The Method of Support

Key

May 10 2022

We will work with data on the brain size/body size relationship in a selection of mammals, some of which are primates, along with a small number of dinosaurs. We will fit a variety of models that represent hypotheses about the species that share a common scaling relationship, and will use the Method of Support to decide which model is best supported by the data.

Import the data into a dataset called brains:

```
library(readxl)
read_excel("brain_body.xls") -> brains
```

```
## Registered S3 methods overwritten by 'tibble':
##   method     from
##   format.tbl pillar
##   print.tbl  pillar
```

Make a model list object:

```
models.list <- list()
```

Assign the first model to models.list, a simple linear regression of log.brain on log.body:

```
models.list$body.lm <- lm(log.brain ~ log.body, data = brains)
```

Fit the remaining models (2 through 10):

```
models.list$taxa.lm <- lm(log.brain ~ log.body + Taxa, data = brains)
models.list$taxa.int.lm <- lm(log.brain ~ log.body * Taxa, data =
brains)
models.list$dino.lm <- lm(log.brain ~ log.body + Dino.nodino, data =
brains)
models.list$dino.int.lm <- lm(log.brain ~ log.body * Dino.nodino, data
= brains)
models.list$primate.lm <- lm(log.brain ~ log.body + Primate.noprimate,
data = brains)
models.list$primate.int.lm <- lm(log.brain ~ log.body *
Primate.noprimate, data = brains)
```

```
models.list$dpo.lm <- lm(log.brain ~ log.body + Dino.prim.other, data
= brains)
models.list$dpo.int.lm <- lm(log.brain ~ log.body * Dino.prim.other,
data = brains)
models.list$intercept.only.lm <- lm(log.brain ~ 1, data = brains)
```

Check that all the models are there in the model.list:

```
names(models.list)
```

```
##  [1] "body.lm"           "taxa.lm"           "taxa.int.lm"
##  [4] "dino.lm"           "dino.int.lm"       "primate.lm"
##  [7] "primate.int.lm"    "dpo.lm"            "dpo.int.lm"
## [10] "intercept.only.lm"
```

Question: Based on the graphs that pop up for each model, which grouping looked like it would account for the greatest amount of variation among species (and would thus have the highest likelihood)?

> Your call, but for me it was taxa.int.lm, because it had the greatest number of lines, each with their own slopes and intercepts, which let it come as close as possible to all the points.

Question: Do you see any possible problems with fitting a separate line for each species grouping (particular when you color by Taxa)? Specifically focus on sample size issues - do we really have the sample sizes to get robust estimates of slopes and intercepts for all taxa?

> There are groups that only have two species in them. Any two points define a line, but such a small number of points is a poor basis for drawing any conclusions about differences in general between taxa.

Extract the AIC statistics from the models:

```
data.frame(t(sapply(models.list, extractAIC))) -> model.aic
```

Change the column names to something better than X1 and X2 - they are the number of parameters (K) and the AIC values:

```
colnames(model.aic) <- c("K","AIC")
```

Calculate the second-order AIC, or AICc:

```
model.aic$AICc <- with(model.aic, AIC + 2*K*(K+1)/(26-K-1))
```

Calculate the delta AICc values:

```
model.aic$dAICc <- model.aic$AICc - min(model.aic$AICc)
```

Calculate the AICc weights:

```
model.aic$wts <- with(model.aic, exp(-0.5*dAICc)/sum(exp(-0.5*dAICc)))
```

Sort the output by dAICc, and display without scientific notation to make it easier to compare the weights:

```
format(model.aic[order(model.aic$dAICc),], digits = 2, scientific = F)
```

```
##                      K   AIC  AICc dAICc                   wts
## dpo.lm               4 -76.3 -74.4   0.0 0.6252493597029067374
## dpo.int.lm           6 -77.2 -72.8   1.6 0.2788778324035202094
## taxa.lm              8 -79.1 -70.7   3.8 0.0957303581929313807
## dino.lm              3 -58.2 -57.1  17.4 0.0001063172432656724
## dino.int.lm          4 -56.8 -54.9  19.5 0.0000361320928347135
## taxa.int.lm         14 -70.1 -31.9  42.5 0.0000000003622611060
## primate.lm           3 -22.2 -21.1  53.3 0.0000000000016731133
## primate.int.lm       4 -20.2 -18.3  56.1 0.0000000000004115334
## body.lm              2 -17.4 -16.8  57.6 0.0000000000001955819
## intercept.only.lm    1   4.4   4.6  79.0 0.0000000000000000043
```

Question: Interpret models with ΔAIC_c less than 4. What do they have in common, and how are they different? Do they include the same predictors?

