EAAP 2008, Session 10

# Genetic evaluation for days-open in Danish Holstein using different models

Guosheng Su, Yali Hou, Per Madsen, Mogens S. Lund

University of Aarhus, Faculty of Agricultural Sciences, Denmark

Guosheng.su@agrsci.dk

# Challenge for statistical model

- Censored records
- Non-Normal distribution

## Objective:

Evaluate models for genetic evaluation of days-open (**DO**).

# Data:

Lactation: First lactation.

Period: 1995 to 2004.

Herd: Having records in all the 10 years

Herd-year: Minimum 5 records

Sire: Minimum 5 records

In total: 476,000 records

# Definition of censored records:

Unknown date of conception: **DO** is calculated as days from calv. to last insem., censored

**DO**> 365: replaced with 365, censored

16.6% censored records

# Statistical models (5 alternative models)

1. **Conventional linear model (LM):** Add 21 d to censored records

2. **Threshold-linear model (TLM):** A threshold model for censoring status (0=CS, 1=UNCS) and a linear model for **DO**.

3. **Right censored linear Gaussian model (CLM)**

4. **Weibull proportional hazard model (SMW):** Ducrocq and Casella, 1996

5. **Cox proportional hazard model (SMC)**: Piecewise constant baseline hazard function with constant length of 21 d

# Statistical models (5 alternative models)

## Basic model

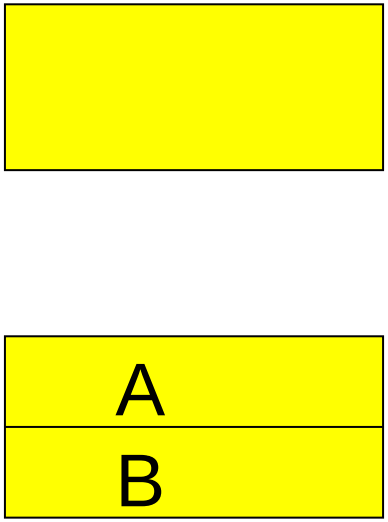
$$
\begin{aligned}
Y = \;& \text{Year\_month} \\
& + \text{Herd\_year} \\
& + \text{Age\_group} \\
& + b_b.\text{Breed\_prop} \\
& + b_h.\text{Heterozygosity} \\
& + \text{Sire} \\
& + \text{residual}
\end{aligned}
$$

# Model validation

## Datasets for model validation

The whole data:

Subset A and B:



A

B

(divided by herd)

# Validation criteria

1. Cor($EBV_A$, $EBV_B$): Test model stability

2. $X^2$ based on cross validation: Test predictive ability

# Procedure of cross validation

1). **DO** → 5 intervals (<66, 67-95, 96-130,131-188, >188)

2). Calculate daughters frequency in each interval for each sire

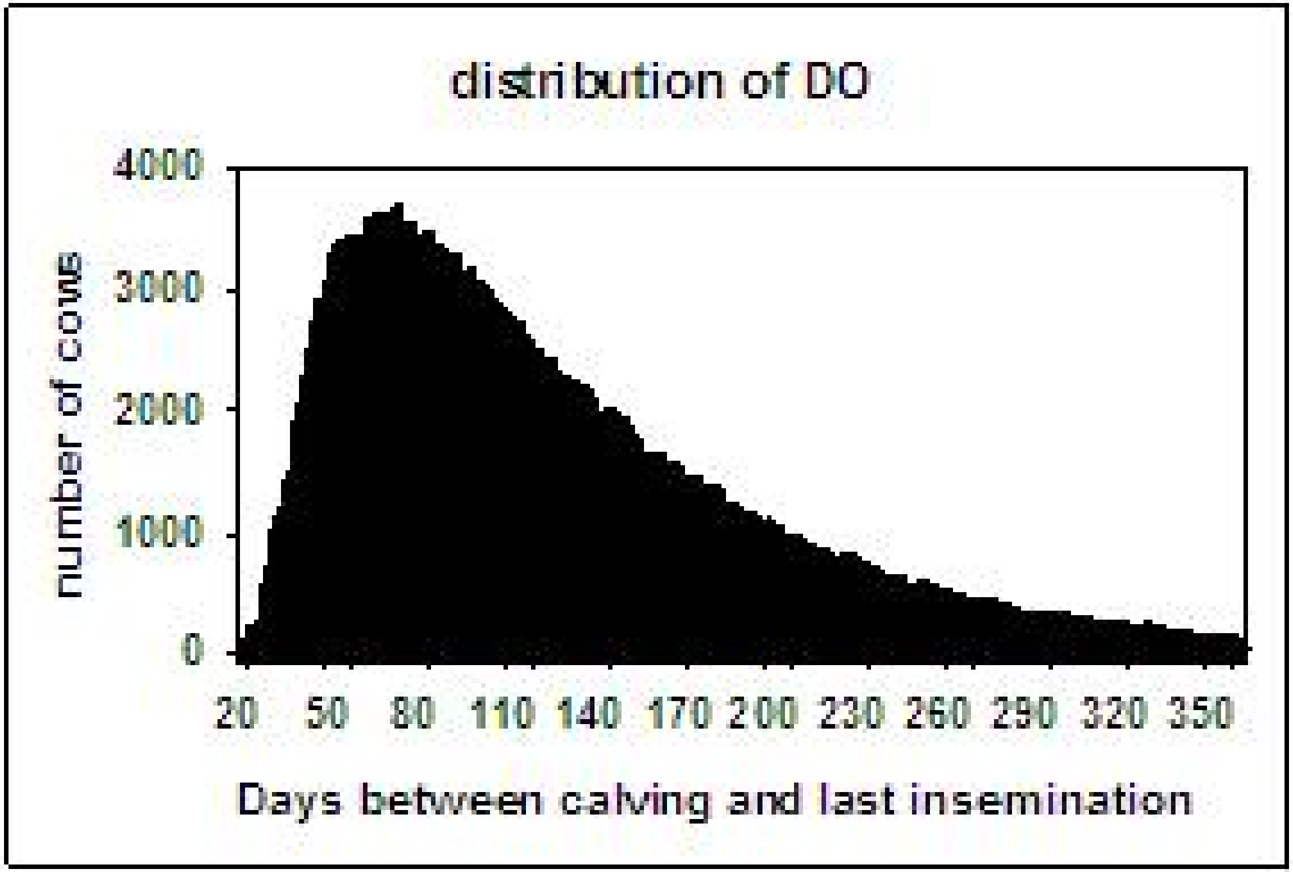
3). Estimate daughters probability of conception in each interval, using logistic regression on EBV, based on dataset A

