# Survival analysis:
## Clustered data

**Mark Stevenson**

Faculty of Veterinary and Agricultural Sciences

The University of Melbourne, Parkville Victoria 3010 Australia

mark.stevenson1@unimelb.edu.au

# Roadmap

- Background
- Robust variance estimators
- Frailty models

# Background

- The ordinary Cox model treats each outcome event as being independent

- When the assumption of independence doesn't hold the variance of the regression coefficients will be underestimated

- Underestimation of the variance of the regression coefficients makes us more likely to declare the effect of an explanatory variable as having a significant influence on time-to-event when, in reality, it has no effect (a Type II error)

# Background

- There are a number of ways to deal with clustered data
  - fixed-effects models (e.g. inclusion of the herd as a fixed effect)
  - stratified Cox proportional hazards models
  - robust variance estimators
  - frailty models

# Roadmap

- Background
- Robust variance estimators
- Frailty models

# Robust variance estimators

- A method for dealing with the situation when subjects experience multiple events
- Observed data are re-sampled (with replacement) to achieve a sample of the same size each time, and to use the variation in the estimated parameters across the set of samples to obtain a value for the sampling variability of the estimates
- With correlated data the sample needs to be drawn with replacement from the set of independent subjects (not observations) so that intra-subject correlation is preserved in the samples that are taken

# Robust variance estimators

- Example
  - a study of bladder cancer in humans
  - many subjects had recurrences of bladder cancer, sometimes as many as four, and were followed beyond the fourth recurrence
  - multiple rows for each subject

Wei LJ, Lin DY, Weissfeld L (1989) Regression analysis of multivariate incomplete failure time data by modelling marginal distributions. Journal of the American Statistical Association 84: 1065 - 1073

```
library(survival); setwd("D:\\TEMP");
dat <- read.table("bladder2.csv", header = TRUE, sep = ",");
dat;
```

```
id rx size number start stop event obsnum
1  1     3       1     0    1     0      1
2  1     1       2     0    4     0      1
3  1     1       1     0    7     0      1
4  1     1       5     0   10     0      1
5  1     1       4     0    6     1      1
5  1     1       4     6   10     0      2
```

```
id: patient identifier.
rx: 1 = placebo, 2 = thiopet.
size: size of the largest initial tumour.
number: number of initial tumours.
start: entry into the study or the time of last recurrence.
stop: time to event (months).
event: 0 = censored, 1 = event.
obsnum: event number.
```

# Robust variance estimators

- Cox proportional hazards model, with treatment, tumour size and tumour number treated as fixed effects …

```
bladder.cph01 <- coxph(Surv(start, stop, event) ~ rx + size +
number, data = dat);
summary(bladder.cph01);
```

```
           coef exp(coef) se(coef)      z       p
rx      -0.4116     0.663   0.1999 -2.059 0.03900
size    -0.0411     0.960   0.0703 -0.584 0.56000
number   0.1637     1.178   0.0478  3.426 0.00061

        exp(coef) exp(-coef) lower .95 upper .95
rx          0.663      1.509     0.448       0.98
size        0.960      1.042     0.836       1.10
number      1.178      0.849     1.073       1.29

Rsquare= 0.074   (max possible= 0.992 )
Likelihood ratio test= 14.7  on 3 df,   p=0.00213
Wald test            = 15.9  on 3 df,   p=0.00119
Score (logrank) test = 16.2  on 3 df,   p=0.00104
```

