Bivariate models

Julia Schroeder

Variance-covariance analysis

Before we begin, we clear our workspace. Never forget!

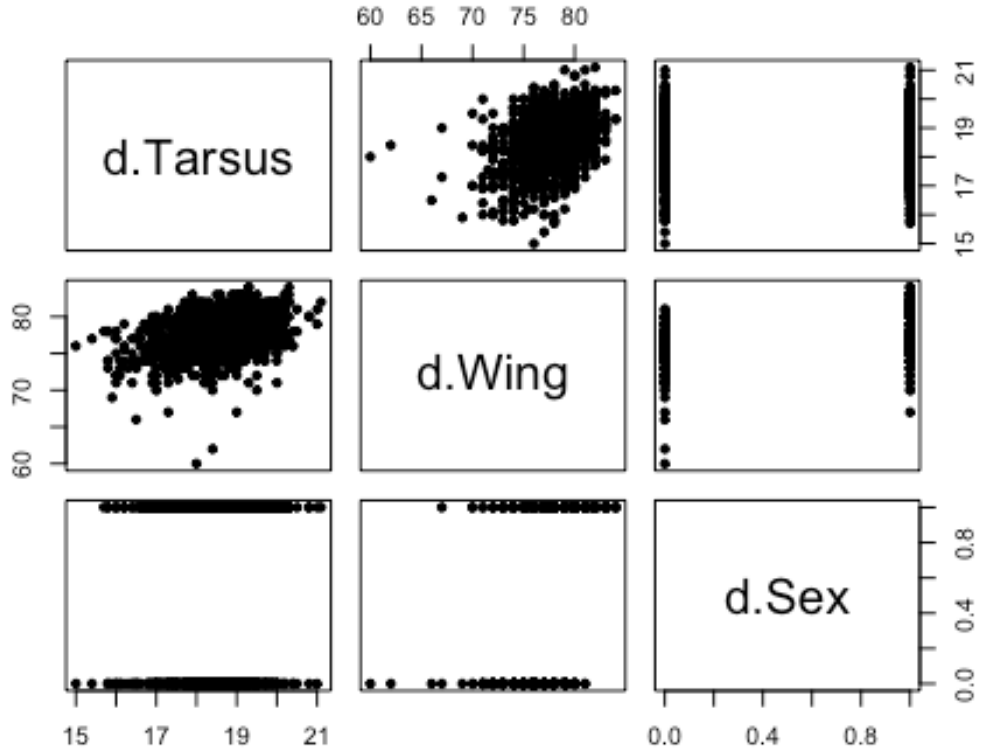
```
rm(list=ls())
setwd("~/Box Sync/Teaching/MagicStats")

d<-read.table("SparrowSize.txt", header=TRUE)
str(d)

## 'data.frame':    1770 obs. of  11 variables:
##  $ BirdID     : int  4401 4401 4405 4405 4405 4409 4409 4409 4409 4409 ...
##  $ Cohort     : int  1991 1991 1994 1994 1994 1994 1994 1994 1994 1994 ...
##  $ CaptureDate: Factor w/ 414 levels "01-Aug-06","01-Dec-07",..: 272 18
254 41 88 303 174 18 159 164 ...
##  $ CaptureTime: Factor w/ 293 levels "04:00","04:30",..: NA NA NA NA NA NA
NA NA NA NA ...
##  $ Year       : int  2000 2000 2000 2000 2000 2000 2000 2000 2001 2001 ...
##  $ Tarsus     : num  18.9 18.8 19.1 19 19.1 ...
##  $ Bill       : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ Wing       : num  82 79 77 78 77 76 76 73 79 77 ...
##  $ Mass       : num  29.4 31.6 29.9 31.6 31 ...
##  $ Sex        : int  1 1 0 0 0 1 1 1 1 1 ...
##  $ Sex.1      : Factor w/ 2 levels "female","male": 2 2 1 1 1 2 2 2 2 2
...

names(d)

##  [1] "BirdID"      "Cohort"      "CaptureDate" "CaptureTime" "Year"
##  [6] "Tarsus"      "Bill"        "Wing"        "Mass"        "Sex"
## [11] "Sex.1"

head(d)

##   BirdID Cohort CaptureDate CaptureTime Year Tarsus Bill Wing Mass Sex
## 1   4401   1991   21-Jun-00        <NA> 2000   18.9   NA   82 29.4   1
## 2   4401   1991   02-Oct-00        <NA> 2000   18.8   NA   79 31.6   1
## 3   4405   1994   20-Jun-00        <NA> 2000   19.1   NA   77 29.9   0
## 4   4405   1994   04-Oct-00        <NA> 2000   19.0   NA   78 31.6   0
## 5   4405   1994   07-Oct-00        <NA> 2000   19.1   NA   77 31.0   0
## 6   4409   1994   23-Mar-00        <NA> 2000   18.0   NA   76 28.1   1
##    Sex.1
```

```
## 1   male
## 2   male
## 3 female
## 4 female
## 5 female
## 6   male
```

