Mugla Journal of Science and Technology

LINEAR REGRESSION MODEL ESTİMATİON FOR RIGHT CENSORED DATA

Ersin YILMAZ[1], Dursun AYDIN[1]

[1]Muğla Sıtkı Koçman Üniversitesi Fen Fakültesi İstatistik Bölümü, 48000, Muğla, Türkiye
yilmazersin13@hotmail.com, daydin@mu.edu.tr

Received: 19.10.2015, Accepted: 09.04.2016
*Corresponding author

Abstract

In this study, firstly we will define a right censored data. If we say shortly right-censored data is censoring values that above the exact line. This may be related with scaling device. And then we will use response variable acquainted from right-censored explanatory variables. Then the linear regression model will be estimated. For censored data's existence, Kaplan-Meier weights will be used for the estimation of the model. With the weights regression model will be consistent and unbiased with that. And also there is a method for the censored data that is a semi parametric regression and this method also give useful results for censored data too. This study also might be useful for the health studies because of the censored data used in medical issues generally.
Keywords: Right-Censored Value, linear Regression, Kaplan-Meier

SAĞDAN SANSÜRLÜ VERİLER İÇİN DOĞRUSAL REGRESYON MODELİ TAHMİNİ

Özet

Bu çalışmada öncelikle sağdan sansürlü verinin ne olduğu tanımlanacaktır. Kısaca söylemek gerekirse, veri seti içerisinde belirlenmiş bir değerin üzerinde gözlemlenen değerler sansürlenir. Daha doğrusu belli nedenlerden ötürü gözlemlenmesi de mümkün olmayabilir. Bu verilere sağdan sansürlü veriler denir. Bu çalışmada sağdan sansürlenmiş açıklayıcı değişkenlerden elden edilmiş cevap değişkeni elde edilecek ve sansürlü veriler için de Kaplan-Meier ağırlıkları kullanılarak doğrusal regresyon modeli tahmin edilecektir. Burada kaplan-meier ağırlıklarının kullanılması tahmin edilecek olan regresyon modelinin parametrelerinin tahminlerinin tutarlı ve yansız olmalarını sağlamak içindir. Bu çalışmanın sağlık verilerinin analiz edilmesinde kullanılmasının yarar sağlaması öngörülmektedir. Genellik hasta gözlem verilerinin çoğunluğu sansürlü veriler olduğundan yansız ve tutarlı tahmincilerin uygun olduğu bu konuda açıktır.
Anahtar Kelimeler: Sağdan sansürlü veri, doğrusal regresyon, Kaplan-Meier

1 Introduction

In statistics', engineering, economics and medical studies, when the observing data have known partially, censoring has to be obligatory. It could be explained an example like this; The device or scale that you measure the thing, is not be able to measure all of the interval about the observing data. For example; suppose that there is a weighing machine that does not measure above 150 kilos. So this machine can't measure the 160 kilos. And if there are things that above 160 kilos, so they have to be right censored. Of course, this example could expand to another censoring methods but in this study, there is no need. The problem about the censored variables is; partial knowing about the variable. Censored data have various forms;

Left-Censored Data: An observation smaller than an exact value but value of the observation is unknown.

Interval-Censored Data: An observation have a value that in an exact interval.

Right-Censored Data: An observation value that above some determined value but, what is the value is not known.

Type I Censored Data: suppose that there is an experiment that founded several object started the experiment and the time of first spoiled object is observed and another times of broken of other objects is censored.

Type II Censored Data: Once again suppose that there is an experiment that formed several objects. There is exact value for the number of spoiled objects. When exact number objects spoiled, experiment is stopped and another objects is censored [5].

For the analyze of the censored variable, several methods have been improved. One of the individuals that have studied about this subject is Daniel Bernoulli. He has interested in the people that have vaccinated smallpox vaccine and he has collected that data in 1766. For that he has tried to find usefulness of smallpox vaccine. And then In 1989, Quesenberry has used Kaplan-Meier weights for the patient that disabled immune system data.

In medical studies, patients have joint to study in determined time interval, that patients would not observed all of the time that during their illness. In this situation, data that acquainted from that patients are right censored. On the other hand, censoring could be exist by many causes. A patient still alive although the study has been done or a patient may be resign from study.

First of all, in this study, apart from regression, Kaplan-Meier estimator has to be estimate. This estimator is a step function.

It estimates that the survival probability in exact time point. It has done that from before death to after that. And in this interval survival probability is the same. Because of that is step function and it has illustrated in Figure 1;

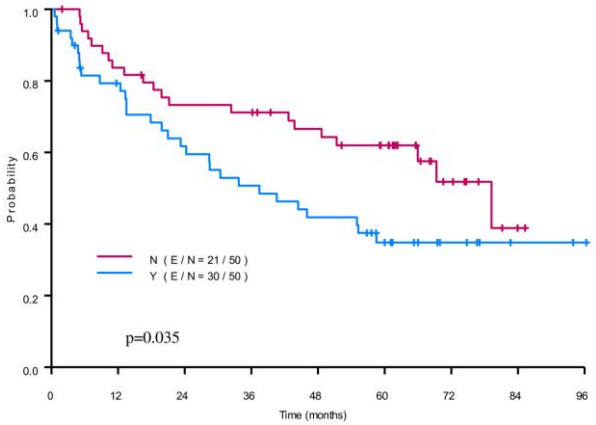

Figure 1. Kaplan-Meier Estimator.