# Robust variance estimators

- Now include a robust variance estimator …

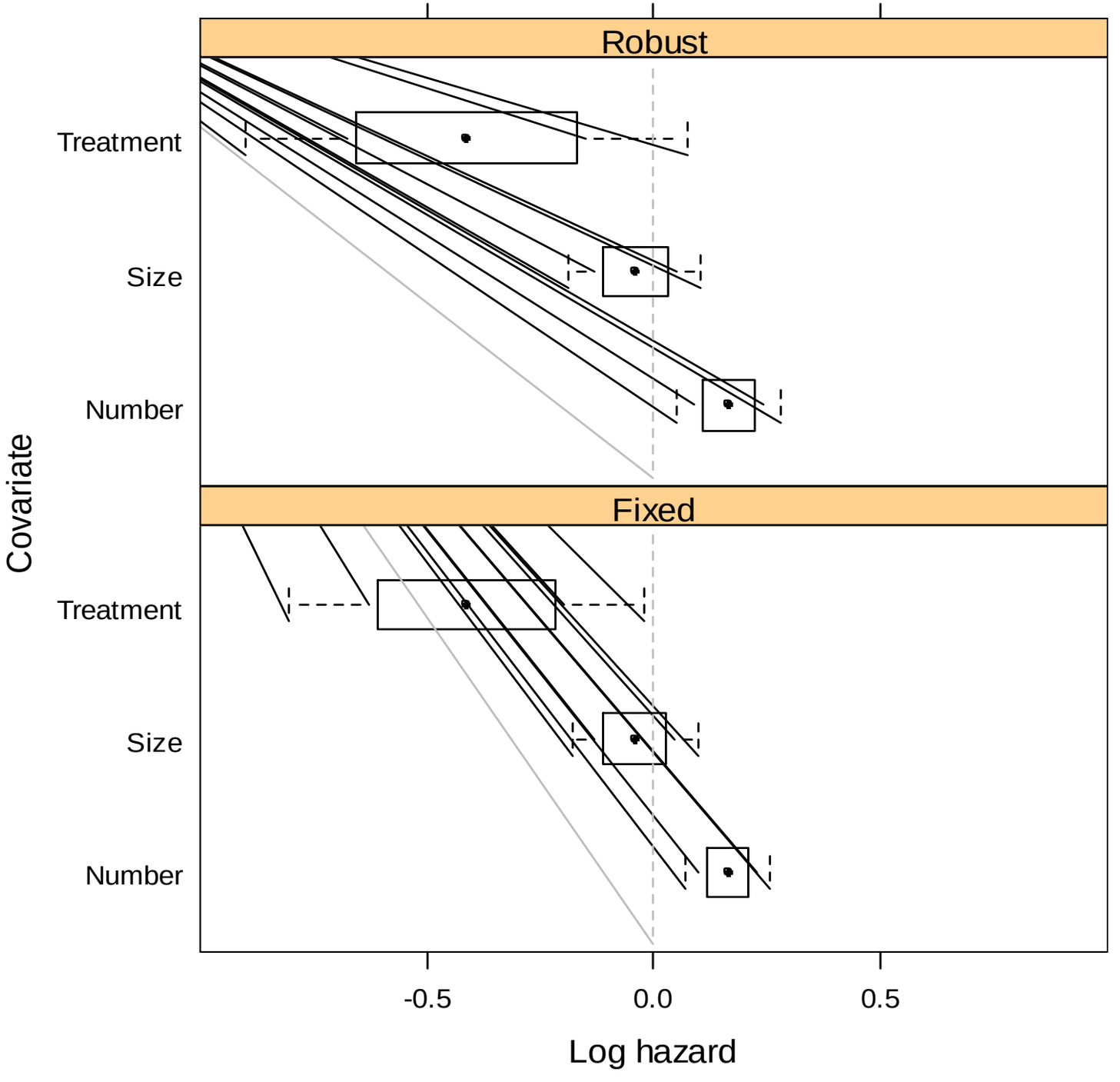
```
bladder.cph02 <- coxph(Surv(start, stop, event) ~ rx + size + number
+ cluster(id), data = dat);
summary(bladder.cph02);
```

```
  n= 190
            coef exp(coef) se(coef) robust se      z      p
rx      -0.4116     0.663   0.1999    0.2488 -1.655 0.0980
size    -0.0411     0.960   0.0703    0.0742 -0.554 0.5800
number   0.1637     1.178   0.0478    0.0584  2.801 0.0051

        exp(coef) exp(-coef) lower .95 upper .95
rx          0.663      1.509     0.407      1.08
size        0.960      1.042     0.830      1.11
number      1.178      0.849     1.050      1.32

Rsquare= 0.074   (max possible= 0.992 )
Likelihood ratio test= 14.7  on 3 df,   p=0.00213
Wald test            = 11.2  on 3 df,   p=0.0107
Score (logrank) test = 16.2  on 3 df,   p=0.00104,   Robust = 10.8
p=0.0126
```

Box and whisker plots showing the variability of estimated log hazard for covariates number of tumours, size of tumours, and treatment type.



# Roadmap

- Background
- Robust variance estimators
- Frailty models

# Frailty models

- We've just talked about the situation where there can be multiple events for each subject (e.g. bladder cancer)

- Consider now the situation where individuals are organised into groups ('clusters') where the probability of event may depend (in part) on characteristics of the cluster
  - reproduction in herds of dairy cattle
  - onset of genetic disorders in families

# Frailty models

- The presence of unobserved (or unobservable) risk factors attributable to the cluster leads to unobserved heterogeneity in the hazard

- Unobserved heterogeneity in survival analysis is also called frailty (i.e. more ‘frail’ individuals have a higher probability of experiencing the outcome of interest)

# Frailty models

- Two types of frailty
  - individual frailty: something special about an individual that makes them more likely to fail (e.g. genetics)
  - shared frailty: something common to a group of individuals (e.g. herd) makes them more likely to fail

# Frailty models

- If there are individual-specific unobserved factors that influence the hazard, the observed form of the hazard function at the aggregate population level will tend to be different from those at the individual level

- Even if the hazards of individuals in a population are constant over time the aggregate population hazard may be time-dependent, typically decreasing: this may be explained by a selection effect operating on individuals

# Frailty models

- Selection effects:
  - high risk individuals will tend to experience the event of interest first, leaving lower risk individuals in the population
  - as time progresses the population is increasingly depleted of individuals most likely to experience the event, leading to a decrease in the population hazard
  - because of this selection, we may see a decrease in the population hazard even if the individual hazards are constant (or even increasing)

# Frailty models

- We can introduce a random effect term which represents individual-specific unobserved effects:

$$\log h_{ij}(t) = \alpha(t) + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + U_j$$

