이학석사 학위논문

Development of High-value Traits of Dairy Cattle using Frailty Model

Frailty Model을 이용한 고부가 젖소 형질 개발

2014년 2월

서울대학교 대학원

통계학과

이 슬 기

Development of High-value Traits of Dairy Cattle using Frailty Model

지도교수 김 용 대

이 논문을 이학석사 학위논문으로 제출함

2013년 10월

서울대학교 대학원

통계학과

이 슬 기

이슬기의 이학석사 학위논문을 인준함

2013년 12월

위 원 장	장 원 철	(인)
부위원장	김 용 대	(인)
위 원	임 요 한	(인)

Development of High-value Traits of Dairy Cattle using Frailty Model

By

Seulgi Lee

A Thesis
Submitted in fulfillment of the requirements
for the degree of
Master of Science
in Statistics

The Department of Statistics
College of Natural Sciences
Seoul National University
February, 2014

Abstract

In order to increase profit on dairy industry in Korea, reducing milk production costs can be a good way. In this study, we will develop high-value traits of dairy cattle using frailty model with data collected from Rural Development Administration in Korea. Survival analysis is used as there are censored observations. In particular, we use frailty model to consider the correlation between dairy cattle in the same farm. Since general variable selection criterions are not applied in the semi-parametric model like frailty model, we propose a criterion for variable selection. We evaluate the performance of the proposed criterion by simulation. And, we construct a model using the criterion and predict the economic life of dairy cattle. We can think that dairy cattle whose predicted economic life will be long have a high-value traits. Finally, we look at the increase rate of the economic life of dairy cattle when selecting dairy cattle with a high-value traits.

Keywords: Survival Analysis, Frailty model, semi-parametric model, variable selection, High value traits of dairy cattle.

Student Number: 2012-20228

Contents

List of Tables

List of Figures

Chapter 1

Introduction

In order to increase profit on dairy industry in Korea, we can think several ways. Among that, reducing milk production costs can be a good way. To reduce the costs, it will be important that extending the economic life of the dairy cattle more than anything else. Low economic life leads to a deterioration of the profitability of dairy farmers due to increase of the number of heads in order to maintain the proper milking cows. In addition, it can be a reduction factor of incomes. In case of well-developed countries in the dairy industry such as Europe and North America, basic research on the economic life is well carried already.

Ducrocq (1994) found a meaningful factor to affect on the economic life through a survival analysis having a baseline Weibull hazard function using 10 million Normande cows from 1979 to 1989. Ducrocq (1997) developed a method of Ducrocq (1994) by applying time-varying explanatory variables to proportional hazard model(Cox, 1972). Roxstrom et al. (2003) proposed a new model of longevity of dairy cattle on a lactation basis, it was shown that the performance is better than the model using period from first childbirth to death as economic life. Sewalem et al. (2004) conducted a survival analysis using the Weibull proportional hazards model for 1 million Holstein species in Canada, as in Ducrocq (1994). Sewalem et al. (2005) analyzed the period between the lactation using Jersey species and Ayrshire species similar to Roxstrom et al. (2003). Sewalem et al. (2006a), Sewalem et al. (2006b) and Sewalem et al. (2008) investigated the effect on the economic life of somatic cell score, inbreeding, reproductive traits, respectively.

In case of well-developed countries in the dairy industry such as Europe and North America, they have set the improvement indicators for the extension of the economic life of dairy cattle. And, they procured the genetic evaluation technique for economic life of dairy cattle and utilize it. In case of Japan, since the mid-2000s, they are trying to extend the economic life with a goal for growing improvement of livestock.

In this study, we investigate the various factors affecting on the survival rate of dairy cattle using survival analysis of the data from birth to death. The data is collected from the Rural Development Administration in Korea. We grasp the influence of each variable on the economic life using frailty model. Based on this, we find out the various factors that affect the economic life and propose a criterion to select dairy cattle having good economic life. In addition, we propose a method to evaluate the traits of individual cows on the basis of the developed model and to verify it.

This paper is organized as follows. In chapter 2, we describe the data used in this paper. In chapter 3, we review gamma frailty model and propose a new criterion for variable selection. And, we evaluate the performances of the proposed criterion by simulation. In chapter 4, we construct a frailty model using the criterion proposed and evaluate model performance. Finally, the conclusion is given in chapter 5.