> There are three, dpo.lm, dpo.in.lm, and taxa.lm. The only difference between dpo.lm and dpo.int.lm is that dpo.int.lm includes an interaction between log.body and Dino.prim.other, but both include log.body and Dino.prim.other. So, the uncertainty is about whether each of these three groups should have lines with different slopes, or instead all have the same slope but with different intercepts. The taxa.lm model has dinosaurs and primates separated, like the dpo models do, but also splits the remaining mammal taxa. According to the taxa.lm model these different mammal taxa also differ in their intercepts, and there is modest support for that hypothesis. That dinosaurs and primates should be separated from other mammals is very well supported, though, since all of the best-supported models have this feature, and none of the others do (aside from taxa.int.lm, which is poorly supported due in large part to the fact that it requires 14 parameters, and is thus the most complex model in the set).

Question: what do the AIC weights add to your interpretation of the results, compared with using just the AIC_c values alone?

The AIC weights give you a number that is interpretable in a more intuitive way than the ΔAIC_c values. While we know that a ΔAIC_c value of less than 4 indicates a model that is plausible, and should be interpreted, we don't know how a ΔAIC_c of 1.6 compares with a ΔAIC_c of 0 or 3.7. The weights tell us that if we repeated the analysis with a new data set, dpo.lm would be selected as best 62.5% of the time, whereas dpo.int.lm would be selected as best 27.8% of the time. It helps us better understand the degree of uncertainty in our conclusions.

To calculate the weights for each variable first make vectors indicating which models each variable is included in:

```
log.body <- c(1,1,1,1,1,1,1,1,1,0)
taxa <- c(0,1,1,0,0,0,0,0,0,0)
dino <- c(0,0,0,1,1,0,0,0,0,0)
primate <- c(0,0,0,0,0,1,1,0,0,0)
dpo <- c(0,0,0,0,0,0,0,1,1,0)
```

Make a data frame from these vectors:

```
data.frame(log.body, taxa, dino, primate, dpo) -> variables.in.model
row.names(variables.in.model) <- names(models.list)
variables.in.model
```

```
##                   log.body taxa dino primate dpo
## body.lm                  1    0    0       0   0
## taxa.lm                  1    1    0       0   0
## taxa.int.lm              1    1    0       0   0
## dino.lm                  1    0    1       0   0
## dino.int.lm              1    0    1       0   0
## primate.lm               1    0    0       1   0
## primate.int.lm           1    0    0       1   0
## dpo.lm                   1    0    0       0   1
## dpo.int.lm               1    0    0       0   1
## intercept.only.lm        0    0    0       0   0
```

Multiply the weights in model.aic by these columns of 0's and 1's, and sum the products to get the weights for each variable:

```
format(colSums(model.aic$wts * variables.in.model), scientific = F)
```

```
##               log.body                   taxa
dino
## "1.000000000000000000" "0.095730358555192491"
```

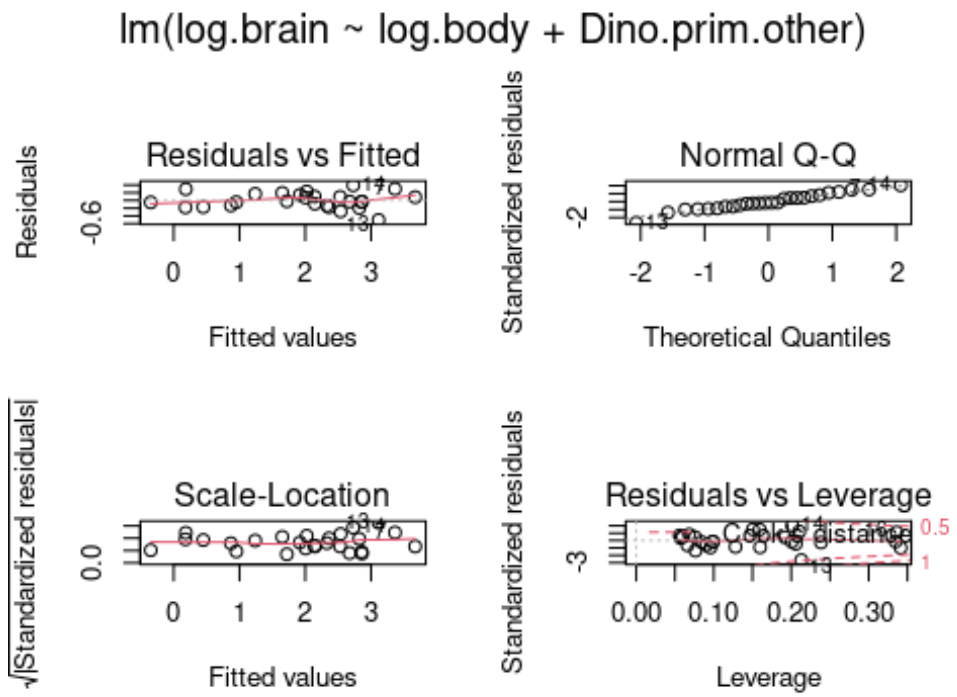
```
"0.000142449336100386"
##                  primate                      dpo
## "0.000000000002084647" "0.904127192106426891"
```

