Summer Projects of Lumber problem

Jiahua Chen

April 22, 2018

1 Introduction

For the past many years, as a member of Lumber Project research team, I have been using DRM for long term monitoring purpose.

Real World Problem: The real world problem is conceptually very clear: there is a virtual population of wood product of each grade every year on market. Do they meet the quality standard announced/established by the industry? Is there a declining trend for the lumber of a specific grade?

Specific interpretation: For each mechanical property such as DOE or DOR, their values over the population form a probability distribution. One measurement of the population quality is the 5% (lower) quantile of this distribution. So one real world problem is: does the 5% quantile/percentile of the population on market meet the industrial standard? Is there any declining trend?

Abstract statistical approach: Collect a representative sample from the target population each year, construct a 95% confidence interval of this parameter. Address the real world problem by conducting relevant statistical significance test.

Complications: Constructing confidence intervals is not hard. How to make it short so that it has satisfactory precision is the true question. This task is not difficult if we have unbounded resources in time and money. We are luckier than medical studies where there are many more restrictions.

The practical consideration leads to many issues. We do not usually have i.i.d. observations though some of our initial development may assume such a data structure. The data in applications are clustered, have penal structure, become available in batches, not available for all products on the market and so on.

This summary gives a description of what we have achieved and the research problems we envisage and wish to address in the near future.

2 Chen's developments

The first project we worked out is based on **cross sectional data**. Suppose i.i.d. observations on strength of individual wood pieces are obtained from the population every year on the quality index of interest. The real world sampling plan is more complex than i.i.d. . The term "cross sectional" refers to the fact the samples from different years are collected independently. In comparison, **penal data** collects data from the same sampling units year after year. Taking blood pressures of the same collection of individuals over years result in typical penal data. In this application, we do not measure mechanical strength of the same piece of wood. However, we may take samples from the same set of mills over the years. Hence, penal data situation can occur from this perspective.

DRM for cross sectional data Denote the data as $\{\{x_{ij} : j = 1, 2, \ldots, n_i\} : i = 0, 1, \ldots, m\}$. Chen and Liu suggest that there may exist a latent structure in these population so that their corresponding distributions share some properties. The DRM is suggested based on this consideration:

$$\log\{f_i(x)/f_0(x)\} = \exp(\alpha_i + \beta_i^\tau q(x)\}$$

for $i = 1, \ldots, m$ and $f_i(\cdot)$ is the density function of the ith distribution. Namely, $f_i(x)$ is the distribution of $X_{i1}, X_{i2}, \ldots$.

This model assumption permits a nice EL and EL based data analysis on quantiles.

EL under DRM with cross sectional data The likelihood contribution of x_{ij} under DRM is identified as

$$\log p_{ij} + \alpha_i + \beta_i^\tau q(x_{ij})$$

where $p_{ij} = F_0(\{x_{ij}\})$.

The profile likelihood for $\theta = \{(\alpha_i, \beta_i) : i = 1, \ldots, m\}$ has the form

$$\ell_n(\theta) = \sup\{\sum \log p_{ij} + \sum\{\alpha_i + \beta_i^\tau q(x_{ij})\} : \text{constraints: } \sum p_{ij} g(x_{ij}, \theta) = 0.\}$$

Model fitting and properties The MELE of θ is found asymptotic normal. The value of $\hat{\theta}$ can be computed using a multinomial logistic regression software. The fitted values

$$\hat{p}_{ij} = \frac{1}{\sum_r \rho_r \exp\{\hat{\alpha}_r + \hat{\beta}_r q(x_{ij})\}}.$$

This leads to fitted distribution functions:

$$\hat{F}_i(x) = \sum \hat{p}_{ij} \exp(\hat{\alpha}_i + \hat{\beta}_i q(x_{ij})) \mathbb{1}(x_{ij} \leq x).$$

Fitted distributions lead to DRM-EL quantiles: $\hat{F}_i^{-1}(t)$ for $t \in (0, 1)$.

Claim: $\hat{F}_i^{-1}(t)$ for given set of i and t are also asymptotically normal. The asymptotic variances can be estimated in some way. These lead to means of testing hypothesis regarding to quantiles.

Much of the technical development has been on Bahadur representation. With such a representation, the joint distributions of the EL-quantiles are easy to determine. Hypothesis tests and confidence intervals in the corresponding paper are carried out by Wald's method.

Density estimation must be utilized as an intermediate step.

3 Other developments

Multi-parameter one-sided tests Guangyu Zhu under supervision of Jiahua, examined the problem of conduction multi-parameter one-sided tests. The method is developed based on several observations.

First, the parameter estimators are generally asymptotic normal. Hence, existing or new multi-parameter one-sided tests based on normal model are applicable.

Second, to control the type I error, existing methods determined the critical value based on the least possible variance matrix configuration. This practice hurts the power of the test though it is the theoretical must.