Chapter 2

Data Description

The data set used in this paper was collected from Rural Development Administration in Korea. In this paper, our purpose is developing high-value traits of dairy cattle and looking for a method to increase the economic life of dairy cattle. We define the economic life of dairy cattle as two ways, a calving number until death and a month age from first childbirth to death. In case that dairy cattle are alive at a investigation time, the economic life of dairy cattle is defined as a calving number until the investigation time and a month age from first childbirth to the investigation time. And, these dairy cattle are treated as censored data. Milk production will increase when the calving number is high or the month age is long. Hence, dairy cattle with high calving number or long month age will have a good economic value.

Input attributes to explain the economic life of dairy cattle consist of 3 groups : conformation score, characteristics related milking and breeding value. Conformation Score is given based on some criteria, between 1 and 9. Characteristics related milking is a value indicating a quality of milking. And, breeding value is related to a hereditary ability. All input attributes are summarized in

Table 2.1.

The original data consists of $645,672$ dairy cattle characteristics. Of these, dairy cattle having explanatory variables are only used in analysis. Outliers were deleted after talking with expert. Finally, $23,080$ dairy cattle are used in analysis.

A Censoring probability per dairy cattle's calving number is in Table 2.2 and Figure 2.1. As seen Table 2.2 and Figure 2.1, censoring probability increases as dairy cattle's calving number increases. Hence, we should consider the censoring information. To analyze such a censored data set, we use Survival Analysis. Furthermore, a cow may has a correlation with other cows in the same farm. We will use a gamma frailty model not to ignore such a correlation.

Table 2.1: Input Attributes

Group	Traits
Conformation	STA, CWI, BD, ANG, RAN, RWI, RLSS, FAN, FUA, RUH, RUW, USU, UDE, FTP, FTL, RLRV
Milk	305day milk yield, 305day fat yield, 305day protein yield, 305day snf yield, mature milk yield, mature fat yield, mature protein yield, mature snf yield
Breeding	milk yield, fat yield, protein yield, somatic cell score, STA, CWI, BD, ANG, RAN, RWI, RLSS, FAN, FUA, RUH, RUW, USU, UDE, FTP, FTL, OCS, LST, UTX, RTP, HDE, BQL, RLRV, HHE, LOC, BCS

[Abbreviation]

STA(Status), CWI(chest width), BD(body depth), ANG(angularity), RAN(rump angle), RWI(rump width), RLSS(rear leg set), FAN(foot angle), FUA(fore udder attachment), RUH(rear udder height), RUW(rear udder width), USU(udder support), UDE(udder depth), FTP(front teat placement), FTL(front teat length), RLRV(rear leg rear view), OCS(overall conformation score), LST(loin strength), UTX(udder texture), RTP(rear teat placement), HDE(heel depth/foof height), BQL(bone quality), HHE(hfe height), LOC(locomation), BCS(Body condition score)

Table 2.2: Censoring probability per calving number

Calving number	1	2	3	4	5	6	7
Censoring probability(%)	0.04	0.06	21.55	45.85	50.04	54.69	75.52

