Animal Behaviour: normally distributed? Some use of graphics and statistics in animal behaviour studies

Lorenz Gygax
Department of Applied Mathematics
University of Zürich-Irchel
Winterthurerstr. 190
CH-8057 Zürich / Switzerland

June 6, 1996

Introduction

Leafing through the most recent issues of Animal Behaviour it struck me that almost all problems of parametric statistics are addressed in one paper or another, but only very rarely all of them in one. The same points pose problems to peolpe whom I advice on statistics. This and the fact that I am an ethologist by training is why I would like to make some commentaries on the use of statistics in studies of animal behaviour.

First of all I would like to recommend to summarize data at the start of an evalution as little as possible. The smallest units that can be reasonably assumed to be independent should at least be kept (summaries per individual, per time period, per observation, per experiment). Sticking to these smallest units will often help to think about how to display the data and what kind of statistics are possible to use.

These units will define the variability of our data which is basic to statistics. It will often help to see what the replicates of an experiment could be, and thus help to find a test. Often one can also see a way of testing a result avoiding a χ^2-test, which is very sensitive to statistically

dependent data and should be avoided whenever possible (lit ?).

In principal the replicates (or the smallest observational units) should be statistically independent of each other. This is often not possible in studies of animal behaviour. Thus we can either use matched tests where e. g. different individuals are tested in all possible conditions or we have to assume that our observational units are independent eventhough the same animals are observed several times (pseudo-replication). In the latter the generality of the results is questionable and the results might be specific to the observed set and might not easily be extrapolated to other individuals or settings.

A general point is that in many papers dozens of tests are conducted. We should be aware that with a probability of an error $\alpha = 0.05$ we will find one significant result within 20 tests by pure chance. Thus we should either focus on a few important questions while testing or adjust the α to a more conservative value (at least mentally, i. e. only cautiously interpret results relying on a few significant test when many have been conducted).

Graphics

The first step of a good evaluation should always be a thorough graphical investigation of the data. A good visualization, graphics display, of the data is often more convincing and intuitive to the reader than the actual statistical result. To illustrate data, mean and standard deviation are often given in figures and tables (where sometimes even raw data is shown) without the claim of the data being normally distributed. But only in the latter case can these measures describe the data adequately.

There is one very condensed form to visualize the true distribution of a data set: boxplots (Fig. 1 C–F). Boxplots show the lower and upper quartile in form of a box, the median as a straight line within the box and the range of the data with bars extending from the box. These

bars are often restricted to extend no farther than 1.5 times the inter-quartile range from the median. All data that extend farther are treated as extreme values and shown as individual dots or horizontal bars (Fig. 1 D–E).

Even if different samples have the same mean and standard deviation (Fig. 1 A, B) the boxplots show the differences in the distribution of a normally distributed random sample, a sample with extreme values (or long tails), or a sample with a skew distribution (Fig. 1 C–E). The latter can be brought to resemble a normal distribution by taking the log (Fig. 1 F, see below).

Non-parametric tests

If there are just two groups which one would like to compare it is strongly advised to use the non-parametric Wilcoxon test for matched samples (i. e. individuals are all tested in two conditions) and the Man-Whitney-U test for non-matched samples (two different groups have been tested in the two conditions). These tests correspond to the dependent and independent t-test but do not assume any specific distribution of the data. The statistical power of these tests is almost as good as the one of the t-test (Lehmann, 1975 or ?) thus even if the data are normally distributed the non-parametric tests are almost as likely to pick up a difference as the t-test. If the data is not normally distributed or if the sample is small, and thus it is difficult to decide whether the data is normally distributed (see below), we conduct a wrong analysis if we use the t-test. Both these restrictions (non–normal distribution and/or small sample size) often occur in the study of animal behaviour.

It is true that the t-test can still be conducted and yield a significant result with a sample so small that a non-parametric test can not be statistically significant anymore. This use has to be questioned, however, because again we can not decide whether the data is normally distributed

(see below) and it is dubious to conduct a statistical test with such few data in general.

The strangest thing in this regard are papers that use both t-tests AND their non-parametric equivalents; depending on whether they give the expected result or not?

