WILD 502

Lab 2 – $\hat{S}$ from Known-fate Data with Individual Covariates

Today's lab presents material that will allow you to handle additional complexity in analysis of survival data. The lab deals with estimation using models that include individual covariates. You'll gain experience with working with the Design Matrix and with model output where covariates are used.

Individual Covariates

Below, you can see the header information and first 10 rows of data in 'lab2fawn.inp', which is a file with data from 1 time interval that treats all animals as belonging to 1 group but that also includes individual covariates as listed. Because we have individual covariates, we have to have a row of data for each individual rather than the collapsed format that we used last week. The fate of each individual is entered using a live-dead format: the 1st number indicates if the animal was alive and studied in a given time interval (1 = yes it was alive and studied) and the 2nd number indicates if the animal died or not in that interval (1 = yes, it died): 10 indicates it was studied and survived; a 11 indicates it was studied and died. Here, there is only 1 time interval, so it wouldn't make sense to have any animals entered as 00, i.e., not studied in 1st interval. But, if you had multiple intervals (say 3 years), you could enter new animals as the study goes along, e.g., 001011 (not studied the 1st year, studied and survived the 2nd year, and died in the 3rd year) or 000010 (not studied the 1st or 2nd years; studied and survived the 3rd year).

```
/* Fawn survival for 1 interval,
Comments provide (1) radio frequency and (2) days lived.
Live-dead encounter history for 1 occasion comes next.
Number of animals with this encounter history comes after encounter history.
Individual Covariates are:
    1) Area: 0=Control 1=Treatment
    2) Sex:  0=Female 1=Male
    3) Mass (kg)
    4) Length (cm)*/
/*  191.62 10    */    11    1    1    0    34     126    ;
/*  192.6  11    */    11    1    1    0    33.4   128.5  ;
/*  191.64 12    */    11    1    1    0    29.9   130    ;
/*  192.79 16    */    11    1    1    0    31     113    ;
/*  192.22 16    */    11    1    1    0    26.7   116    ;
/*  192.53 13    */    11    1    1    0    23.9   124    ;
/*  192.66 28    */    11    1    1    1    33.5   126    ;
/*  192.38 27    */    11    1    1    1    37     129.5  ;
/*  192.68 26    */    11    1    1    0    26.5   113    ;
/*  192.3  31    */    11    1    1    1    41     117.5  ;
```

1) Open MARK and choose File, New, and use "lab2fawn.inp" – when you examine the input file you should notice the following:
   a. There is 1 encounter occasion and 1 group.
   b. There are 4 individual covariates. So, for models with no interactions, there are 2^4 possible models, which is 2 x 2 x 2 x 2 = 16 models. In a real study, you would want to develop a set of *a priori* models. Here, you will run a subset of those 16 models plus several others that are of interest (see below) to gain experience with building models with individual covariates.

2) When specifying the features of this new analysis, be sure to
   a. select "Known Fate" as the data type,
   b. provide a meaningful title – perhaps "Lab 2 Fawns with covariates"
   c. choose the correct # of occasions, groups, and individual covariates,
   d. ENTER INDIVIDUAL COVARIATES NAMES (use button shown below). This doesn't have to be done, but it's easier to keep track of things later if you do. Go ahead and use 'Area', 'Sex', 'Mass', & 'Length'


   e. Next, check out the Parameter Index Matrix for this dataset. Because we entered data for a single time interval and single group for a known-fate analysis, there's only a PIM for survival rate and it has a single cell, which means you'll only have 1 row in the design matrix.
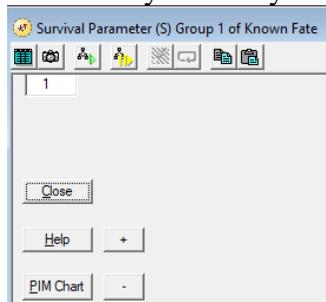
3) To run each model of interest, you'll click on Design, request as many columns as needed, and then fill in values in each column to create the model of interest.

   a. Run an S(Length) model, which should have a Design Matrix with 2 columns: a 1 in the first column (intercept) and the word "Length" in the 2nd column (if you named it "length" or used some abbreviation, type that version of the word)
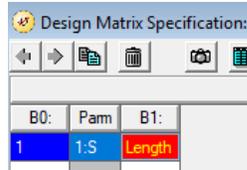
i. Accept the defaults for the numerical estimation run
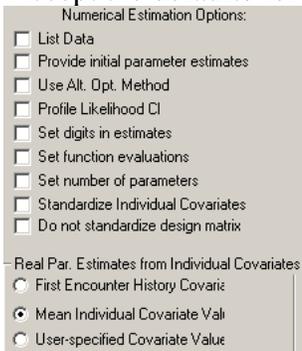
i. After you've run the model, go to the Results Browser and examine the beta estimates for this model. Is the point estimate for the effect of length positive, zero, or negative? What evidence supports your claim? Consider what the CI indicates about the effect of *Length*: all positive, positive and negative, or all negative.