The aim here is to explore variance in tarsus and wing among and between individuals and cohorts in house sparrows. The prediction is that both traits are repeatable, that means, in both traits, variation between individuals should be larger than variation within individuals. Therefore, we expect that BirdID explains a lot of variance. We expect that cohort explains some variance, too, because individuals from the same cohort are exposed to the same conditions during growing up, which is when adult size is determined. We know that we have some NAs in the data, and therefore we should remove these. We also know that sex is important in size variables, so we will want to correct for that.

To do this, we will use a range of linear models, including bivariate models.

```
dat <- d[ which( d$Tarsus!="NA" & d$Wing!="NA" & d$Sex!="NA" &
d$Cohort!="NA") , ]
# this is a different way of subsetting - we can give multiple conditions at
once, and combine them with a logical operator &, which means AND. We could
also use | which means OR.
d1<-data.frame(d$Tarsus,d$Wing,d$Sex)
pairs(d1, pch=19, cex=0.7)
```




```
cor.test(dat$Wing,dat$Tarsus)
```

```
##
##  Pearson's product-moment correlation
##
## data:  dat$Wing and dat$Tarsus
## t = 17.019, df = 1675, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.3423907 0.4240387
## sample estimates:
##      cor
## 0.383965
```

```
var(dat$Tarsus)
```

```
## [1] 0.7403658
```

```
var(dat$Wing)
```

```
## [1] 5.836133
```

```
mean(dat$Wing)
```

```
## [1] 77.39744
```

```
mean(dat$Tarsus)

## [1] 18.52501
```

We will first build the maximal model without covariance (yet) - we know we want Cohort and BirdID as random, and we know we need Sex as fixed effect. We plot the model to check if it makes any sense...

```
library(MCMCglmm)

mMaxNoCov<-MCMCglmm(cbind(Tarsus,Wing)~trait-1+trait:Sex,
random=~idh(trait):BirdID+idh(trait):Cohort, rcov=~idh(trait):units,
family=c("gaussian","gaussian"),data=dat,verbose=FALSE)

plot(mMaxNoCov)
```

You must plot the plots yourself. You can see that the two intercepts make a lot of sense - they are in the right ballpark. There seems to be a significant sex effect on tarsus, and surely one on wing. We can see that from the first plot, because the density plot does not (or marginally not in case of tarsus) overlap zero. One plot further, we can see that the birdID plots for both traits also look sensible, although a little thin, they could need a longer chain. However, the plot for cohort is horrendous, it's worse for tarsus, and a little less bad, but still bad, for wing. The residuals (units) look ok. So it's a good call to run this model a bit longer, and maybe keep in mind whether cohort is a good thing to have in here at all. We set the number of iterations to 100.000.

```
mMaxNoCov<-MCMCglmm(cbind(Tarsus,Wing)~trait-1+trait:Sex,
random=~idh(trait):BirdID+idh(trait):Cohort, rcov=~idh(trait):units,
family=c("gaussian","gaussian"),data=dat,nitt=100000, verbose=FALSE)

plot(mMaxNoCov)
```