  - we usually assume $U_j \sim N(0, {\sigma_u}^2)$; ${\sigma_u}^2$ represents the variance of the frailty for cluster $j$
  - an alternative is to specify a gamma distribution for the frailty term, which allows it be asymmetric (allowing for groups displaying exceptionally low or high risk, as in the case of genetic disease studies where the presence of a high-risk allele markedly increases the hazard of failure)

# Frailty

- Example
  - mortality in patients with advanced lung cancer
  - Loprinzi CL, Laurie JA et al. (1994) Prospective evaluation of prognostic variables from patient-completed questionnaires Journal of Clinical Oncology 12: 601 - 607

```
library(survival); setwd("D:\\TEMP");
dat <- read.table("lung.csv", header = TRUE, sep = ",");
head(dat);
```

```
inst time status age sex ph.ecog ph.k pat.k meal.cal wt.loss
3     306  2      74  1   1        90   100   1175       NA
3     455  2      68  1   0        90    90   1225       15
3    1010  1      56  1   0        90    90     NA       15
5     210  2      57  1   1        90    60   1150       11
1     883  2      60  1   0       100    90     NA        0
12   1022  1      74  1   1        50    80    513        0


inst: enrolling institution.
time: day of event.
status: 1 = alive, 2 = dead.
age: patient age at enrolment.
sex: 1 = male, 2 = female.
ph.k: physician's estimate of Karnofsky score.
pat.k: patient's estimate of Karnofsky score.
meal.cal: calories consumed at meals.
wt.loss: weight loss in the last six months.
```

# Frailty

- There are 18 separate institutions that enrolled at least one subject in this trial

- Because the enrolling institutions range from community practices to a large tertiary care centre, differences in the baseline risk of enrollees might be a concern

- A fixed-effects Cox proportional hazards model (i.e. one that ignores intra-institutional correlation) would be called as follows ...

```
lung.cph01 <- coxph(Surv(time, status) ~ sex + ph.karno + pat.karno,
data = dat);
summary(lung.cph01);
```

```
  n=224 (4 observations deleted due to missing)
              coef exp(coef) se(coef)      z      p
sex        -0.51188     0.599  0.16927 -3.024 0.0025
ph.karno   -0.00616     0.994  0.00682 -0.902 0.3700
pat.karno  -0.01702     0.983  0.00654 -2.604 0.0092

           exp(coef) exp(-coef) lower .95 upper .95
sex            0.599       1.67      0.43     0.835
ph.karno       0.994       1.01      0.98     1.007
pat.karno      0.983       1.02      0.97     0.996

Rsquare= 0.097   (max possible= 0.999 )
Likelihood ratio test= 22.9  on 3 df,   p=4.21e-05
Wald test            = 22.6  on 3 df,   p=4.79e-05
Score (logrank) test = 23.2  on 3 df,   p=3.67e-05
```

# Frailty

- Now account for enrolling institution as fixed effect ...

```
dat$inst <- factor(dat$inst);
contrasts(dat$inst) <- contr.treatment(18, base = 1, contrasts =
TRUE);

lung.cph02 <- coxph(Surv(time, status) ~ sex + ph.karno + pat.karno
+ inst, data = dat);
summary(lung.cph02);
```

```
coef exp(coef) se(coef)         z      p
sex         -0.5197  5.95e-01 1.76e-01 -2.94734 0.0032
ph.karno    -0.0108  9.89e-01 7.26e-03 -1.49391 0.1400
pat.karno   -0.0190  9.81e-01 6.86e-03 -2.76976 0.0056
inst2        0.5159  1.68e+00 5.43e-01  0.95078 0.3400
inst3       -0.4168  6.59e-01 3.31e-01 -1.25751 0.2100
…
inst18     -13.9514  8.73e-07 2.27e+03 -0.00614 1.0000

           exp(coef) exp(-coef) lower .95 upper .95
sex         5.95e-01   1.68e+00     0.421     0.840
ph.karno    9.89e-01   1.01e+00     0.975     1.003
pat.karno   9.81e-01   1.02e+00     0.968     0.994
inst2       1.68e+00   5.97e-01     0.578     4.852
inst3       6.59e-01   1.52e+00     0.344     1.262
…
inst18      8.73e-07   1.15e+06     0.000       Inf

Rsquare= 0.173   (max possible= 0.998 )
Likelihood ratio test= 42.4  on 20 df,   p=0.00242
Wald test            = 41.8  on 20 df,   p=0.00296
Score (logrank) test = 44  on 20 df,   p=0.00150
```

# Frailty

- Including the cluster variable as a fixed effect is OK when there are only a small number of levels of the variable (< ~6)

- Presentation of the model becomes clumsy when the number of levels of the factor is large (as in the example above)

# Risk factors for culling and deaths in eight dairy herds

MA STEVENSON[a] and IJ LEAN
Department of Animal Science, University of Sydney, Camden, New South Wales 2570

**Objectives** To identify risk factors for culling of dairy cows from eight New South Wales dairy herds.

**Design** A longitudinal population study of dairy cow culling in eight non-seasonally calving dairy herds in the Camden district of New South Wales. Cox's proportional hazards model was used to evaluate various risk factors for culling for a specific reason (sales, deaths, reproductive failure, disorders of the udder and low milk production).

assessed since 1968.[9] The purpose of this longitudinal population study was to identify factors that influenced the risk of culling from eight dairy herds in the Camden district of New South Wales. A description of the pattern of reason-specific culling with respect to length of productive life and length of time after calving is given in an accompanying paper.[11]

## Materials and methods

*Study population and data collection*

*Aust Vet J* Vol 76, No 7, July 1998

Stevenson, M., Lean, I., 1998. Risk factors for culling and deaths in eight dairy herds. Australian Veterinary Journal 76, 489 - 494.

