Homework

Lluis Ramon

20 de abril de 2015

1 Introduction

Cattle are an important economic resource and also a major health factor in many countries. Therefore preventing diseases in a herd is vital.

In this study we focus our interest on the parasite of trypanosomosis which can lead to the death of a cow. This disease, transmitted by the tsetse flies, causes an infection characterized by fever, loss of appetite and anemia, which can lead to death depending on different factors.

One medicine, the Berenil, is used to cure the infected cattle. The aim of this research is to determine the efficiency of different doses of Berenil. Finding the most efficient dose, if such a dose exists, is critical when it comes to save both cattle and money. Here lies all the interest of our study.

To determine if a cow is ill or not a binary variable will be studied. This indicator can be 1 for a healthy cow or 0 for a sick one.

2 Objective

The aim of this study is to assess the efficacy of different doses of Berenil in cattle infected with the trypanosomosis parasite.

3 Dataset

A cohort of 10 different cows infected by trypanosomosis parasite was selected for the study. Each Berenil dose (low, medium and high) was administrated three times (time 1, 2 or 3) for each animal. PCV was reported each time as well as the number of calves it had before being infected.

The variables reported for this study are presented below.

- Id: Each cow has its own id. From 1 to 10. *(id)*
- PCV: Binary variable. *(pcv.b)*
- Dose: H High, M Medium L low. *(dose)*
- Time: From 1 to 3. *(time)*
- Number of birth: From 2 to 8. *(nbirth)*

Since the gathering process of the data was unknown, several assumptions were needed.

- Each time PCV is obtained before the treatment. Therefore the effect of the third dose could not be evaluated.
- Dose is assigned randomly in time to the cow.
- For a given cow, the previous treatments (high, medium or low) do not affect the following ones.
- Time intervals are the same and fixed.

4 Outcome categorization

TODO(Mathieu): Improve text

The binary response was related with a cow being healthy or unhealthy. Taking this into considertation, the outcome was categorized using the following criteria:

- Literature review*: A healthy cow is estimated to have a PCV value ranging from 24 to 46.
- Practical Modeling: As the binary response should be modeled in following sections, a suitable one was searched. To this end, a trial and error with a cutoff ranging from 20 to 24 was explored.

The threshold between healthy and unhelathy cow was set at a PCV value of 20. If the PCV value was bigger it was categorized as healthy, if it was lower or equal it was cathegorized as unhelathy.

Table 1 shows a contingence table for the dichotomised response variable. Healthy category increases in time while unhealthy category decreases. A Missing value category for healthy/unhealthy cow i is also included. It can be seen that the number of missng values increases with time. This topic will be studied in further detail in section 7 Missing Data Analysis.

	Time 1	Time 2	Time 3
Unhealthy	28	15	2
Healthy	1	12	18
Missing value	1	3	10

Table 1: Contingence table for the dichotomised response variable (rows) and times.

5 Statistical methods

The methods used in the statistical analyses are detailed in this section. First a Generalized Estimating Equations (GEE) and latter a Generalized Linear Mixed Model. An exploratory data analysis was performed before those regression methods.

5.1 GEE

The response variable was the binary PCV to detect healthy and unhelathy cows. As we wanted to estimate the effect of dose in healthy/unhelathy cows, these two variables were included in the model.

From this initial model, a forward step-wise method was carried for the model selection, including additional covariates or interactions between them. The models were compared quasilikelihood ratio test when nested and with QuasiLikelihood Information Criteria (Pan 2001) when non-nested.

A classification table with predicted values and original data was created to asses how the model performed. Sensitivity and specificity are calculated for the final model.

5.2 GLMM

TODO(Mathieu)

IN order to take into account the specific effect due to the cow itself we had to introduce random effects in our model. To do that we performed a Generalised Linear Mixed Model. We kept the fixed effects found in the previous part and add random effects on the intercept and the time. Since we have two way of grouping our data we have a random effect for the id and among the values of a given cow we have random effects for the kind of dose that was used.

5.3 Transition Models

TODO(Lluis): Add model information. TODO(All): Include pros/cons with our data. Decide if our data could be modeled using this kinds of models.

6 Results

6.1 GEE

TODO(Mathieu): Revise text.

As explained in the section 5.1, a starting model for PCV was fitted with dose. Following a step-wise aproach time was included in the model. The number of births and its interaction with time or dose did not imporved the model. Neither it did to include the interaction between time and dose.

The final model was:

$$logit(p) = \beta_0 + \beta_1 time + \beta_2 doseMedium + \beta_3 doseHigh$$

Several working correlation matrices were used and an AR1($\alpha = 0.171$) was chosen because of its smaller QIC.

Table X shows the coefficient estimate for the final model. The odds ratio of having a high dosage is $exp(3.382)$ times of having Low dosage adjusted for time covariate. An increase of one unit in time increases the odds ratio by $exp(3.205)$ times adjusted for dose covariate. Even medium dose was not significant in the model, we keept with it in the model because High dose was significant. As part of a further research 8 is let the models with Low and Medium dose categories merged as a single category.