4). Predict daughters frequency in dataset B, using the probability from dataset A

5) Calculate $X^2$ statistic

$$x^2 = \sum_{i=1}^{N} \sum_{j=1}^{5} \frac{(E_{ij} - O_{ij})^2}{E_{ij}}$$

# Results



Mean=132.0

STD=78.4

# **Table 1**. Spearman rank correlation between EBV from different models (EBV$_{time}$ in LM, TLM and CLM, EBV$_{hazard}$ in SMW and SMC)

| Model | TLM | CLM | SMW | SMC |
|---|---|---|---|---|
| LM | 0.997 | 0.983 | -0.906 | -0.826 |
| TLM |  | 0.970 | -0.891 | -0.817 |
| CLM |  |  | -0.930 | -0.826 |
| SMW |  |  |  | 0.661 |

- Different models could result in different ranking

# **Table 2.** Spearman rank correlation between EBV from subset A and subset B

| Dataset | LM | TLM | CLM | SMW | SMC |
|---|---|---|---|---|---|
| A - B | 0.620 | 0.624 | 0.594 | 0.384 | 0.841 |

- SMC is the best in stability

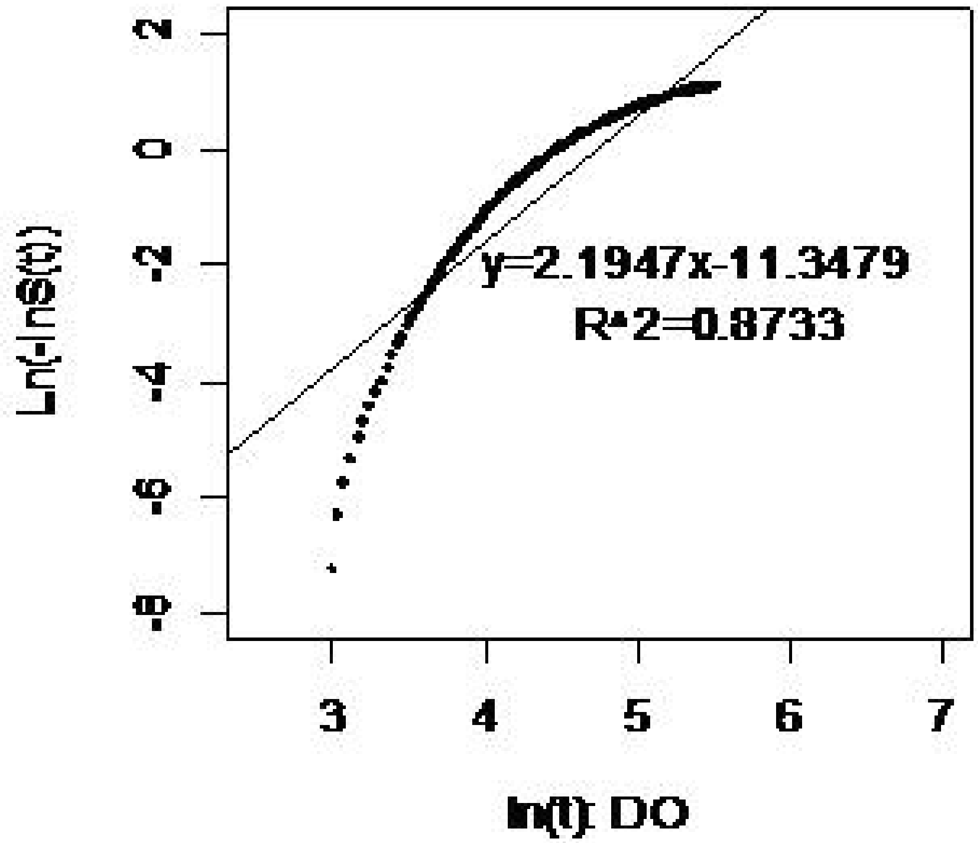
## **Table 3.** $x^2$ statistics for the sires with at least 20 daughters, calculated from the expected and observed frequency of daughters getting conception in five intervals

| Cross validation | LM | TLM | CLM | SMW | SMC |
|---|---|---|---|---|---|
| A → B | 5920 | 5876 | 6055 | 6407 | 5750 |
| B → A | 6109 | 6091 | 6135 | 6478 | 5885 |

- SMC shows best prediction ability

**[We have found a mistake in analysis of $X^2$ statistic after EAAP, the figures in this table is waiting to be verified !!!]**

# Why did Weibull model (SMW) not perform well?



- Weibull function does not fit the data

**Figure 2.** Plot of ln[-ln*S(t)*] against ln(*t*). *S(t)* = Kaplan-Meier estimates of the survival function at time t

# Conclusions

1. Genetic evaluation of **DO** using different models could result in different ranking of candidates

2. Cox proportional Hazard model (SMC) is a good alternative to genetic evaluation of **DO**.