We decided to estimate the variance matrix of the parameter estimators. The critical value of the test is computed accordingly. This approach does not fully control the type I error under the nominal level. Yet we believe that in practical situations, the level of the test will not be overly large. The simulation experiments seem to support this claim. A paper about this method has been published on Technometrics. The method is relevant to forestry project.

Random effect In real world sampling plan, for the least, the data are clustered.

In simplistic fashion, the observations are grouped into clusters each of size 5 or 10. They are taken from the same "plot" from the same mill. Hence, these observations are more alike than observations from different plots/mills. If this effect is ignored, the variance of various estimators will be under estimated. This leads to inflated type I error for monitoring purposes.

For instance, under normal model, the type I error is typically inflated to 10% when the nominal is 5% under various model assumptions.

The DRM-EL based method has the same problem if we assume all observations are independent of each other. In the latest paper, we show that the method leads to asymptotically unbiased parameter estimators. However, their asymptotic variances are larger in the presence of cluster structure than it would be if the observations were i.i.d. / A cluster based bootstrap approach is found effective to get the “right” variance estimator. This leads to ‘valid’ confidence regions and test methods.

4 Something in progress

R-function package Currently, Song Cai has written a R-package for fitting DRM and many related functions to his thesis work. His code should include methods for censored data, enables hypothesis test on the value of DRM parameters. We also have codes for the results in the paper for the DRM-EL approaches, they are not organized in nice R-package format. Some diagnostic features and graphic tools can also be nice to have. Thus, Boyi Hu has been asked to develop a complete R-package based on results in these papers. This is a on-going project.

Likelihood interval Chen and Liu (2013AOS) use Wald’s method for confidence interval construction. This method relies on asymptotic normality of the estimator. For quantiles, the asymptotic variance depends on density function. Thus, their method involves a density estimation step.

Chen, Li, Liu and Zidak (2017) use bootstrap to construct confidence intervals. The paper focuses on clustered data and the method is also applicable to independent observations. There is nothing particularly bad about their method.

In general, the likelihood interval is superior. It does not require estimating the variance of parameter estimators needed in normal approximation based Wald method. The shape of the region is not the restrictive oval. Often, the approximation of the chisquare limiting distribution is more accurate compared to normal approximation used in Wald method. In the current application, the target parameter is quantile. When DRM-EL approach involves estimating functions for quantiles, these estimating functions are not smooth in quantiles. Also, the properties of the likelihood ratio statistic for quantiles under DRM have not been investigated. These properties are not directly implied by the standard or existing results. At the same time, the related technical problems seem manageable to a smart and hard working beginner. For this reason, Archer has been asked to work on this project. The initial study shows that the DRM-EL setting is free from empty-solution problem. The LRT for fixed set of quantile values is free from non-smooth issues. The limiting distribution appears to be standard chisquare. It is promising that some results can be obtained without much technical difficulties.

5 Possible research projects for new MSc students

R-package. YanChao has been instructed to join Boyi to get his R-functions completed and fully implemented. The final goal is to make it easy to show-off. Some specific tasks include getting familiar with the statistical methods involved, understand the R-functions developed by Boyi, test all functions from many angles he can image. For instance, YanChao can test these functions with data with various sample sizes, different number of populations, various data set from different distributions. Ideally, YanChao may discover new ways to apply these functions, new approaches to the statistical problem, and new statistical problems.

Comparison between DRM-EL and artificially censored approaches. Suppose of we have a random sample $x_1, \ldots, x_n$ from a population under investigation. In principle, a Weibull distribution with two parameters is assumed for the population distribution:

$$f(x, \alpha, \eta) = \alpha(x/\eta)^{\alpha-1} \exp\{-(x/\eta)^{\alpha}\}.$$

Once α and η are properly estimated, we get a corresponding estimate of the lower quantiles such as that of 5%.

In this respect, ASTM D5457 recommends a seemingly strange approach. Given a random sample $x_1, \ldots, x_n$, let $\hat{\xi}$ be the 0.3th sample quantile. Let

$$y_i = \min(x_i, \hat{\xi})$$

for $i = 1, \ldots, x_n$. Regarding $\hat{\xi}$ as non-random, one can work out the distribution of Y_i under Weibull distribution assumption. Maximum likelihood estimates of α and η are then obtained based on data $y_1, \ldots, y_n$. In addition, ASTM D5457 includes two numerical procedures for this purpose.

If the true distribution of $x_1, \ldots, x_n$ is Weibull, then ASTM D5457 conceptually loses efficiency compared to full likelihood approach. However, such parameter estimates can be heavily influenced by large valued observations. Because the Weibull distribution is unlikely a good fit to true population distribution, such heavy influence may distort the shape of the estimated density function at the lower end. This may lead to heavy bias in lower quantile estimation. This explains why ASTM D5457 recommends somewhat a statistically strange estimation method.

The work of Yang Liu et al. shows that this Artificial Censored (AC) method based on Weibull distribution assumption has acceptable performance. They placed their focus on the best possible amount of censorship in terms of MSE. In addition, they used bootstrap method to provide a variance estimate of the resulting quantile estimators.