The same is true for the simplest cases of analyses of variance (ANOVA) to test differences among more than two groups: the non-parametric equivalents are available as the Friedman test for matched and the Kruskal-Wallis test for non-matched data. In the realm of non-parametric tests one is even able to test for a monotonous trend between the groups: The Page–test for matched and the Jonkhere–test for non-matched data. It is advisable to test a data set first with a Kruskal–Wallis or a Friedman and only after establishing a significant difference with these "two–sided" tests use the "one–sided" Page or Jonkheere to test for the trend.

Parametric statistics: theory

If one wants to include more than one variable (multiple and multivariate tests) we can only rely on parametric statistics to date. Parametric statistics have, as mentioned already, something to do with the normal distribution. But what exactly is normally distributed?

Almost all parametric tests (all linear models) can be written in the form $y_i = \alpha + \beta^{(1)} x_i^{(1)} + \beta^{(2)} x_i^{(2)} + \cdots + \beta^n x_i^{(n)} + \epsilon_i$, where the ϵ_i are assumed to be random errors independently taken from a normal distribution with mean 0 and some sample-variance σ^2. This is the same as saying that the data split according to all variables is normally distributed. But if we split our data like that we usually get very small samples for each class of data and find it thus difficult to judge whether the data are normally distributed. It is easier to work with all the errors at the same time in an analysis of residuals (see below). The errors are estimated by the residuals, i. e. by the difference of the expected (based on the estimated coefficients a and $b^{(j)}$) and the observed values of our response variable. Before we can have a look at the distribution of our

residuals we have to conduct a first attempt of a statistical evaluation.

The first step to do so is the transformations of the response and the predictive variables. Tukey recommends the following "first–aid transformations" which should always be used if there are no theoretical considerations oposing it (as e. g. when a formula underlying a phenomenon is known):

absolute values (time, concentrations, etc.): $\tilde{y} = log(y)$

counts: $\tilde{y} = \sqrt{y}$

proportions (percentages/100): $\tilde{y} = arcsin(\sqrt{y})$

The next step is a first round of evaluations. With the nowadays available computer power it is recommended to start out with a model including all variables (and their interactions) that might be interesting and then follow a step-wise backwards elimination strategy, i. e. dropping the most nonsiginificant variable one at the time, with the restriction that we shoudn't drop simple terms if those variables are involved in statistically significant interactions as well. Thus we end up with a preliminary model. Now we have to have a look at the residuals and might see that they deviate from normality, this can often be helped with further transformation of the data, the inclusion of other predictive variables (e. g. the power of a considered variable), or the exclusion of a few extreme values or a group of values which behaves differently than the rest of the data. Then a new circle of reduction and residual analysis starts. Interacions appear if two variables influence the response variable in a non-additive way. One can imagine groups of animals that react differently to the increase of a variable (see example below).

Now I finally come to the mystic analysis of residuals (= estimate of errors) of which we assumed that they were independently sampled from one specific normal distribution with mean zero and variance σ^2.

It is rather difficult to test a sample for normality in a statistical sense because statistical tests are designed to reject a hypothesis and can not really proove a hypothesis to be true. Thus if we test a sample for normality and can't reject the zero hypothesis, we can only say that the deviations from normality are not big enough to be picked up by the test for the given sample size. Thus one often gives up formal statistical testing and uses graphical methods instead.

In the "normal–plot" the residuals are drawn against the corresponding quantiles of a normal distribution. If data are normally distributed a straight line results in this plot (Fig. 2 A, D). Different aberrations can easly be detected in the plot: long tails (outliers) or a squew distribution (Fig. 2 B, C). These deviations from a normal distribution can be dealt with by exculding (a few) outliers from analysis, transformation of data and/or including other variables or derivates of already included variables (e. g. a variable to the power of two). With this plot we can thus see whether our data is normally distributed. This is easier seen in this plot than in a histogramm because deviations from a straight line can be more easily perceived by the human eye (Fig. 2, 3).

