we can use the complete sufficient statistic $T(\mathbf{X}) = \sum_{i=1}^{k} t_i(\mathbf{x})$ to estimate the parameter θ. This is done as follows:

> *Expectation (E-Step):* Estimate $\mathbf{t}(\mathbf{x})$ by finding $\mathbf{t}^{(p-1)} = E\left(\mathbf{t}(\mathbf{x})|\mathbf{y}, \theta^{(p-1)}\right)$ where $\mathbf{x}$ represents the complete data and $\mathbf{y}$ represents the censored, or incomplete data.
>
> *Maximization (M-Step):* Determine $\theta^{(p)}$ as the solution to $E\left(\mathbf{t}(\mathbf{x})|\theta\right) = \mathbf{t}^{(p-1)}$

Example (Exponential Distribution)

For a definitive example, suppose we have data from an exponential distribution with unknown parameter θ. For each sample, we are giving a vector of values, $(c_{(1,i)}, c_{(2,i)}, x_i)$, where $c_{(1,i)}$ represents a left-censoring point, $c_{(2,i)}$ represents a right-censoring point, and x_i is the value the sample. If the i^{th} sample is present, or rather if $c_{(1,i)} < x_i < c_{(2,i)}$, we will call define $y_i = x_i$. Analogously, if the i^{th} sample is left-censored, or $c_{(1,i)} \geq x_i$, we will define $l_i = x_i$ and if it is right-censored, or $x_i \geq c_{(2,i)}$, we will define $r_i = x_i$. Hence, $\mathbf{x}$ is a vector of our complete data, $\mathbf{y}$ is a vector of our incomplete data, $\mathbf{l}$ is a vector of our left-censored data, and $\mathbf{r}$ is a vector of our right-censored data. Our object is to use the EM algorithm to estimate θ using only y and the censoring points $c_{(1,i)}$ and $c_{(2,i)}$ corresponding with the values known in $\mathbf{l} and \mathbf{r}$.

To begin, note that a random sample for the exponential family has a complete sufficient statistic of $\sum_{i=1}^{n} x_i$. Hence for the *E-Step* we must find the expectation of $E(\sum_{i=1}^{n} x_i|\mathbf{y}, \theta^{(p-1)})$, which we can expand as:

$$E\left(\sum_{x_i \in \mathbf{y}} y_i + \sum_{x_i \in \mathbf{l}} l_i + \sum_{x_i \in \mathbf{r}} r_i\right)$$

Now, we need to estimate our censored data, l_i and r_i. To do this, we will use the memoryless property of the exponential distribution to get $l_i = \min\{\theta^{(p-1)}, c_{(1,i)}\}$ and $r_i = c_{(1,i)} + \theta^{(p-1)}$.

Note that

$$\begin{aligned} E(l_i) &= \frac{\int_0^{c_{(1,i)}} \min\{\theta^{(p-1)}, c_{(1,i)}\} \frac{1}{\theta^{(p-1)}} e^{-x_i/\theta^{(p-1)}}\, dx_i}{\int_0^{c_{(1,i)}} \frac{x_i}{\theta^{(p-1)}} e^{-x_i/\theta^{(p-1)}}\, dx_i} \\ &= \frac{\min\{\theta^{(p-1)}, c_{(1,i)}\} - \min\{\theta^{(p-1)}, c_{(1,i)}\} e^{-c_{(1,i)}/\theta^{(p-1)}}}{\theta^{(p-1)} - (c_{(1,i)} + \theta^{(p-1)}) e^{-c_{(1,i)}/\theta^{(p-1)}}} \end{aligned}$$

Therefore, we can simplify our expectation of $E\left(\sum_{x_i \in \mathbf{y}} y_i + \sum_{x_i \in \mathbf{l}} l_i + \sum_{x_i \in \mathbf{r}} r_i\right)$ to

$$\sum_{x_i \in \mathbf{y}} y_i + \sum_{x_i \in \mathbf{l}} \frac{\min\{\theta^{(p-1)}, c_{(1,i)}\} - \min\{\theta^{(p-1)}, c_{(1,i)}\} e^{-c_{(1,i)}/\theta^{(p-1)}}}{\theta^{(p-1)} - (c_{(1,i)} + \theta^{(p-1)}) e^{-c_{(1,i)}/\theta^{(p-1)}}} + \sum_{x_i \in \mathbf{r}} c_{(1,i)} + \theta^{(p-1)}$$

And our maximization step is the solution to

$$\begin{aligned} E(\mathbf{t}(\mathbf{x})|\theta^{(p)}) &= E(\sum_{i=1}^{n} x_i|\theta^{(p)}) = n\,\theta^{(p)} \\ &= \sum_{x_i \in \mathbf{y}} y_i + \sum_{x_i \in \mathbf{l}} \frac{\min\{\theta^{(p-1)}, c_{(1,i)}\} - \min\{\theta^{(p-1)}, c_{(1,i)}\} e^{-c_{(1,i)}/\theta^{(p-1)}}}{\theta^{(p-1)} - (c_{(1,i)} + \theta^{(p-1)}) e^{-c_{(1,i)}/\theta^{(p-1)}}} + \sum_{x_i \in \mathbf{r}} c_{(1,i)} + \theta^{(p-1)} \\ \implies \theta^{(p)} &= \frac{1}{n}\left(\sum_{x_i \in \mathbf{y}} y_i + \sum_{x_i \in \mathbf{l}} \frac{\min\{\theta^{(p-1)}, c_{(1,i)}\} - \min\{\theta^{(p-1)}, c_{(1,i)}\} e^{-c_{(1,i)}/\theta^{(p-1)}}}{\theta^{(p-1)} - (c_{(1,i)} + \theta^{(p-1)}) e^{-c_{(1,i)}/\theta^{(p-1)}}} + \sum_{x_i \in \mathbf{r}} c_{(1,i)} + \theta^{(p-1)}\right) \end{aligned}$$

Simulation (Exponential Distribution)

In the actuarial context, we could encounter data that includes the deductible, policy limit, and losses for each customer, where the losses would be unreported if they exceed the policy limit or are below the deductible.

