Armidale Practical Materials

Michael Morrissey

02 February, 2020

The data

Soay sheep are a neolithic breed of sheep that have been unmanaged in the St Kilda (Hebrides, Scotland) archipelago since prehistoric times. Since 1985, an individual-based study has collected life history, pedigree, morphological, phenological and parasitological data on the Soay sheep population in Village Bay on Hirta, St Kilda. The example dataset contain records of selected phenotypes and fitness measures for female Soay lambs with normal horns from selected cohorts that had relatively high first year survival. The phenotypic traits are all morphological measures for mass (`MASS`, kg), right hind metatarsal length (`HINDLEG`, mm) and horn length (`HORNLEN`, mm) measured in August of the first year of life (age: approximately 4 months). Fitness (component) measures are first year survival (`SURV`) coded as whether or not an individual survived its first winter (1: survived, 0: died), and total lifetime reproductive success (`LRS`), calculated as the number of live births recorded for each ewe throughout her life.

Here is code that might be useful for reading from the data file, assuming that it is in your working directory:

```
d<-read.table("./soay_ewe_lamb_selection_data.csv",sep=',',header=TRUE)
```

```
str(d)
```

```
## 'data.frame':    213 obs. of  6 variables:
##  $ ID     : int  2378 2110 2181 2248 2293 2311 2314 2323 2339 2351 ...
##  $ MASS   : num  15 10.2 11.4 14.3 12.2 15.7 14.4 16.6 16 13.5 ...
##  $ HINDLEG: int  168 144 161 167 165 171 165 176 169 156 ...
##  $ HORNLEN: int  101 65 81 19 92 106 100 120 72 85 ...
##  $ SURV   : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ LRS    : int  3 0 0 0 0 1 0 15 5 2 ...
```

Before setting off on selection analyses, you might want to explore the data a bit. Have a brief look at the means and variability of each trait. Histograms might be useful.

Univariate selection

(1) What is the mean mass in the unselected population, and among those individuals that survive their first year?

(2) What is the selection differential for mass acting through first year survival?

(3) Convert the selection differential that you just calculated into mean-standardised and variance-standardised differentials.

(4) In the previous question, you could have worked out what operation (e.g., multiplication, division), with the mean and standard deviation of mass, was necessary to apply to the selection differential, in order to make the re-scalings. Alternatively, you could have re-scaled the phenotype and then re-calculated the selection differentials on those re-scaled phenotypes. Whichever you didn't do above, do it now to verify your initial calculation.

(5) What is the selection gradient for mass, acting through first year survival (do the calculation without implementing a linear regression, i.e., without using the `lm()` function).

(6) Check that your calculation of the selection differential for mass (via first year survival) agrees with the OLS-based method (i.e., from Lande and Arnold 1983).

(7) What are the units of the selection gradient that you just calculated?

(8) Calculate the selection differential and gradient for hind leg length (also acting through first year survival)?

(9) Which is stronger, selection of hind leg length or mass?

(10) How much does first year viability selection change the variance of mass and leg length?

(11) How much does first year viability selection change the variance of mass and leg length, over and above the effect of strictly directional selection to change the variance?

Multivariate selection

(12) Calculate the multivariate selection gradients of body mass, hind leg length, and horn length, acting through first winter viability.

(13) Do our conclusions about the strength of directional selection on mass and leg length change when we consider the direct component of selection, rather than the overall association?

```
# a slight leg up on getting the vectors of the mean and sds that you'll need (hint, hint)
z<-d[,c("MASS","HINDLEG","HORNLEN")]
mu<-apply(d[,c("MASS","HINDLEG","HORNLEN")],2,mean)
sigma<-apply(d[,c("MASS","HINDLEG","HORNLEN")],2,sd)
```

(14) [Advanced] Say you wanted to use a GLM to characterise the function relating mass, hind leg length, and horn length to fitness, and you particularly wanted to then recover the selection gradient estimate that belong to that specific model. Say your model was a binomial GLM, like this:

```
glm_lambs<-glm(SURV~MASS+HINDLEG+HORNLEN,data=d,family="binomial")
summary(glm_lambs)$coefficients
```

```
##                Estimate Std. Error    z value    Pr(>|z|)
## (Intercept) -5.89665169 4.38857578 -1.3436367 0.179065880
## MASS         0.58621352 0.18654017  3.1425592 0.001674778
## HINDLEG      0.02339123 0.03820471  0.6122604 0.540365476
## HORNLEN     -0.03863961 0.01428608 -2.7047035 0.006836539
```

And the expected fitness of all indivdiuals (according to the GLM), and the corresponding mean fitness of the popualtion would be calculated like this:

```
inv.logit<-function(x){exp(x)/(1+exp(x))}
exp_fit<-inv.logit(predict(glm_lambs,newdata=data.frame(MASS=d$MASS
                                          ,HINDLEG=d$HINDLEG,HORNLEN=d$HORNLEN)))
barW<-mean(exp_fit)
```