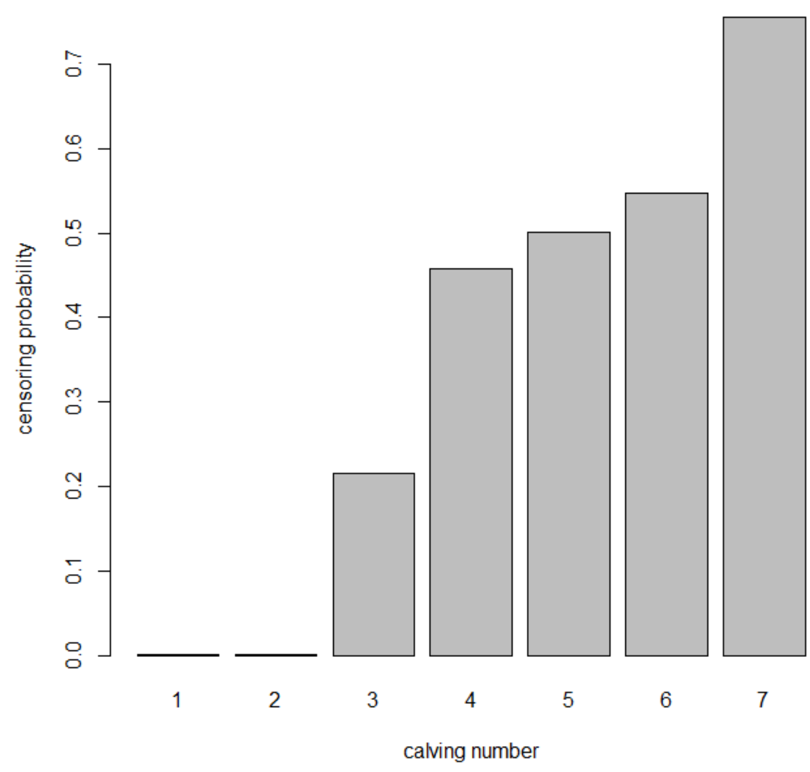

Figure 2.1: Censoring probability per calving number

Chapter 3

Methodology

3.1 Gamma Frailty Model

A frailty model is useful for modeling association between individual survival times within subgroups. The most common model for frailty is the shared frailty model. In this model, all the units within each category share a common frailty, each unit belongs to precisely one category, and frailties of different categories are independent.

Let T_i, $i = 1, ..., n_j$, $j = 1, ..., G$ be the observed times. Assume that the data for subject i, who is a member of the jth of G groups, follows a proportional hazards shared frailty model. The hazard can be written as

$$\lambda_i(t) = \lambda_0(t) u_j \exp(\mathbf{X}_i \beta), \tag{3.1}$$

where u_j is the frailty of group j, $\mathbf{X}_i$ is the covariate matrix of $\imath$th subject, dimension 1 by p, and β is a vector of regression coefficients. And (3.1) can be represented as

$$\lambda_i(t) = \lambda_0(t) \exp(\mathbf{X}_i \beta + \mathbf{Z}_i u), \tag{3.2}$$

where $\mathbf{Z}$ is a matrix of G indicator variables such that $Z_{ij} = 1$ when subject i is a member of family j and 0 otherwise, and each individual belongs to only one family.

To estimate β and u, we should maximize a penalized partial log-likelihood for model (3.2), which is given by

$$PPL = l_{partial}(\beta, u) - g(u; \theta). \tag{3.3}$$

Here $l_{partial}(\beta, u)$ is the log of the usual Cox partial likelihood,

$$l_{partial}(\beta, u) = \sum_{i=1}^{n} \int_0^\infty \left[Y_i(t)(\mathbf{X}_i\beta + \mathbf{Z}_i u) - \log \left\{ \sum_k Y_k(t) \exp(\mathbf{X}_k\beta + \mathbf{Z}_k u) \right\} \right] dN_i(t) \tag{3.4}$$

and g is a penalty function.

In gamma frailty model, it is assumed that each e^{u_j} has i.i.d. gamma distribution with mean 1 and a common unknown variance θ; the probability density function is thus :

$$g(u; \theta) = -\frac{1}{\theta} \sum_{i=1}^{G} (u_j - e^{u_j}). \tag{3.5}$$

So PPL is given by

$$PPL = l_{partial}(\beta, u) + \frac{1}{\theta} \sum_{i=1}^{G} (u_j - e^{u_j}). \tag{3.6}$$

The profile marginal log-likelihood of this shared frailty model is given by

$$l_m = PPL + \sum_{j=1}^{G} \nu - (\nu + d_j) \log(\nu + d_j) + \nu \log \nu + \log \left(\frac{\Gamma(\nu + d_j)}{\Gamma(\nu)} \right), \tag{3.7}$$

where d_j is the number of events in the jth family and $\nu = 1/\theta$.

The fitting program for a shared gamma frailty model consists of two step. First, for any fixed θ, Newton-Raphson algorithm is used to solve the penalized model, and return the corresponding value of the PPL. Next, θ is chosen to maximize the profile likelihood in Equation (3.7).

3.2 Variable Selection for Frailty Model

To find out high-value traits of dairy cattle, we should select important variables. In parametric model, we can use various criterions such as Mallow's Cp, AIC and BIC for variable selection. However, such methods are not available in semi-parametric model like frailty model. In this paper, we will propose new variable selection criterions like AIC and BIC. Proposed variable selection criterions are defined as

$$\texttt{s\_AIC} = -2l_m(\theta) + 2p \tag{3.8}$$

$$\texttt{s\_BIC} = -2l_m(\theta) + p \log n. \tag{3.9}$$

3.3 Simulation

This simulation study looks at the performances of the criterions (3.8) and (3.9). We simulated 100 data sets consisting of G=40, $n_j = 5$ for all j. We let $p = 4$ and the $x_{i1}, ...x_{i4}$ are independently generated from $Pr(x_{ij} = 0) = Pr(x_{ij} = 1) = 1/2$. For the true regression coefficients and baseline hazard function, we set two situations. First, we let $\beta = (1, -0.5, 0, 0)$ and $\lambda_0(t) = 1$. Second, we let $\beta = (2, -1, 0, 0)$ and $\lambda_0(t) = 1$. The distribution of the censoring time is an exponential distribution with mean 1. We consider the shared gamma frailty model for the frailty term with $\theta = 0.5$.