**Table 4. Effect of farm, first lactation milk production and previous calving interval on the risk of removal for udder disorders after the second calving.**

| Explanatory variable | Regression coefficient (SE) | P | Relative risk (95% CI) |
|---|---|---|---|
| *Farm comparison* | | 0.02[a] | |
| Farm 2 vs farm 1 | +0.3526 (0.3290) | | 1.42[b] (0.75 - 2.71) |
| Farm 3 vs farm 1 | -0.2339 (0.3416) | | 0.79 (0.41 - 1.55) |
| Farm 4 vs farm 1 | -0.5927 (0.4045) | | 0.55 (0.25 - 1.22) |
| Farm 5 vs farm 1 | -0.1527 (0.3754) | | 0.86 (0.41 - 1.79) |
| Farm 6 vs farm 1 | +0.3804 (0.3561) | | 1.46 (0.73 - 2.94) |
| Farm 7 vs farm 1 | -0.1093 (0.3892) | | 0.90 (0.42 - 1.92) |
| Farm 8 vs farm 1 | +0.3140 (0.3453) | | 1.37 (0.70 - 2.69) |
| PRO1 (500 L increments) | +0.1000 (0.0500) | 0.01 | 1.11[c] (1.00 - 1.22) |
| ICI (10 day increments) | -0.0370 (0.0140) | 0.01 | 0.96[d] (0.94 - 0.99) |

Data consisted of 157 culled cows and 788 censored observations (cows that were either present at the end of the study, were transferred to a new herd, or were lost to follow up).

[a]The significance of inclusion of all seven farm variables in the model.

[b]For every day of productive life, cows from farm 2 have a 1.42 times risk of removal for udder disorders compared to cows from farm 1 ($P < 0.02$).

[c]Compared to cows with population mean PRO1, for every 500 L increase in first lactation milk yield, there was a 1.11 times risk of removal for udder disorders ($P = 0.01$).

[d]Compared to cows with population mean ICI, for every 10 day increase in the previous calving interval, there was a 0.96 times risk of removal for udder disorders ($P = 0.01$).

# Frailty

- Including the cluster variable as a fixed effect is OK when there are only a small number of levels of the variable (< ~6)

- Presentation of the model becomes clumsy when the number of levels of the factor is large (as in the example above)

- An alternative is to treat the cluster variable as a frailty term …

```
lung.cph03 <- coxph(Surv(time, status) ~ sex + ph.karno + pat.karno
+ frailty(inst), data = dat);
summary(lung.cph03);
```

```
                coef    se(coef) se2     Chisq DF   p
sex            -0.5098 0.16953  0.16950 9.04  1.00 0.0026
ph.karno       -0.0062 0.00685  0.00684 0.82  1.00 0.3700
pat.karno      -0.0170 0.00654  0.00654 6.77  1.00 0.0093
frailty(inst)                           0.25  0.21 0.3600

           exp(coef) exp(-coef) lower .95 upper .95
sex            0.601       1.67     0.431     0.837
ph.karno       0.994       1.01     0.981     1.007
pat.karno      0.983       1.02     0.971     0.996

Iterations: 5 outer, 16 Newton-Raphson
Variance of random effect= 0.00149   I-likelihood = -711.9
Degrees of freedom for terms= 1.0 1.0 1.0 0.2
Rsquare= 0.099   (max possible= 0.998 )
Likelihood ratio test= 23.2  on 3.21 df,   p=4.7e-05
Wald test            = 22.6  on 3.21 df,   p=6.43e-05
```