This looks much improved, although cohort remains a worry. Maybe it is indeed a useless variable? We check the summary:

```
summary(mMaxNoCov)

##
##  Iterations = 3001:99991
##  Thinning interval  = 10
##  Sample size  = 9700
##
##  DIC: 7464.612
##
##  G-structure:  ~idh(trait):BirdID
##
##                    post.mean l-95% CI u-95% CI eff.samp
## traitTarsus.BirdID    0.6439   0.5711   0.7265     9700
## traitWing.BirdID      2.7268   2.3245   3.1744     9700
##
##                ~idh(trait):Cohort
```

```
##
##                    post.mean  l-95% CI u-95% CI eff.samp
## traitTarsus.Cohort  0.004128 2.895e-78  0.02724    152.1
## traitWing.Cohort    0.235842 2.802e-02  0.56940   8632.6
##
##  R-structure:  ~idh(trait):units
##
##                   post.mean l-95% CI u-95% CI eff.samp
## traitTarsus.units    0.0961  0.08796   0.1044     9700
## traitWing.units      1.6492  1.50782   1.7871    10750
##
##  Location effects: cbind(Tarsus, Wing) ~ trait - 1 + trait:Sex
##
##                 post.mean l-95% CI u-95% CI eff.samp  pMCMC
## traitTarsus      18.43739 18.33416 18.54271     2991 <1e-04 ***
## traitWing        76.03859 75.66287 76.44040     9817 <1e-04 ***
## traitTarsus:Sex   0.16123  0.02633  0.29106     9700 0.0159 *
## traitWing:Sex     2.71258  2.42536  3.02938     9700 <1e-04 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We focus on the cohort bit. The variance explained by cohort is nearly zero for tarsus, but present in wing. Ok, that means two things. First, it explains some variance in wing, thus should remain in the model. Second, we cannot have covariance in cohort - because is variance of wing is zero - if there is no variance in one of the variables, there cannot be covariance between them! So our full model really should not include any covariance on the cohort term. So, let's go and get the full model. We use us(trait): for birdID and the units, but idh (meaning no covariance estiamted for this term) for cohort. Because we deal with covariances now, it can take quite a long time to compute. To get an idea of how long, use *verbose=TRUE* - then you get an idea of the speed with which the iterations take place. We went for 100.000 iterations -- you can gauge yourself whether you have time enough for a coffee run!

```
mFull<-MCMCglmm(cbind(Tarsus,Wing)~trait-1+trait:Sex,
random=~us(trait):BirdID+idh(trait):Cohort, rcov=~us(trait):units,
family=c("gaussian","gaussian"),data=dat,nitt=100000, verbose=TRUE)

plot(mFull)
```

Now, these are plots we are very happy with! Let's check the autocorrelation just to make sure that we're on the right track: First for the fixed effects (*Sol* is for solutions - or at least that's how I try to remember these odd terminiology)

```
autocorr(mFull$Sol)
```

```
## , , traitTarsus
##
##          traitTarsus    traitWing traitTarsus:Sex traitWing:Sex
## Lag 0    1.000000000  0.254254105   -0.7170948699  -0.310443874
## Lag 10   0.006041820  0.009390245   -0.0038679663  -0.010538286
```