We consider a sequence of nested submodels $\mathcal{P}_1 \subset \mathcal{P}_{12} \subset \mathcal{P}_{123} \subset \mathcal{P}_{1234}$, where $\mathcal{P}_I$ denotes the model using only covariates whose indices belong to I. For each simulated data, we apply the s_AIC and the s_BIC to select the submodel among $\mathcal{P}_1, \mathcal{P}_{12}, \mathcal{P}_{123}$ and $\mathcal{P}_{1234}$.

Results from the simulation can be seen in Table 3.1. The frequencies of the selected model by the criterion s_AIC and s_BIC in each situation are

Table 3.1: Simulation Result

true coefficients	criterion	Selected models			
		$\mathcal{P}_1$	$\mathcal{P}_{12}$	$\mathcal{P}_{123}$	$\mathcal{P}_{1234}$
$(2, -1, 0, 0)$	`s_AIC`	0	76	20	4
	`s_BIC`	2	95	3	0
$(1, -0.5, 0, 0)$	`s_AIC`	20	55	19	6
	`s_BIC`	59	40	1	0

presented. By increasing the variation of coefficients, the frequency of selecting the true model, $\mathcal{P}_{12}$, increase. And the `s_BIC` tends to select smaller models than `s_AIC`, which is similar to the relation between AIC and BIC in parametric model. We can see that `s_AIC` and `s_BIC` are work as AIC and BIC, so the `s_AIC` and `s_BIC` can be used for selecting semi-parametric model, shared gamma frailty model. In this paper, we will use only `s_AIC`.

Chapter 4

Data Analysis

4.1 Variable Selection

All variables in conformation and breeding group of Table 2.1 were used to predict the economic life of dairy cattle. As seen in Table 4.1, there are high correlations between variables in milking group of Table 2.1. So we will use only one variable, the mean of standardized variables in milking group, instead of all variables in milking group.

We split the data into two groups : training set and test set. We select a model using forward selection method in training set. For selecting variables, we use `s_AIC` in (3.8). The results of variable selection are summarized in Table 4.2. The variables are sorted by selection order.

As seen in Table 4.2, somatic cell score is the most effective variable on increasing the economic life of dairy cattle, both month age and calving number. We can also see that a lot of variables are overlapped in two economic life. Thus, we can effectively develop high-value traits of dairy cattle using only overlapping variables such as somatic cell score, rear leg rear view, the mean of

Table 4.1: The Correlation between Variables in Milking Group

	MILK 305	FAT 305	PROT 305	SNF 305	MAT MILK	MAT FAT	MAT PROT	MAT SNF
MILK305	1.00	0.75	0.95	0.98	0.99	0.74	0.94	0.97
FAT305	0.75	1.00	0.79	0.78	0.75	0.99	0.79	0.77
PROT305	0.95	0.79	1.00	0.97	0.94	0.78	0.99	0.97
SNF305	0.98	0.78	0.97	1.00	0.98	0.77	0.97	0.99
MATMILK	0.99	0.75	0.94	0.98	1.00	0.76	0.95	0.98
MATFAT	0.74	0.99	0.78	0.77	0.76	1.00	0.79	0.78
MATPROT	0.94	0.79	0.99	0.97	0.95	0.79	1.00	0.97
MATSNF	0.97	0.77	0.97	0.99	0.98	0.78	0.97	1.00

standardized variables in milking group, breeding value related to rump angle and udder depth.