# Frailty

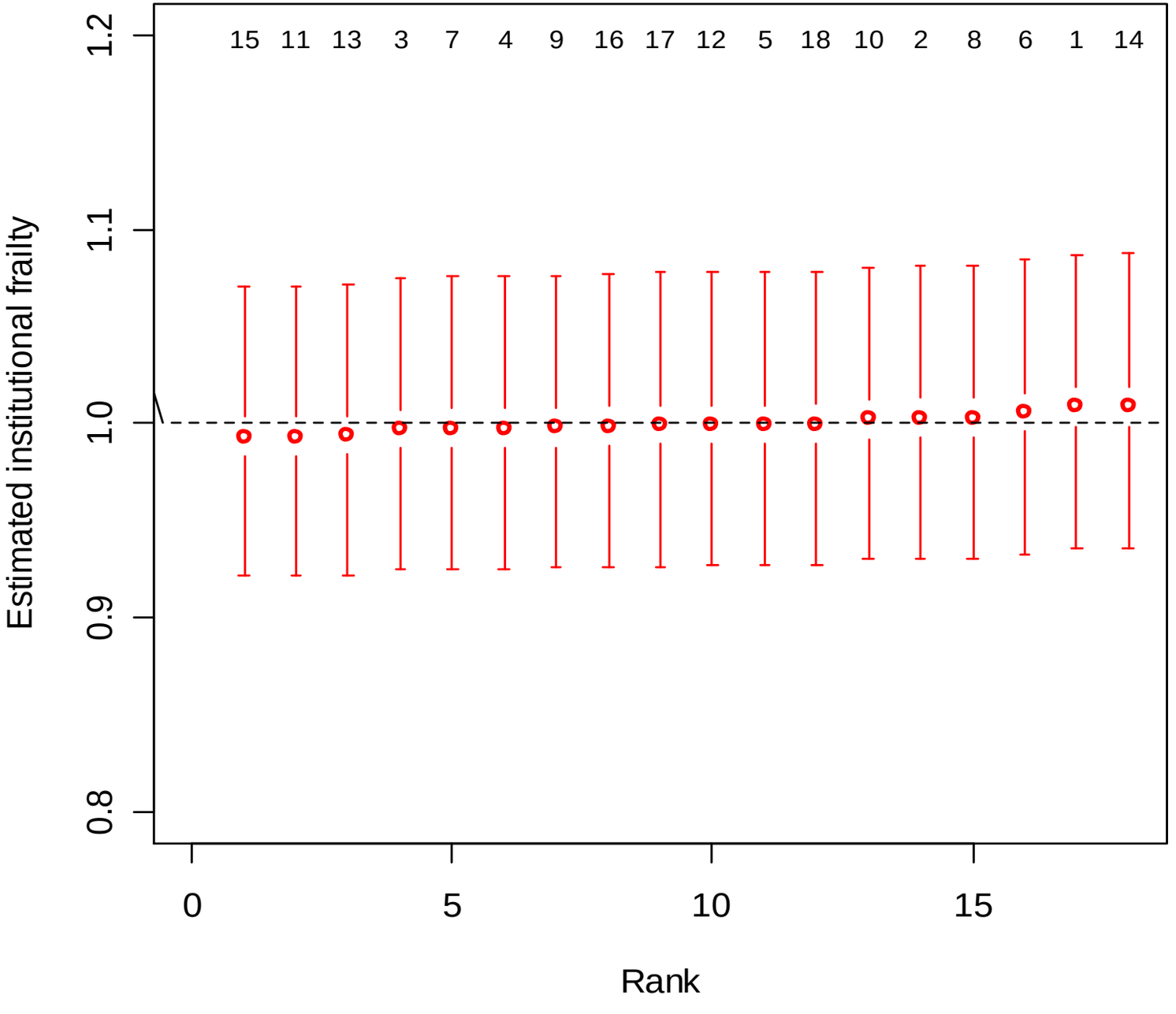
- We can plot the frailty term for each institution to better-appreciate those institutions that are prone to ‘earlier failures’ …

- Extract the frailty estimates using the `predict` function:

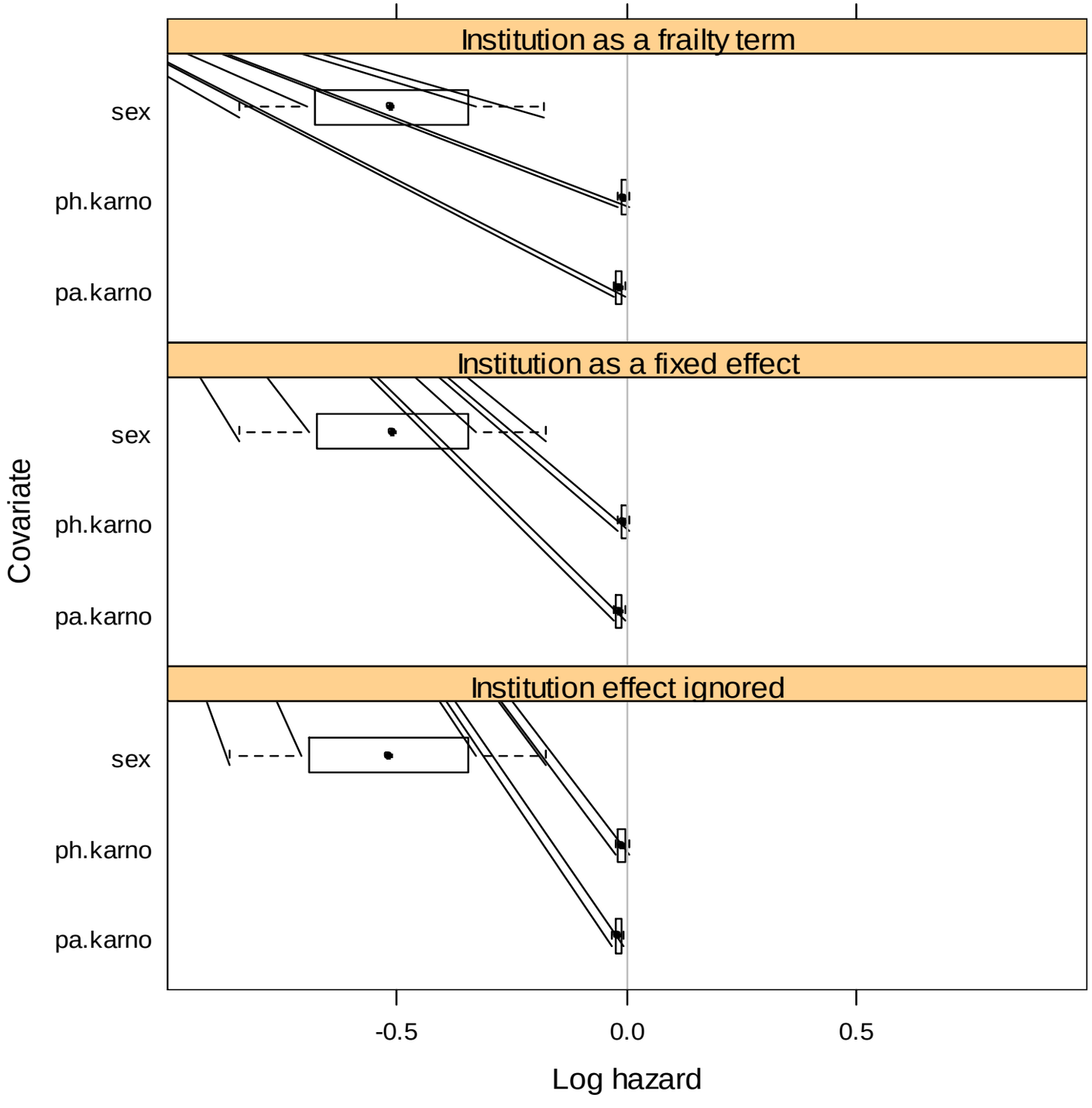
```
predict(lung.cph03, type = "terms", se.fit = TRUE)
```

- The argument `se.fit = TRUE` returns the standard errors, allowing you to calculate a 95% confidence interval for the frailty terms

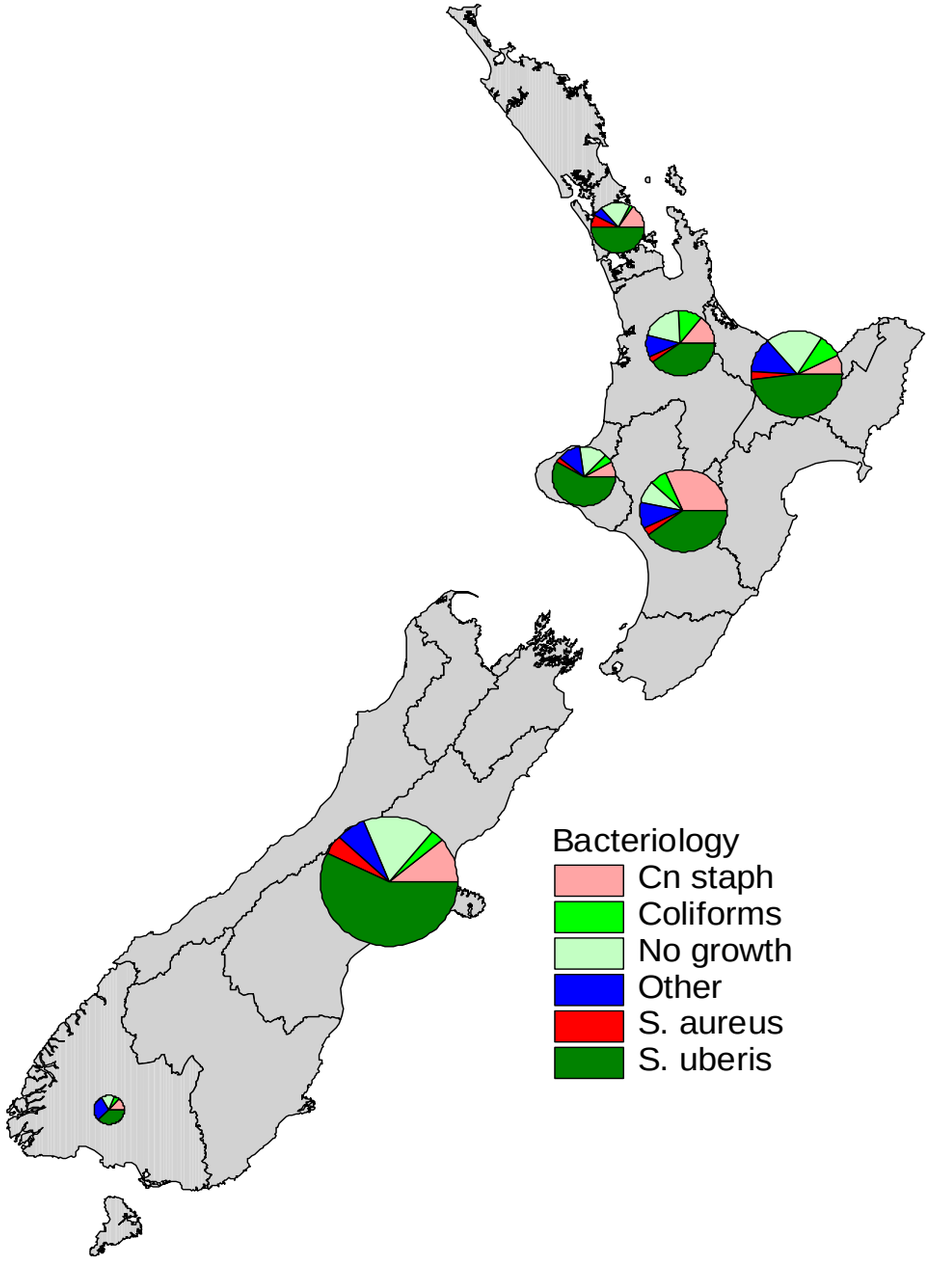
Influence of institution on the daily hazard of death for patients enrolled in a study of advanced lung cancer patients, conducted by the North Central Cancer Treatment Group. Institution identifiers are shown at the top of the plot. The effect of institution on hazard of death is subtle, with institutions 15, 11, and 3 demonstrating the lowest daily hazard of failure. Confidence intervals are wide, demonstrating no clear evidence that one institution is associated with lower or higher hazards of failure over another.