```
## Lag 50   0.007975677  0.005968396    0.0012921141   0.006527112
## Lag 100 -0.008307459 -0.005175334    0.0102711253   0.014655733
## Lag 500 -0.001620455  0.001659957    0.0009680898   0.011817889
##
## , , traitWing
##
##          traitTarsus    traitWing traitTarsus:Sex traitWing:Sex
## Lag 0    0.254254105  1.000000000   -1.781735e-01  -0.430760064
## Lag 10   0.018275763 -0.006646763   -2.007974e-06   0.006744509
## Lag 50  -0.001265081  0.016297306    3.533510e-03  -0.013294669
## Lag 100 -0.013279741 -0.006719287    2.427686e-02   0.023182169
## Lag 500 -0.003737914 -0.002582290    5.372697e-04  -0.006643450
##
## , , traitTarsus:Sex
##
##         traitTarsus    traitWing traitTarsus:Sex traitWing:Sex
## Lag 0   -0.71709487 -0.178173465    1.0000000000  0.4273040403
## Lag 10   0.01048924 -0.006450509    0.0026127231  0.0021792602
## Lag 50  -0.01590328 -0.009830962    0.0083097991  0.0101015447
## Lag 100  0.01437127  0.008642206   -0.0075545538 -0.0046157389
## Lag 500 -0.00183839  0.001801019    0.0002591423 -0.0004535829
##
## , , traitWing:Sex
##
##          traitTarsus    traitWing traitTarsus:Sex traitWing:Sex
## Lag 0   -0.310443874 -0.430760064     0.427304040   1.000000000
## Lag 10  -0.010612910 -0.005122557     0.003595326  -0.006111817
## Lag 50  -0.018140289 -0.021966333     0.008904181   0.025292052
## Lag 100  0.003689099 -0.009429466    -0.008621132  -0.016150265
## Lag 500  0.002988202  0.005936441    -0.013222976   0.006526624
```

And then for the random effects and residual structure (*VCV* stands for **V**ariance**C**o**V**ariance structure). I didn't print them as they took up too much space.

```
autocorr(mFull$VCV)
```

Both are looking good! Remember, we want super low, uncorrelated values for lag10 and higher.

Now it is time to look at the summary, and interpret the results. Let's go step-by-step, and start with the fixed effects, which are on the bottom of the summary, and called *Location effect* (and I have no clue why). But at least we're given the fixed effect structure, which is nice.

```
summary(mFull)
```

```
##
##  Location effects: cbind(Tarsus, Wing) ~ trait - 1 + trait:Sex
##
##                 post.mean l-95% CI u-95% CI eff.samp  pMCMC
## traitTarsus       18.4326  18.3346  18.5220     9700 <1e-04 ***
```

```
## traitWing         76.0025  75.6554  76.3585     9700 <1e-04 ***
## traitTarsus:Sex    0.1601   0.0303   0.2978     9700 0.0196 *
## traitWing:Sex      2.6865   2.3864   2.9880     8674 <1e-04 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

So, we can see, we added the term *trait-1*, which is stand-in for estimate intercepts for each variable separately (you can also only get one intercept, but that would not make sense if you have two variables that mean different things). Then we wanted a sex effect, also seperately for each variable.

The intercepts make a lot of sense and are fairly similar to the grand-total means of tarsus and wing. They are statistically significantly different from zero, but that is no surprise. The sex effect in tarsus is 0.16, so males are a bit larger than females. Males also have a longer wing than females. For both effects, the 95CI do not span zero, that means they are statistically significantly different from zero. There is a p-value that we can use, although it's not the frequentist one, but one estimated by MCMC procedures. It should not bother us too much. The 95CI are more important anyways.