ii. Use MARK's "Individual Covariate Plot" tool to build a plot of Survival Rate vs Length. Open the tool, click on "1:S" as the parameter to plot, choose "Length" as the covariate to use in the plot, accept the default range of values (min and max in the dataset), output the data to EXCEL, and click on the "OK" button. Once you've done this, examine the plot and view the EXCEL output. Next, enter the following R code into a script in R, copy the data from Excel to the clipboard and execute the R commands.

```
s_length <- read.table("clipboard", header = F,
                col.names = c("Length", "S",
                              "se_S", "lcl_S", "ucl_S"))
```

iii. Use R and the *dplyr* package to calculate the log-odds of survival & associated 95% LCI and UCI values using these R commands.

```
library(dplyr)
s_length <- s_length %>%
  mutate(log_odds_S = qlogis(S),
         lcl_log_odds_S = qlogis(lcl_S),
         ucl_log_odds_S = qlogis(ucl_S))
```

iv. Finally, use the *ggplot2* package and the following R commands to plot (a) the log-odds of survival vs. length and (b) survival vs. length.

```
library(ggplot2)
ggplot(s_length, aes(x = Length, y = log_odds_S)) +
  geom_ribbon(aes(ymin = lcl_log_odds_S, ymax = ucl_log_odds_S),
              fill = "grey80") +
  geom_line()

ggplot(s_length, aes(x = Length, y = S)) +
  geom_ribbon(aes(ymin = lcl_S, ymax = ucl_S), fill = "grey80") +
  geom_line() +
  ylim(0, 1)
```

      b. Run an S(Area) model.
          i. Go to the Results Browser and examine the beta estimates for this model. Is the point estimate for the effect of *Area* positive, zero, or negative?
         ii. What does the CI for effect of *Area* indicate?
        iii. What are the estimates of Survival for each area (include SE's)? One way to get MARK to provide these values along with proper SE's is to (a) select the S(Area) model, i.e., highlight it, and (b) use the "Run" menu's option to "Regenerate Real and Derived Estimates". The first time you do this, set area=0 and record the real estimates that are produced. Repeat, set area=1, and record the estimate again (Note: values for other covariates don't matter as they don't appear in the S(Area) model). If you use this method, you'll have 2 copies of the model in your results browser, which messes up your model-selection results. So, you can run them both, record the estimates, and then delete all but 1 of the copies of the S(Area) model. Another way of proceeding is to use the individual covariate plot tool in MARK to output estimates to Excel for values of Area from 0 to 1, which will mean you'll have to delete all but the 1st and last rows of output in Excel to get what you want.

      c. Run the S(Sex+Length) model.
          i. Next, go to the Results Browser and examine the beta estimates for this model. Are the point estimates of the 2 effects positive, zero, or negative? Be sure you can determine which beta estimate goes with each covariate. If you're stumped, look at the Design Matrix.
         ii. What do the CIs for effects indicate? All positive, positive and negative, all negative?

iii. Use MARK's individual covariate plot and look at plots for (a) females (sex = 0) and for (b) males (sex = 1).
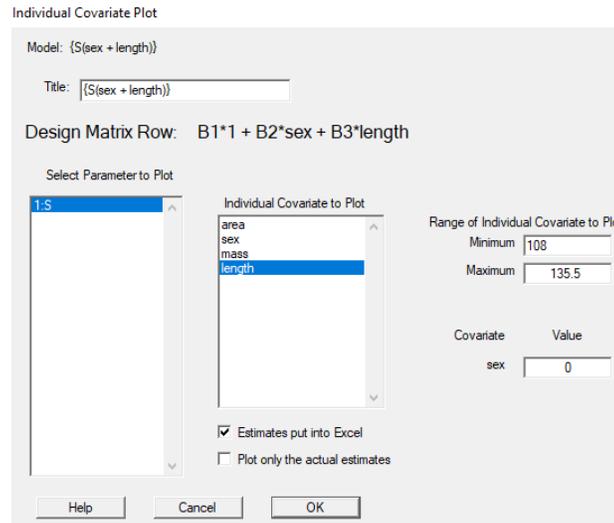
iv. You can then go to each of the resulting Excel spreadsheets and add a column for 'sex' where you then enter the appropriate value for each dataset. Next, you can combine the 2 spreadsheets into 1 and copy them over to R using the below to bring in the data and make useful plots.

```
s_sex.length <- read.table("clipboard", header = F,
                    col.names = c("Length", "S", "se_S",
                                    "lcl_S", "ucl_S", "sex"))
s_sex.length <- s_sex.length %>%
  mutate(Sex = factor(sex, levels = c(0, 1),
                            labels = c("female", "male")),
          log_odds_S = qlogis(S),
          lcl_log_odds_S = qlogis(lcl_S),
          ucl_log_odds_S = qlogis(ucl_S))

# use of 'alpha' on ribbons makes them opaque
ggplot(s_sex.length, aes(x = Length, y = log_odds_S, group = Sex)) +
  geom_ribbon(aes(ymin = lcl_log_odds_S, ymax = ucl_log_odds_S),
                alpha = 0.2) +
  geom_line(aes(color = Sex))

ggplot(s_sex.length, aes(x = Length, y = S, group = Sex)) +
  geom_ribbon(aes(ymin = lcl_S, ymax = ucl_S), alpha = 0.2) +
  geom_line(aes(color = Sex)) + ylim(0, 1)
```