TODO(Gerard): Create table with coefficients and alpha

```
##
## Call:
## geeglm(formula = pcv.b ~ dose + time, family = binomial, data = cows.com,
##     id = idDose, corstr = "ar1", scale.fix = TRUE)
##
##  Coefficients:
##             Estimate Std.err   Wald Pr(>|W|)
## (Intercept)  -8.5234  1.7975 22.484 2.12e-06 ***
## doseM         1.1988  1.2221  0.962   0.3266
## doseH         3.3817  1.1873  8.112   0.0044 **
## time          3.2046  0.6289 25.966 3.47e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Scale is fixed.
##
## Correlation: Structure = ar1  Link = identity
##
## Estimated Correlation Parameters:
##       Estimate Std.err
## alpha   0.1707 0.04126
## Number of clusters:   29   Maximum cluster size: 3
```

Table X shows classification table for the choosed model. The model has a sensitivity of 0.911 and a specificity of 0.839.

TODO(Gerard): Create table with classification

```
##
##       0  1
##    0 41  5
##    1  4 26
```

6.2 Random effects model

TODO(Mathieu)

7 Missing Data Analysis

TODO(Gerard): Missing exploration, in processs

In this section, the patterns of missing data will be discussed and analyzed, in order to assess its influence in the analyses performed in previous sections.

Figure 1 and Table 2 show the missing data distribution over time and for each dose. Clearly, the fact that an observation is missing is associated with the dose and the time (fisher tests p-values= 0.005 and 0.007, respectively). Actually, there are more missings in low and medium dose than in high dose, and also, there are more missings as time increases. Missings in the PCV are not produced completely at random, then.

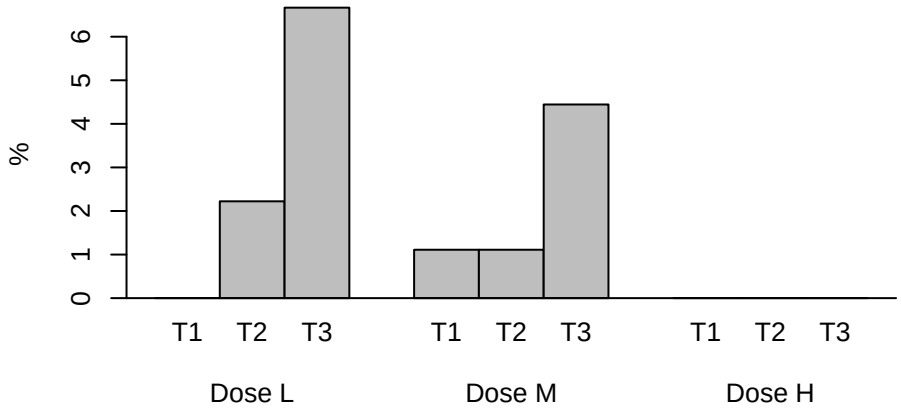

Figure 1: Percentage of missings for each time and dose with respect to the whole sample.

For all this, it is pretty clear that there is some systematic mecanism that produces the missing data. Probably, given that 8 of the 9 cows that have some missings have the first missing after being unhealthy, it would be logic to understand that the missings are produced because of the death or that the cow is dropped out from the study due to having too many health issues. This makes sense also when looking at Figure ????: the cow with ID 5, for example, would have died (or left the study) before taking the time 3 low dose measure, and if one assumes that its sequence of doses was H-L-M, this explains the missings for the medium dose. Following

Dose	Time	Missing	Available
L	1	0	10
L	2	2	8
L	3	6	4
M	1	1	9
M	2	1	9
M	3	4	6
H	1	0	10
H	2	0	10
H	3	0	10

Table 2: Number of missings and available data in the outcome.

this rationale, cows number 1, 2 and 4 would have died at the time 3 of dose L (their sequence would end with L, since there are no more missings); cows number 3, 8 and 10 before time 3 of the medium dose (ending with M dose) and cows with ID 6 and 9 before time 2 of the L dose. In this sense, our sample could be easily biased because we only have data in the low and medium dose for the cows with a less severe infection (due to the death of the cows that have a worse prognostic and that they are more likely to die with a dose other than high). Therefore, the pattern for missings in this data is, probably, missings not at random (MNAR).

Since all the analyses performed previously were done with the complete cases (missing PCV's were omited), the results obtained could be far from the reality (or not). The complete cases analysis can only deal with MCAR type of missing data. Therefore, several approaches will be used to analyse the data taking this into account.

It is natural to think, that if the animals that give up the study are due to deaths, and thus are unthealthy cows. This could justify (in some way) subtituting the missings for the last value observed (last observation carried forward).

8 Limitations and Further research

- Jacknife estimators for small sample better than sandwhich estimator

  Jackknife variance estimators are preferable to the sandwich estimator in case of a small number of clusters. (Højsgaard, Halekoh, Yan 2006)

- Compre a new factor L and M dose against H dose
- Perfect Separation Problem with gee

  It apeared when using a 22 PCV value threshold for healthy or unhealthy cow.

- Compare GEE model with GLMM. Coeficients and SE.

9 Bibliography

TODO(Gerard): Normalize citations

- Pan 2001, Biometrics, Akaike's Information Criterion in Generalized Estimating Equations.
- Højsgaard, S., Halekoh, U. & Yan J. (2006) The R Package geepack for Generalized Estimating Equations Journal of Statistical Software, 15, 2, pp1–11

- Hardin & Hilbe 2002, Generalized Estimating Equations
- Kamil Barton (2015). MuMIn: Multi-Model Inference. R package version 1.13.4. http://CRAN.R-project.org/package=MuMIn