The work of Song Cai and Jiahua Chen (Canadian J of Stat) is on applying DRM-EL to censored data. Compared to AC method, the DRM-EL is non-parametric, yet it requires multiple samples to make up the loss of without a Weibull distribution assumption. Hence, it should be less efficiency if the true distribution is approximately Weibull, at least when the Weibull is a good match of the true distribution at the lower end. Yet by giving away the Weibull distribution assumption, we may also gain in other respect. Thus, it is worthwhile to make a comparison between these two methods.

Let

$$q(x;\alpha,\beta) = \beta_0 + \beta_1 x^{\alpha} + \beta_2(\log x)$$

and revise the DRM to

$$\log\{f_i(x)/f_0(x)\} = \exp(q(x;\alpha,\beta)\}.$$

If the population distributions are indeed Weibull, then the above choice of $q(x)$ in DRM will have these population distributions correctly linked. Hence, the approach of Song and Chen has Weibull-AC approach covered from this angle. At the same time, this $q(x)$ is different from the DRM specified in our DRM papers. It contains unknown parameters in a nonlinear fashion in terms of α. Thus, the numerical problem will be different.

May we use simulation study to find out the pros and cons of two methods? The students working on this project should actively think about various ways to make the comparison.

Study of penal data due to rotating sampling plan Canada forestry industry uses a rotating sampling plan. In the beginning, 36 mills from a population of over 200 mills are randomly selected with probability proportional to their production. Each of them will contribute a sample of 5 or 10 pieces of lumber. (Somehow, I heard the total sample size is 360). They are divided into 6 groups each of 6 mills. In the subsequent year, one group of mills will be replaced by a fresh new 6 mills. This sampling plan will continue in such a rotational fashion.

When the corresponding measurements are normally distributed such as MOE (suggested by Jim), the hypothesis test and confidence interval problems are investigated by Carolyn Taylor (and Jim?). See page 247 in the review paper on “rotating panels of mills”. Based on real data which leads to parameter configurations in the normal model, they show that rotating sampling plan leads to more accurate parameter estimations. In addition, Carolyn has an effective Bootstrap method for variance estimation. Jim suggested that Jiahua may help on looking into theoretical issues behind the bootstrapping.

Jiahua is interested to learn the effect of rotating sampling plan to DRM-EL method. Namely, if we do not make any changes in data analysis aspect, will the type I error be lower or higher than the nominal level? Under the normal model, the rotating sampling plan makes the comparison in population means partly like the comparison under paired-experiment. Hence, the corresponding analysis is more efficient just like paired-t test is more powerful. However, DRM-EL method does not have any structure to directly take advantage of the paired-experiment arrangement. Hence, this suggestion is largely exploratory.

6 Cells of lumber products

Lumbers in the market are divided in terms of their sizes, grades and so on. My understanding is that there are at least 240 types of lumbers because combination number grows very fast. These types are referred to as cells.

The standard sample size drawn every year in current practice is 360. Yet only of these 240 cells is sampled in Canada. One question asked by Conroy is: if we wish to move from sampling in one cell to another cell, which other cell should be sampled?

Do you have any suggestions?

7 References

I have already cited many papers above.

1. Applied Statistics. 2018:

Using artificial censoring to improve extreme tail quantile estimates Yang Liu, Matias-Barrera, Ruben H. Zamar and James V. Zidek UBC.

2. title=Monitoring test under nonparametric random effects model, author=Chen, Jiahua and Li, Pengfei and Liu, Yukun and Zidek, James V, journal=arXiv preprint arXiv:1610.05809, year=2016

3. Verrill, S. and Kretschmann, D. E. and Evans, J. W., title = Simulations of strength property monitoring tests, note = Unpublished manuscript. Forest Products Laboratory, Madison, Wisconsin. Available at `http://www1.fpl.fs.fed.us/monit.pdf`, year = 2015,

4. Zhu,Guangyu and Chen,Jiahua, title = Multi-parameter One-Sided Monitoring Tests, journal = Technometrics, year = 2017, doi = 10.1080/00401706.2017.1371081,

5. Hypothesis testing in the presence of multiple samples under density ratio models, author=Cai, Song and Chen, Jiahua and Zidek, James V, journal=Statistica Sinica, volume=27, pages=716–783, year=2017

6. Chen, Jiahua and Liu, Yukun, Quantile and quantile-function estimations under density ratio model, journal=The Annals of Statistics, volume=41, number=3, pages=1669–1692, year=2013, publisher=Institute of Mathematical Statistics

7. astm2002standard, title=Standard practice for establishing allowable properties for visually-graded dimension lumber from in-grade tests of full-size specimens, author=ASTM, D, journal=American Society for Testing and Materials, West Conshohocken, PA, year=2002

8. Annual Review of Statistics and Its Application. Statistical Challenges in Assessing the Engineering Properties of Forest Products. James V. Zidek and Conroy Lum