d. Run all of the models in the table below

Number	Model
1	{S(.)}
2	{S(Area)}
3	{S(Mass)}
4	{S(Length)}
5	{S(Length + Length-squared)}
6	{S(Sex)}
7	{S(Area + Length)}
8	{S(Length + Mass)}
9	{S(Sex + Length)}
10	{S(Length + Length x sex)}
11	{S(Sex + Length + Length x sex)}

Think carefully about what the differences are among models 9, 10 & 11, especially in terms of how many intercepts and slopes are present in each. You'll probably need a little help running models like the *S(Length+Length-squared)* model, which require you to either have (i) products of covariates (e.g., *Length^2* or *Length x Sex*) in the list of covariates (we don't have that) or to (ii) learn about MARK's *Design Matrix Functions.* From the MARK help file, "Twelve special functions are allowed as entries in the design matrix: add, product, power, min, max, log, exp, eq (equal to), gt (greater than), ge (greater than or equal to), lt (less than), and le (less than or equal to)." We need to use the *product* function here. Your design matrix needs to look like the one below.

For the "S(Sex + Length + Length x sex)" model, the design matrix needs to look like:

e. You can obtain useful graphics from any of the models using the strategies provided above for getting estimates to Excel and then over to R for plotting using R code that you modify as needed for other models (e.g., change the name of the data frame to indicate which model you're working with; choose to use a data with or without the 'sex' variable in it depending on whether the model includes that covariate or not). Be sure you know how the *beta* estimates and values of *Length* and *Sex* are used to generate the predicted values for all models and especially for models 9, 10, and 11.

Formal assignment:
Note: feel free to make any of the plots requested in MARK (if that's possible), Excel, or R depending on what works best for you.

1) A description of the data available for the analysis.
2) A table that contains 1 column with model names and another column that has the actual equation for the log-odds of survival.

Model Name	Equation for ln(S/(1-S))
S(.)	B1
S(Area)	B1 + B2*Area
…	…
S(Sex + Length)	B1+B2*Sex+B3*Length
S(Length + Length*Sex)	B1+B2*Length+B3*Length*Sex
…	…

3) Table of model-selection results. <u>Note:</u> if you ran any of the models more than once to obtain estimates of survival rate for different covariate conditions, you'll need to delete the repeats in order to obtain a valid table of AIC results.
4) Statements about the best model and model-selection uncertainty.
5) Features of the best model in terms of beta estimates (with SE's).
6) Graphics of survival (along with 95% CIs) as a function of the covariates contained in the best model. (Hint: you can provide a separate graphic for each sex obtained directly from Program MARK if you are having trouble using the R commands).
7) A graphic of log-odds of survival (along with 95% CIs) as function of length from the S(Length) model (you can make this in R or Excel, but nor MARK).
8) A graphic of survival as function of length (along with 95% CIs) and an explanation of why this plot is curvilinear whereas the plot of $\ln(\frac{S}{1-S})$ versus length was linear.
9) For the S(Area) model, what do the estimates of beta's and real parameters (along with their uncertainty indicate about possible differences in survival between the 2 areas?
10) For S(Sex+Length) model, what can be inferred about the effects of sex and length on survival rate. (Hint: for the sex effect, it's important to know how sex was coded in the original data, and this is explained at the head of the input file).
11) For the S(Length+Length-squared) model, please provide a plot of estimated survival rate (along with 95% confidence intervals) versus length. Also, please provide the accompanying equation for the line using the actual estimated beta values you obtained from MARK.
12) Your interpretation of what can be learned about survival of mule deer fawns given this dataset and this set of models. Use quantitative evidence to support your statements and feel free to refer to tables, graphs, and results you provided to earlier answers.
13) Provide any comments and questions that you may have on the analysis that you would like to have discussed at the start of the next lab.