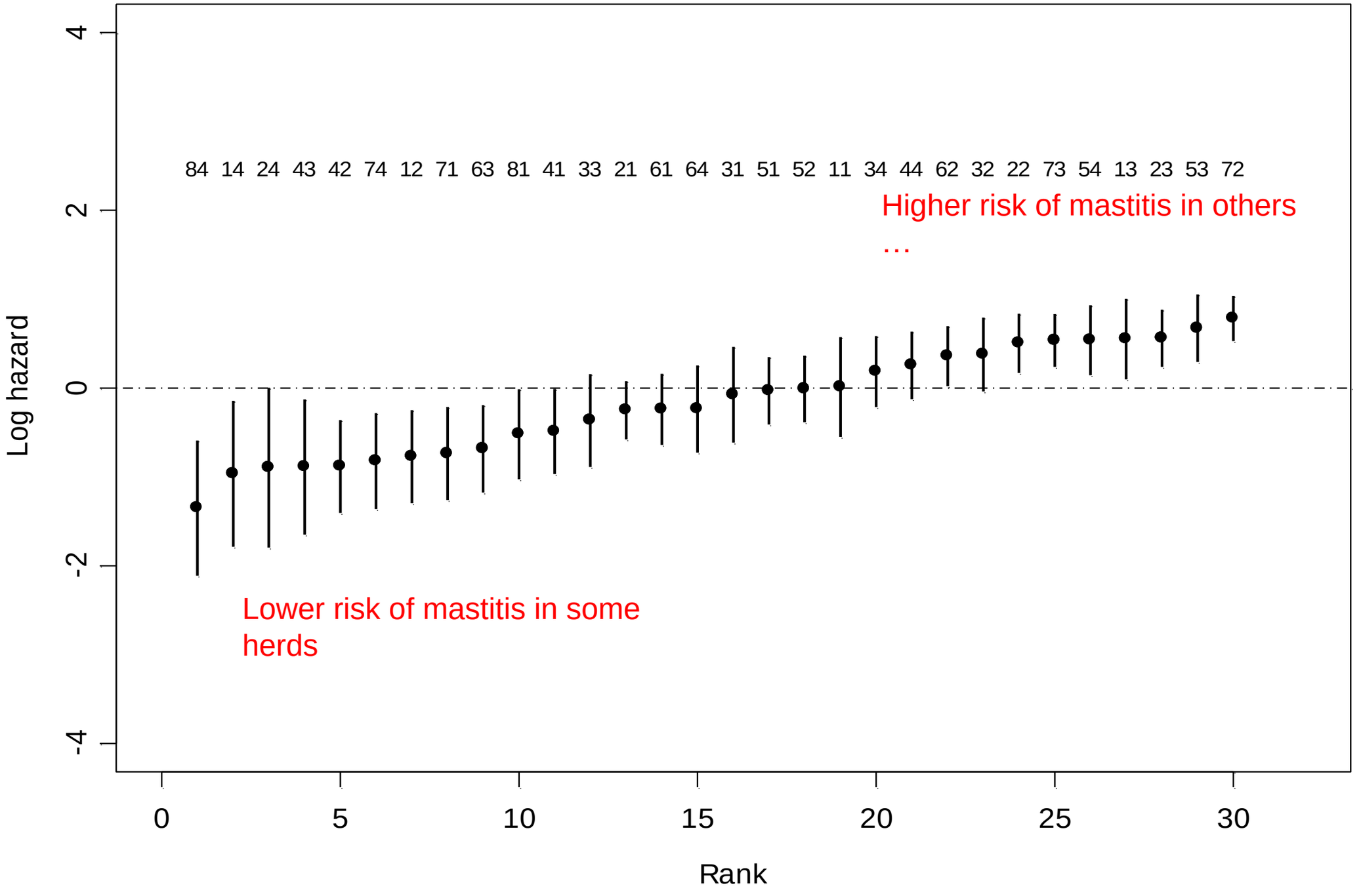
Box and whisker plots showing the variability of estimated log hazard for covariates pa.karno, ph.karno, and sex.