Depending on your *R* skills, you might realise that passing the `newdata` argument to the `predict()` function is unnecessary, as `predict()` uses the original data by default. But I did this for a reason. If you wanted to calculate the mean fitness of an hypothetical population where all individuals were just a bit heavier than they really were, you could do this:

```
h<-0.01 # a small numer of KG (or of mm for that matter),
        # relative to the range in the population
```

```
bar_W1<-mean(inv.logit(predict(glm_lambs,newdata=data.frame(MASS=d$MASS+h,
                                                    HINDLEG=d$HINDLEG,HORNLEN=d$HORNLEN))))
```

I'll leave it at that. Good luck!

(15) [Advanced] I claimed that multiple regression is unbiased by corrleations among predictor variables. But I only backed this up in the lecture with a single numerical example. Here is my code again, but with a third predictor variable as a bonus:

```
library(mvtnorm)
P<-matrix(c(1,0.8,-0.2,0.8,1,-0.2,-0.2,-0.2,1),3,3)
b<-c(0.5,-0.5,0.2)
n<-50
x<-rmvnorm(n,rep(0,3),P)
y<-rnorm(n,x%*%b,1)
summary(lm(y~x[,1]+x[,2]+x[,3]))$coefficients
```

```
##               Estimate Std. Error   t value    Pr(>|t|)
## (Intercept)  0.1693643  0.1338148  1.265662 0.212007229
## x[, 1]       0.7397032  0.2417403  3.059909 0.003687228
## x[, 2]      -0.5133341  0.2353002 -2.181614 0.034285078
## x[, 3]       0.2312069  0.1261479  1.832824 0.073305522
```

Those are not so close to the true values as on my slides. This is because I cranked the sample size way up (on the slides, that is) to make my point in one go. But you might be curious what the situation is for more modest sample size.

Wrap up my example into a simulation that can be run many times, and confirm (or refute) for parameter values of your own choosing whether (a) OLS estimates of linear effects are unbiased (i.e., whether $E[\hat{\beta}] = \beta$), and (b) whether standard errors of OLS regression are rendered incorrect by collinearity (determine the standard error of the estimator is reflective of the standard deviation of the esimates across multiple runs with the same underlying reality).

(15) [Advanced] It's not about multivariate selection anymore (in principle; you can build a univariate or a multivairate analysis into your answer for this question), but if you've gotten this far, you're probably finding it pretty fun. Biostatistics orthodoxy has it that regressing a response variable that will have highly non-normal residuals is criminally naive. Actually, OLS requires absolutely no assumption about the distribution of residuals to yield unbiased estimates of linear effects. Do a simulation to determine how well the OLS regression approach of Lande and Arnold recovers selection gradients, when fitness is non-normal (it might be convenient to take advantage of the fact that $\beta = b$ when $E[W] \propto e^{a+bz}$) in order to simulate non-normal fitness residuals, say from a Poisson distribution under GLM assumptions, which being able to fix the true value of β to whatever value you fancy.

Elaborations

(16) Calculate the directional selection differentials and gradients for the two episodes I used (first year viability selection, and subsequent lifetime selection of survivors), and use the Arnold-Wade-Kalisz mechanics for the partition of total selection into episodes to combine them to recover total selection.

(17a) Estimate the matrix of quadratic (including correlational) viability selection gradients for mass, leg length, and horn length in Soay ewe lambs. (You may or may not have already done this for question 12, though you may want to re-do it with variance standardisation, so it will match the lecture.)

(17b) [tricky] Visualise the two major axes of quadratic selection. It is quite a bit of work to do this in `R`.

(18) Say an ecophysiologist (I fear probably not a terribly competent one, but we'll run with it) told you that mass is almost entirely causally determined by skeletal size (for which we have a good proxy in hind leg length), and that skeletal size and mass are both likely to have direct effects on fitness. This might suggest a path model to you wherein $hind\ leg \rightarrow mass$, $hind\ leg \rightarrow survival$ & $mass \rightarrow survival$. Calculate the direct and extended sense viability selection gradients of hind leg length and mass, and compare these to the direct selection gradients (Lande's β).

(19) The file `soay_ram_lamb_selection_data.csv` contains analogous data for males – what fun! Calculate directional viability selection gradients for mass, hind leg length, and horn length. Compare these with the corresponding selection gradients for females. Consider how these selection gradients are represented under Cheng and Houle's formula for expressing sex-specific selection as concondant and antagonsitic selection gradient vectors.

(20) [Advanced] Fit a GLM of the relationship between lamb August mass of Soay ewes that survive their first winter. In combination with a GLM of viability selection (maybe simplify the one from question 14 down to using just mass as a predictor), make a model of lifetime selection of August lamb mass that is sensitive to the statistial distributions of survival and subsequent total reproductive success.