2 Purpose

Censored regression models are in considerable position in statistics and econometrics. For the parametric estimation of censored regression models, residuals have to be come from parametric distribution family. The estimations that have made with parametric methods; although that, estimations have been inconsistent and biased, this is the opportunity for the comparison to compare better way to estimate and parametric method. In this paper, the better way is perhaps using the semi-parametric regression model for the estimation. This study might be use in fields where censored data have common use like survival analysis, medical surveys etc…

3 Method

Suppose that the T variable have independent values that have unknown "F" distribution. Because of censored data C variable is called censor variable. Let"δ" be the failure indicator. In addiation to this, explanatory variables be the k-ordered vector. This vector have to satisfy some assumptions.

1. T and C have to be independent
2. $P(T_i \leq C_i|T_i, X_i) = P(T_i \leq C_i|T_i), i = 1,2,3, \ldots . n$

The first condition is ordinary censored model condition. And second one has been told that in given time the model cannot provide more information unless there is no censorship.

That second assumption is more weak assumption than the assumption that C and X have to be independent. The model that have a dependent variable and logarithm of time and have explanatory variables is like:

$$lnT_i = \alpha + X_i\beta + \varepsilon_i \quad (1)$$

Residuals have identity independent distribution that zero mean and constant variance. Because of the model includes censored data, Miller(1976), suggested that Kaplan-Meier estimator for minimize the sum of squares of residuals [7].

3.1 Kaplan-Meier Estimator

This estimator have an another name called product-limit estimator. Nonparametric statistical methods use this for estimate to survival function. In medical studies, this method generally use for determine intervals for patients life times.

3.1.1 Estimation of Kaplan-Meier Function And Weights

Kaplan and Meier(1958) have improved the product-limit estimator for the estimation of survival functions. This method appropriate for small or large sample sizes [5].

Basic point in Kaplan-Meier; method is, calculation a new probability for each event of each time. At each step, determined event's(death, quit etc.) probability will increase or not change. Consequently, Kaplan-Meier graph will be a curve like stairs. Some researchers have told the method is useful for great analysis [7].

$$S(t) = (n - d)/n \quad (2)$$

$(n - d)$: is number of individuals in time point *"t"* and greater than *"t"*

n: Total number of Individuals

If the censored data exist, then for the Calculation of the Kaplan-Meier estimator time is divided in intervals. But, there is a key part in here. That is each interval have to include least one interested event. So if the estimation of Kaplan-Meier is doing the way that explained above, there are estimation of each individuals and censored data in Kaplan-Meier curve. It illustrated in Figure 1. All of this estimations are used in regression calculations as a weight matrix. This matrix is a diagonal and square matrix that have only diagonal elements in it.

4 Conclusion

The end of the study, A linear model is acquainted that is taken an account censored data. This model is formed according to parametric assumptions so that except some datasets but in many of the studies estimations are inconsistent and may be entirely wrong. For solving this problem, there are lots of methods have been suggested by Cox, Miller, Buckley and James. Also Semi parametric and nonparametric methods can be used for that.

5 References

[1] AYDIN, D., Estimations And Inferences In Semi parametric Regression Modelling Using Smoothing Spline Approach, Anadolu Üniversitesi İstatistik Bölümü(2005)

[2] CHEN S., KHAN S., Semi parametric Estimation Of a Partially Linear Model, Econometric Theory, 567-590(2001)

[3] ENGLE, R. F., GRANGER, C. W., RICE, J., WEISS A., Semi parametric Estimates Of The Relation Between Weather and Electricity Sales, Journal Of The American Statistical Association, 310-320(1986)

[4] HARDLE, W., LIANG, H., Large Sample Theory Of The Estimation Of The Error Distribution For Parameters In a Semi Parametric Regression Models, Communications In Statistics, Theory And Methods(1999)

[5] KOUL, H., SUSARLA, V., VAN RYZIN, J., Regression Analysis With Randomly Right-Censored Data, Journal Of The Annals Of Statistics , 1276-1285(1981)

[6] LAI, T. L., And YING, Z. L., Asymptotically Efficient Estimation In Censored And Truncated Regression Models, Statistica Sinica, 2, 17-46(1992)

[7] MILLER, R., HALPERN, J. Regression With Censored Data, Biometrika, 521-531(2011)

[8] SHICK, A., On Asymptotically Efficient Estimation In Semi parametric Model, Annals In Statistics, 14, 1139-1151(1986)