We will use computer simulation to see how well the algorithm holds given the parameter $\theta = 1000$.

```
set.seed(53523)

#Simulate Data
n_customers <- 1000
damages <- round(rexp(n_customers, 1/1000),2)
deductible <- round(runif(n_customers, min = 0, max = 600))
limit <- round(runif(n_customers, min = 5000, max = 20000))

#combine data into one data frame
customers <- as.data.frame(cbind(deductible, limit, damages))
#censor data
left <- customers[(damages < deductible),]
right <- customers[(limit < damages),]

observed <- customers %>%
  filter(damages > deductible & limit > damages)


#EM Algorithm
sum_observed <- sum(observed$damages)

#Initialize theta
theta_new <- 500
theta <- 0

#iterate until difference between previous theta and new theta is small
while((theta - theta_new)^2 > 0){
  theta <- theta_new
```

```
  #Expectation step
  m <- min(theta, left$deductible)

  #Maximization step
  theta_new <- (sum((m-m*exp(-left$deductible/theta)) / (theta -(left$deductible + theta)*
    exp(-left$deductible/theta))) + sum(right$limit + theta) + sum(sum_observed))/n_customers
}
theta_new #display outcome
```

```
## [1] 970.7771
```

The output from this simulation estimated the value of θ to be 970.777086, which is 29.222914 from the known value of 1000.

References

Casella, George, and Roger L. Berger. 2002. *Statistical Inference*. Duxbury.

Dempster, A. P., N. M. Laird, and D. B. Rubin. 1977. "Maximum Likelihood from Incomplete Data via the EM Algorithm." *Journal of the Royal Statistical Society: Series B (Methodological)* 39 (1): 1–22. https://doi.org/https://doi.org/10.1111/j.2517-6161.1977.tb01600.x.

Hartley, H. O. 1958. "Maximum Likelihood Estimation from Incomplete Data." *Biometrics* 14: 174.

Klugman, Stuart et al. 2004. *Loss Models : From Data to Decisions, 2nd Edition*. Hoboken, N.J: Wiley Interscience.

Peng, Roger D. 2021. *4.1 EM Algorithm for Exponential Families | Advanced Statistical Computing*. https://bookdown.org/rdpeng/advstatcomp/em-algorithm-for-exponential-families.html.

Poudyal, Chudamani. 2018. "Robust Estimation of Parametric Models for Insurance Loss Data." *Theses and Dissertations*.

Truxillo, Catherine. 2005. "Maximum Likelihood Parameter Estimation with Incomplete Data." In *Proceedings of the Thirtieth Annual SAS (r) Users Group International Conference*, 111–30. Citeseer.

R Packages

Henry, L. and H. Wickham (2020). *purrr: Functional Programming Tools*. R package version 0.3.4. https://CRAN.R-project.org/package=purrr.

Müller, K. and H. Wickham (2022). *tibble: Simple Data Frames*. R package version 3.1.8. https://CRAN.R-project.org/package=tibble.

R Core Team (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. ISBN: 978-3-319-24277-4. https://ggplot2.tidyverse.org.

——— (2022a). *forcats: Tools for Working with Categorical Variables (Factors)*. R package version 0.5.2. https://CRAN.R-project.org/package=forcats.

——— (2022b). *stringr: Simple, Consistent Wrappers for Common String Operations*. R package version 1.4.1. https://CRAN.R-project.org/package=stringr.

——— (2022c). *tidyverse: Easily Install and Load the Tidyverse*. R package version 1.3.2. https://CRAN.R-project.org/package=tidyverse.

Wickham, H., M. Averick, J. Bryan, et al. (2019). "Welcome to the tidyverse". In: *Journal of Open Source Software* 4.43, p. 1686. DOI: 10.21105/joss.01686.

Wickham, H. and J. Bryan (2022). *readxl: Read Excel Files*. R package version 1.4.1. https://CRAN.R-project.org/package=readxl.

Wickham, H., W. Chang, L. Henry, et al. (2022). *ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics*. R package version 3.3.6. https://CRAN.R-project.org/package=ggplot2.

Wickham, H., R. François, L. Henry, et al. (2022). *dplyr: A Grammar of Data Manipulation*. R package version 1.0.10. https://CRAN.R-project.org/package=dplyr.

Wickham, H. and M. Girlich (2022). *tidyr: Tidy Messy Data*. R package version 1.2.1. https://CRAN.R-project.org/package=tidyr.

Wickham, H., J. Hester, and J. Bryan (2022). *readr: Read Rectangular Text Data*. R package version 2.1.2. https://CRAN.R-project.org/package=readr.

Zhu, H. (2021). *kableExtra: Construct Complex Table with kable and Pipe Syntax*. R package version 1.3.4. https://CRAN.R-project.org/package=kableExtra.