Question: which variables have the highest support and which have the lowest? Do any of the variables with high support appear in models with low support (which)?

> The variable with the higest support is log.body, and the one with the lowest support is primate. Note that splitting primates only has very low support, and splitting dinosaurs only has low support, but splitting primates and dinosaurs from non-primate mammals has very high support (0.90).

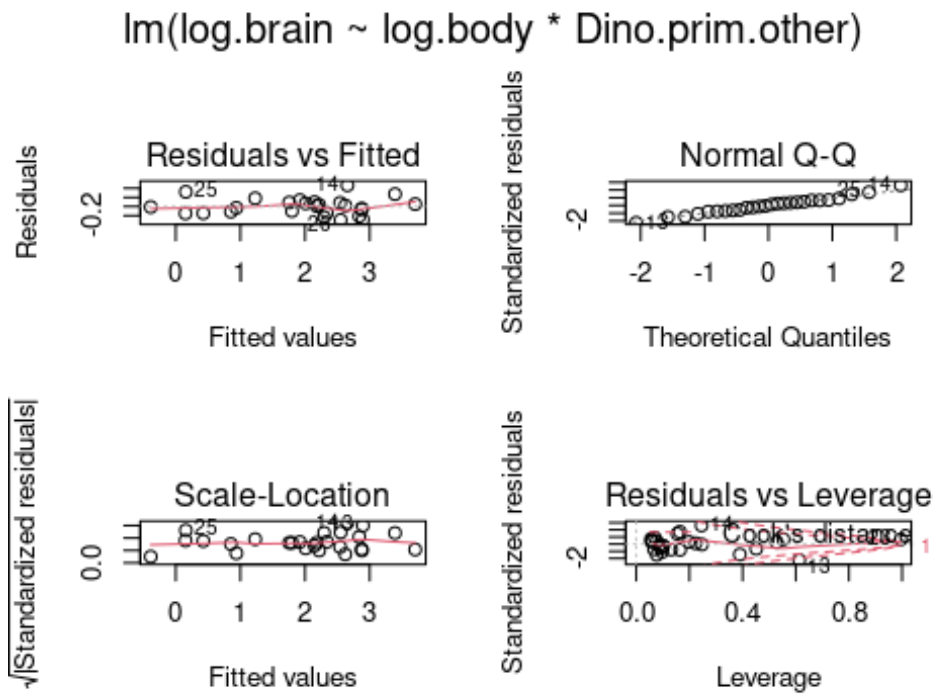
Check that the best-supported model meets GLM assumptions:

```
par(oma = c(0,0,3,0), mfrow = c(2,2))

plot(models.list$dpo.lm)
```




```
plot(models.list$dpo.int.lm)
```

```
## Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced

## Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced
```



Make sure that the best model is also a good model by extracting the R^2 for each model:

```
sapply(models.list, FUN = function(x) summary(x)$r.squared)

##           body.lm           taxa.lm       taxa.int.lm
dino.lm
##         0.5995125         0.9765490         0.9790453
0.9227799
##       dino.int.lm       primate.lm    primate.int.lm
dpo.lm
##         0.9247020         0.6924136         0.6925171
0.9644556
##       dpo.int.lm intercept.only.lm
##         0.9705627         0.0000000
```

Question: How do we know that the best supported model is any good at all? How do we know that it's better than random chance (that is, do we have a null hypothesis that we can reject with this approach)?

The R^2 for the best-supported models is 0.92 for dino.lm, 0.92 for dino.int.lm, and 0.98 for taxa.lm. These are all very close to the maximum amount of explainable variation, so they are good predictors. We know that the models are better than chance because the support for the intercept only model was very poor (it had the biggest ΔAIC_c at 79.0).