Table 4.2: The Result of Variable Selection

month age			calving number		
Variable	Estimate	p-value	Variable	Estimate	p-value
somatic cell score	72.81	<0.001	RLRV	31.92	<0.001
Breeding_CWI	0	0.960	somatic cell score	73.43	<0.001
fat yield	17.16	<0.001	Breeding_HHE	2.71	0.100
FUA	10.16	0.001	FUA	11.52	<0.001
s.milking	22.56	<0.001	BD	9.55	0.002
BD	11.83	<0.001	s.milking	18.90	<0.001
Breeding_RAN	16.34	<0.001	Breeding_RWI	5.77	0.016
RLRV	7.58	0.006	Breeding_RAN	19.15	<0.001
Breeding_RWI	6.72	0.010	Breeding_UDE	14.57	<0.001
Breeding_UDE	14.88	<0.001	DF	8.10	0.004
Breeding_STA	8.67	0.003	fat yield	9.70	0.002
DF	7.26	0.007	Breeding_STA	8.85	0.003
RUW	8.46	0.004	FTP	5.73	0.017
FTP	6.94	0.008	RUW	7.39	0.007
Breeding_LOC	10.61	0.001	RUH	7.65	0.006
Breeding_USU	8.69	0.003	Breeding_RUH	8.43	0.004
Breeding_FAN	3.65	0.056	Breeding_BQL	4.35	0.037
Breeding_RUH	6.95	0.008	Breeding_LOC	4.61	0.032
FA	3.63	0.057	Breeding_USU	5.45	0.020
RUH	3.93	0.048	Breeding_FTP	3.14	0.076
Breeding_FTP	2.66	0.100	Breeding_LST	2.05	0.150
Breeding_BQL	2.46	0.120			
Breeding_RLRR	2.00	0.160			

4.2 Model Performance

To evaluate model performance, economic life of dairy cattle was predicted in test set. A dairy cattle with high score in effective input attributes can be shown that it has high value traits. So we evaluated the traits of dairy cattle with predicted economic life. We ranked the dairy cattle in order of predicted economic life. We calculated medians and means of the true economic life of the dairy cattle, whose predicted economic life is in the top $n\%$, where $n = 10, 20, ..., 100$. The results are presented in Table 4.3. Comparing to the median and mean of economic life of all dairy cattle, increase rates of median and mean in the top $n\%$ are also summarized in Table 4.3.

Median is more reasonable statistic than mean in survival analysis. As seen in Table 4.3, median of monthage of dairy cattle in top 10% increases about 16% comparing to the median of monthage of all dairy cattle. In addition, we can increase median of calving number about 20% comparing to the median of calving number of all dairy cattle by selecting top 10% dairy cattle. Also, increase rates of median grow steadily as increasing predicted economic life, that is, traits score. Increase rates of mean have a tendency like median, except for dairy cattle in top 30% of calving number. We conjectured that the exception will be because of outliers.

Table 4.3: Increase Rate of Median and Mean by Using Constructed Model

	Top%	Median	Increment	Mean	Increment
Month age	10	43.09	15.93	44.62	10.79
	20	41.49	11.64	44.45	10.36
	30	40.97	10.23	44.41	10.27
	40	40.64	9.34	43.72	8.56
	50	40.37	8.62	43.33	7.59
	60	39.93	7.43	42.81	6.29
	70	39.31	5.77	42.3	5.04
	80	38.61	3.88	41.63	3.38
	90	38.07	2.44	41.06	1.97
	100	37.17	0	40.27	0
Calving number	10	3.25	19.84	3.49	12.21
	20	3.15	16.26	3.44	10.59
	30	3.1	14.36	3.58	15.01
	40	3.01	11.1	3.5	12.47
	50	2.97	9.56	3.44	10.51
	60	2.9	6.85	3.36	8.07
	70	2.85	5.16	3.3	6.28
	80	2.81	3.72	3.24	4.09
	90	2.78	2.38	3.18	2.37
	100	2.71	0	3.11	0

Chapter 5

Conclusion

In this paper, we developed high-value traits of dairy cattle. For that, we used the frailty model to contain the correlation between dairy cattle in the same farm. And, to select variables, we proposed variable selection criterion and showed the rationality by the simulation. Using this, we found that the effective attributes that influence the economic life of dairy cattle are somatic cell score, rear leg rear view, the mean of standardized variables in milking group, breeding value related to rump angle, udder depth and so on. Actually, a dairy cattle with good score in these variable has long economic life. If we select dairy cattle whose predicted economic life is in the top 10%, we can increase the economic life up to about 16%. However, to increase the economic life of dairy cattle as much as a advanced country of dairy industry, we need to apply more advanced model or find out more effective attribute. It is expected that an environmental factor of Korea having distinct 4 seasons may affect the life of dairy cattle. In further study, if these environmental factors are considered, it is expected that we will developed more efficient model for increasing the life of dairy cattle.

