Study Note on Truncation and Censoring

1 Introduction

It is common in modeling and data analysis to encounter incomplete data in the form of *truncation* and *censoring*. This is especially true when working with insurance data, however, it also arises in many other areas of data analysis. This study note will provide a discussion on the topics of truncation and censoring as well as discuss some considerations for properly handling truncated and censored data for modeling of insurance data.

2 Background

2.1 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.

2.1.1 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:

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

A common example of left truncation that arises in practice when working with insurance data is with ordinary insurance deductibles. 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[1]. In this situation, the insurer's

[1] There are a variety of ways in which deductibles apply. In lines of business such as Auto Physical Damage there is no incentive for policyholders to report claims below the deductible because insurance coverage does not apply to these losses. These types of deductibles are referred to as ordinary deductibles. Another type of deductible program is called a deductible reimbursement plan. Deductible reimbursement plans exist within commercial insurance where

data is considered to be left truncated for these losses which would be paid by the policyholder and not recorded in the insurer's data systems.

2.1.2 Right Truncation

An observation is right truncated at u if it is recorded at its observed value when the value is less than u and not recorded when the 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 truncation would be represented as:

$$L = \begin{cases} \text{x} & : x < \text{u} \\ \text{not recorded} & : \text{x} \geq \text{u} \end{cases}$$

This form of truncation occurs far less often in practice, however, right truncation is described in this study note for completeness.

2.2 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.

Some insurance policies apply different limits to different types of payments. For example, some liability policies place a limit on payments made toward a settlement with the claimant but not on the payments made toward defense costs; thus, some analyses separate these two components of cost because one is censored and the other is not. Also, many policies have separate limits at various levels, such as per claimant, per claim, per location, or per policy year. It is important to understand the details of how the policy limits work for the policies you are analyzing.

2.2.1 Left Censoring

An observation is left censored at d if it is recorded at d if the observed value is less than d and recorded at its observed value if the observation is greater than or equal to d. Let X be a random

the insurance company will pay for losses that fall under the deductible and seek reimbursement from the policyholder for these loss payments. When analyzing data that has been impacted by a deductible program, it is important to understand how the deductible program operates.

variable representing the size of loss and L be a random variable representing the recorded value. Mathematically, left censoring would be represented as:

$$L = \begin{cases} \mathrm{d}: x < \mathrm{d} \\ \mathrm{x}: \mathrm{x} \geq \mathrm{d} \end{cases}$$

An example of where this type of censoring occurs is with measuring devices. For example, a student is measuring the minimum daily temperature using a thermometer that has a minimum temperature reading of -10° F. In situations where the minimum temperature is actually less than -10° F, the thermometer will measure the temperature as -10° F. In this situation the temperature data measured is considered to be left censored.

2.2.2 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:

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

This situation arises frequently when working with policy limits in insurance data. 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.

3 Considerations for Working with Truncated and Censored Data

When working with data it is critically important for the analyst to understand the impact of truncation and censoring on data. This section will provide some background on practical considerations that arise when working with truncated and censored data in insurance.

3.1 Incomplete Size of Loss Distribution

Understanding the size of loss distribution is an important component of many analyses. In particular in insurance the size of loss distribution is important for determining credits for deductibles and pricing for higher limits. As previously discussed in the examples, data recorded in the insurer's data systems are often impacted by truncation and censoring caused by policy characteristics such as deductibles and policy limits.

It is common to encounter data sets with experience from policies with a variety of deductible amounts. As mentioned during the left truncation example, it is common in lines of insurance such as Auto Physical Damage to only observe losses that are greater than the deductible amount. This is due to the fact that losses below the deductible amount are not reported to the insurance company and therefore not recorded in insurance company data systems. In this situation the data observed by the insurance company would be considered to be left truncated.

It is also common when working with insurance data to encounter losses from policies with a variety of policy limits. For example, in modeling loss data from a particular line of business there may be some policies with limits of $1M and some policies with limits of $5M. This means that losses will be censored at a variety of thresholds.

Both actuarial and statistical methods include techniques for making appropriate inferences from truncated and censored data. See [1] for actuarial methods and any standard reference on survival models for statistical methods. Note that many survival model methods can be applied to insurance loss data with size of the claim being treated as the time dimension.

3.2 Deductible and Policy Limit Correlations with Other Variables

In some cases, policy features such as deductible amounts and policy limits may be correlated and/or aliased with other characteristics of policyholders. When building insurance loss models, it is a best practice to estimate deductible and policy limit factors through a separate analysis rather than estimating insurance losses using policy features such as deductibles and policy limits as covariates. The example below illustrates some of the challenges of using these policy features in insurance loss modeling:

XYZ Insurance offered a competitive personal auto product which launched 20 years ago with a $1000 deductible. XYZ has since retired this insurance product for new customers, but continues to renew the product for existing customers. It is observed today that these products have lower insurance loss relative to products offered today which now have a $2000 deductible. In this situation the insurance loss difference between the two products is likely caused by other covariates (e.g. age) rather than the product features such as deductible. In this example aliasing could become a serious ratemaking issue if deductible was used as an alias for other policyholder characteristics and policyholders can quote multiple deductible levels. For example, in an extreme case this could lead to a nonsensical situation where the $1000 deductible would be cheaper than the $2000 deductible option.

It is a best practice to separately estimate deductible and policy limit factors. One common classical actuarial approach is based on analyzing empirical size of loss distributions. Alternatively, one can arrive at appropriate limit and deductible factors by building a severity model that reflects the size of loss distribution realistically. As one often needs to incorporate covariates, this can be done with tools like GLMs, though not usually with a GLM itself as the size of loss distribution is often not one of those allowed by a GLM (i.e., not in the exponential family).

After arriving at limit and/or deductible factors as above, these can be included as adjustments to the target variable in the next stage of the modeling process (be it a GLM, a random forest, etc.) where other predictors (such as age in the Auto example above) enter.

3.3 Analyst-Imposed Censoring Scheme

In some situations the practitioner may self-impose a censoring scheme to reduce the volatility of data. This is generally done by having the analyst select a particular limit to apply to individual loss amounts. Some lines of insurance such as those covering excess layers can have highly volatile loss experience from year to year.

Increased limits factors arrived at by a separate analysis as described previously can then be used to build back to a full rate for each risk profile and limit combination.

4 Summary

Truncation and Censoring impact many different types of data analyses. This particularly true in insurance where left truncation and right censoring are quite common with deductibles and policy limits, respectively. It is important for the analyst to understand *whether* and *to what degree* data has been impacted by truncation and censoring.

References

[1] Klugman, Stuart et. al *Loss Models: From Data to Decisions, 2nd Edition*. John Wiley & Sons Inc., Hoboken, New Jersey, 2004