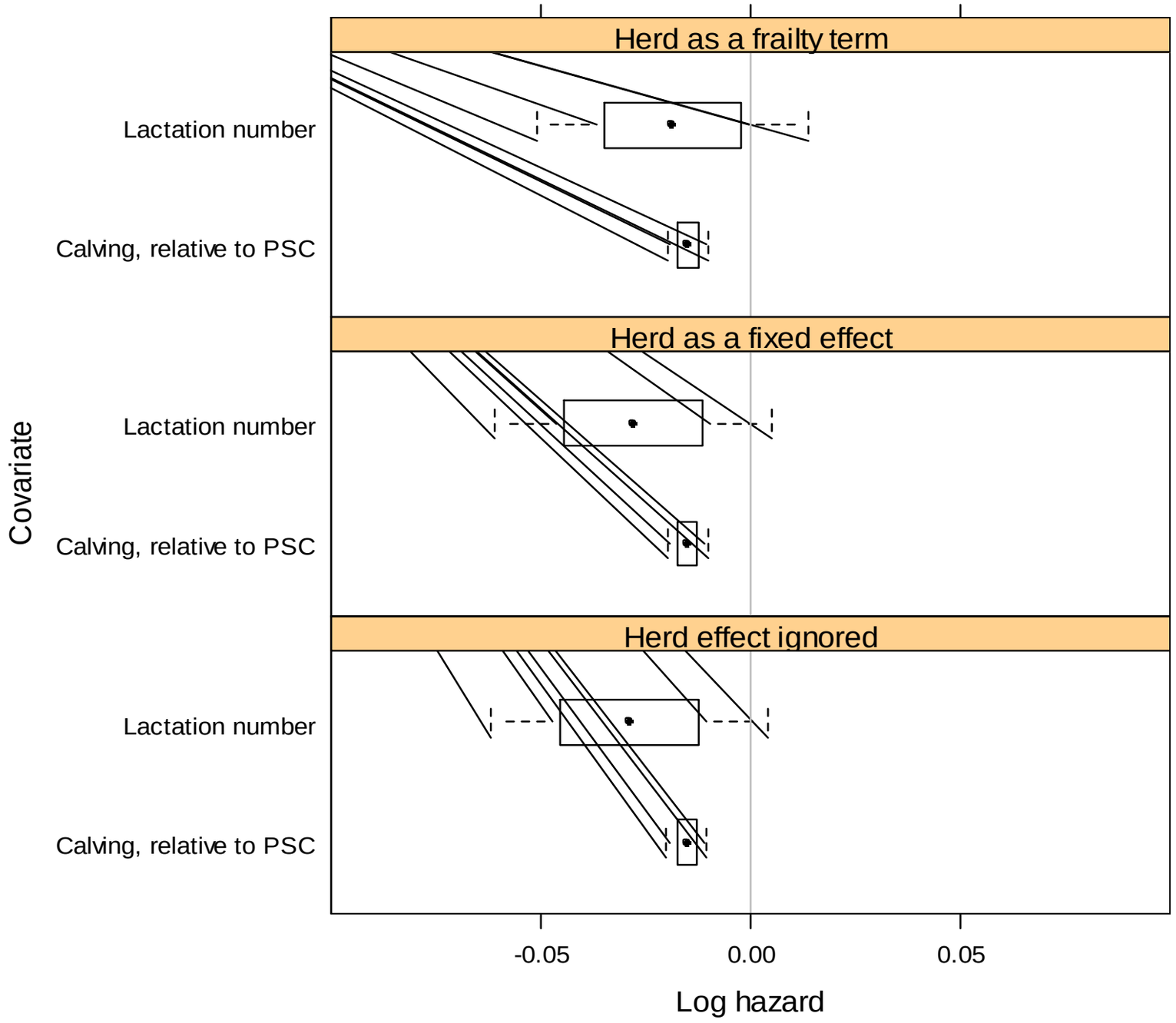


Pie charts showing the breakdown of primary isolates from milk samples taken cows and heifers diagnosed with clinical mastitis in either the 7 days before or the 7 days after calving. Size of pie charts are proportional to the number of mastitis events recorded for each region.



Estimates of herd-level risk of PCM (and 95% confidence intervals) after controlling for the effect of lactation number and day of calving, relative to herd PSC (cows and heifers).




Herd as a frailty term
Herd as a fixed effect
Herd effect ignored
Lactation number
Calving, relative to PSC
Covariate
-0.05
0.00
0.05
Log hazard

# Roadmap

- Background
- Robust variance estimators
- Frailty models