Now let's have a look at the random effects:

```
summary(mFull)
```

```
##
##  Iterations = 3001:99991
##  Thinning interval  = 10
##  Sample size  = 9700
##
##  DIC: 7408.255
##
##  G-structure:  ~us(trait):BirdID
##
##                                post.mean l-95% CI u-95% CI eff.samp
## traitTarsus:traitTarsus.BirdID    0.6475   0.5697   0.7261     9700
## traitWing:traitTarsus.BirdID      0.6486   0.5128   0.7879    10154
## traitTarsus:traitWing.BirdID      0.6486   0.5128   0.7879    10154
## traitWing:traitWing.BirdID        2.7047   2.2971   3.1368     9700
```

First, we get a G-structure (and don't bother about the name - it derives from quantitative genetics). The g-struture gives us information about the random effects BirdID and Cohort. For BirdID we get four parameter estimates - one each for the variance of tarsus and wing, respectively, and two for the covariance. The covariance is shown twice because of the matrix notation - it reminds us that the variance-covariance structure is a matrix and the covariance shows up twice in it - even if it's the same value (0.65). The covariance is statisitcally significantly differnt from zero - we know that even if we can't have a p-value for it because the 95CI does not span zero. For the variances (first and last line of G-structure BirdID), it is not so easy because variances can't be negative, they are always positive. But for both, wing and tarsus, the lower 95CI is a far way from zero and narrow, so

we are confident that these variances are significant. For both traits, there is quite some variance explained by BirdID.

```
##
##                ~idh(trait):Cohort
##
##                     post.mean   l-95% CI  u-95% CI eff.samp
## traitTarsus.Cohort 5.116e-15 1.715e-139 1.079e-31        0
## traitWing.Cohort   1.775e-01  2.385e-02 4.257e-01     9700
```

Now we move on to cohort, which only has two parameter esitmates, that is because we told R to not estmate the covariance. It is clear from the estimates that the variance in tarsus explained by cohort is zero. Even the upper 95CI is pretty much zero. So our assumption was correct. However, there is a bit more variance in wing explained by cohort, but that is also close to zero. Maybe we should run another model without cohort, and see whehter the model fit improves. If you feel confident that you can do that by yourself, you can start that model right now, then the model can run while you read the rest. I'll give you the code further below.

```
##
##  R-structure:  ~us(trait):units
##
##                                post.mean l-95% CI u-95% CI eff.samp
## traitTarsus:traitTarsus.units   0.09623  0.08819  0.10457    10223
## traitWing:traitTarsus.units     0.05993  0.03483  0.08409     9700
## traitTarsus:traitWing.units     0.05993  0.03483  0.08409     9700
## traitWing:traitWing.units       1.65895  1.51692  1.80597     9700
```

But first we'll continue interpreting the results. The R-structure is missing. We see that there is some residual variance left in both traits, although especially in tarsus, it's not a lot. However, there is also significantly positive residual covariance between wing and tarsus (95CI not overlapping zero). So that's ok.

It is a bit confusing that in some cases, we can use the overlap of the 95CIs with zero for assessing statistical significance (fixed parameter estimates, covariances), and sometimes, we can't (variances). It makes sense when you think about whether or not a parameter can go below zero. Variances cannot be negative. So we always need to think when we interpret model results - there are no *one-size-fits-all* rules.

Now to the reduced model - we removed cohort:

```
m2<-MCMCglmm(cbind(Tarsus,Wing)~trait-1+trait:Sex, random=~us(trait):BirdID,
rcov=~us(trait):units, family=c("gaussian","gaussian"),data=dat,nitt=100000,
verbose=FALSE)

mFull$DIC

## [1] 7408.255
```

```
m2$DIC
```

```
## [1] 7415.47
```

From the DIC, it seems as if it's a good idea to keep cohort in, because lower DICs are beter, and 7408 is lower than 7415. So we'll keep the full model as our final model. We now know that there is quite strong covariance between birds (the BirdID covariance), which makes sense as it means that birds with longer wings have longer tarsi, too. There is however some covariance, also positive, in the residuals. That means that within birds, when they have longer tarsi they also have longer wings. This may come from two sources - first, maybe not all measured birds were fully grown yet, and thus, within birds there might be some variation (even though it's not a lot) that covaries. Second, it could also be that this stems from different observers - one observer who might always measure too long would measure a wing and tarsus of the same bird longer than another observer.