Bibliography

Ducrocq, V. (1994). Statistical analysis of length of productive life for dairy cows of the Normande breed. *Journal of Dairy Science*, **77**, 855–866.

Ducrocq, V. (1997). Survival analysis, a statistical tool for longevity data, in: 48th Annual Meeting of the European Association for Animal Production, 25–28 August 1997, Vienna, Austria.

Fan, J. and Li, R. (2002). Variable Selection for Cox's Proportional Hazards Model and Frailty Model. *The Annals of Statistics*, **1**, 74–99.

Gutierrez, R.G. (2002). Parametric frailty and shared frailty survival models. *The Stata Journal*, **2**, 22–4.4

Hougaard, P. (1995). Frailty Models for Survival Data. *Lifetime Data Analysis*, **1**, 255–273.

Jeong, K., Chae, M., Lee, S., Cho, K. and Kim, Y. (2013). Development of High-value Traits of Dairy Cattle using Survival Analysis. *Journal of the Korean Data Analysis Society*, **15**, 2407–2416.

Klein J.P. and Moeschberger M.L. (2003). *Survival Analysis.* Springer.

Ripatti, S. and Palmgren, J. (2000). Estimation of Multivariate Frailty Models Using Penalized Partial Likelihood. *Biometrics*, **56**, 1014–1002.

Roxstrom, A., Ducrocq, V. and Strandberg, E. (2003). Survival analysis of longevity in dairy cattle on a lactation basis, *Genetics Selection Evolution*, **3**, 305–318.

Sewalem, A., Kistemaker, G. J., Miglior, F. and Van Doormaal, B. J. (2004). Analysis of the relationship between type traits and functional survival in Canadian Holsteins using a Weibull proportional hazards model, *Journal of Dairy Science*, **11**, 3938–3946.

Sewalem, A., Kistemaker, G. J., Ducrocq, V. and Van Doormaal, B. J. (2005a). Genetic analysis of herd life in Canadian dairy cattle on a lactation basis using a Weibull proportional hazards model, *Journal of Dairy Science*, **1**, 368–375.

Sewalem, A., Miglior, F., Kistemaker, G. J. and Van Doormaal, B. J. (2006a). Analysis of the relationship between somatic cell score and functional longevity in Canadian dairy cattle, *Journal of Dairy Science*, **9**, 3609–3614.

Sewalem, A., Kistemaker, G. J., Miglior, F. and Van Doormaal, B. J. (2006b). Analysis of inbreeding and its relationship with functional longevity in Canadian dairy cattle, *Journal of Dairy Science*, **6**, 2210–2216.

Sewalem, A., Miglior, F., Kistemaker, G. J., Sullivan, P. and Van Doormaal, B. J. (2008). Relationship between reproduction traits and functional longevity in Canadian dairy cattle, *Journal of Dairy Science*, **4**, 1660–1668.

Therneau, T.M., Grambsch, P.M. and Pankratz, V.S. (2003). Penalized Survival Models and Frailty. *Journal of Computational and Graphical Statistics*, **12**, 159–175.

국문초록

한국의 낙농산업에서 더 많은 수익을 창출하기 위해서는 우유 생산비 절감이 좋은 방안으로 제시될 수 있다. 본 연구에서는 농촌진흥청에서 조사된 자료를 이용하여 젖소의 고부가 형질을 개발할 것이다. 중도 절단된 자료들을 분석하기 위하여 생존 분석이 본 연구에서 사용되며, 특히 같은 농장의 젖소들 사이에 존재하는 상관관계를 고려하기 위하여 frailty model이 사용될 것이다. 일반적인 변수선택 기준이 frailty model과 같은 semi-parametric 모형에서는 적용되지 않아 새로운 변수선택 기준을 제시하고 시뮬레이션을 통해 제시된 변수선택 기준의 성능을 확인한다. 또한, 이를 사용하여 모형을 구축하고 젖소의 경제수명을 예측한다. 예측된 경제수명은 젖소의 형질점수와 비례하기 때문에 긴 경제수명을 가질 것이라고 예측된 젖소의 형질점수가 높다고 생각할 수 있다. 최종적으로 높은 형질점수를 가진 젖소를 선택하였을 때 젖소의 경제수명 증가율을 살펴본다.

주요어: 생존 분석, frailty model, 준모수 모형, 변수 선택, 고부가 젖소 형질.

학 번: 2012-20228