In the "Tukey–Anscombe–plot" we draw the residuals against the estimated values of our response variable. In this plot one can check, whether the number of positie and the number of negative residuals is about the same, whether the positive and the negative residuals are of about the same size and whether they are the same size along the axis of the estimated values. Especially outliers and a distribution that asks for a log transformation can be detected in this plot (in the latter case the residuals grow with growing estimate).

The leverage (or the Mahanalobis distance) is a measure of how much a single data point influences the whole analysis. Thus we don't want to include points in our analysis with a big leverage and a big residual at the same time, because these are points that heavily influence the outcome of the evluation. These points can be detected by plotting the residuals against the leverages.

Finally it is advisable to plot the residuals against all the predictive variables and against time if data was gathered in a certain temporal sequence. There should be no pattern in the residuals along the axis of the predictive variable and the variance along the axis should be constant. If possible a plot of the residuals against two predictive variables can give insight into interactions of two variables that have not yet been considered.

A good way not to overinterpret the data is to run a couple of simulations, i. e. use the formula of the statistical evaluation and add some independent identically normally distributed errors, to see what patterns of residuals might look like by pure chance.

Parametric statistics: an illustrative example

To make the above mentioned points clearer. I will now present a sample evaluation: a multiple regression problem with some simulated data. Imagine we are interested in the number of warning calls per time unit an individual of a species gives in a situation we have defined as being dangerous. We would like to bring the number of calls into relation with the sex (SEX) and age (AGE) of the caller and on the number of animals present (GROUP).

All evaluations and plots in [illegible]

sigificant factors. In this example we end up with the following situation:

```
Coefficients:
              Value Std. Error t value Pr(>|t|)
(Intercept)  0.4646  0.1749     2.6560  0.0093
        SEX  0.6079  0.1538     3.9536  0.0002
        AGE -0.0488  0.0163    -2.9980  0.0035
      GROUP -0.0518  0.0197    -2.6298  0.0100
    SEX:AGE -0.0275  0.0095    -2.8980  0.0047
  SEX:GROUP -0.0341  0.0156    -2.1852  0.0314
  AGE:GROUP  0.0056  0.0018     3.1082  0.0025

Residual standard error: 0.1859 on 93 degrees of freedom
Multiple R-Squared: 0.4107
F-statistic: 10.8 on 6 and 93 degrees of freedom, the p-value is 4.325e-09
```

Except for the small multiple R-squared, which is a measure of how well the data fit to the model, and which is usually quite low in biological problems, all variables and their interactions seem to have a significant influence and appear in all possible interactions.

If we have a look at the residuals we can immediately see that something must be wrong (Fig. 5). Especially the strongly downwards sloped curve in the normal–plot, the structure in the Tukey–Anscombe–plot and the explosion in the size of the residuals towards lager fitted values tells us that we need to transform our data.

Thus in the next step we apply the Tukey–first–aid transformations. I assume here that the number of calls and the age is known rather exactly, i. e. that they can be considered absolute continuous variables, thus we log–transform them (lAGE). The number of animals in a group is a count and thus square–root–transformed (sGROUP).

Again we run the model and exclude the non–significant variables. We find:

```
Coefficients:
              Value Std. Error t value Pr(>|t|)
(Intercept)  9.4509  1.8037     5.2398  0.0000
        SEX  1.8941  0.4946     3.8296  0.0002
       lAGE -3.7116  0.3807    -9.7488  0.0000
     sGROUP -3.4604  0.5582    -6.1987  0.0000

Residual standard error: 2.463 on 96 degrees of freedom
Multiple R-Squared: 0.6003
```

```
F-statistic: 48.06 on 3 and 96 degrees of freedom, the p-value is 0
```

Only the single variables are significant in this approach, but again the residuals show us, that our residuals are not idenpendently normally distributed. In the normal plot (sloped to the top), the Tukey–Anscombe plot and the residuals against age, we see that there seems to be a quadratic relation between number of calls and age (Fig. 6). Thus we include the age by the power of two in the next step of the analysis and run the model again:

```
Coefficients:
               Value Std. Error  t value Pr(>|t|)
(Intercept)  -5.0691   0.6664    -7.6071   0.0000
        SEX   2.0674   0.0685    30.1675   0.0000
       lAGE  14.9191   0.3827    38.9836   0.0000
      lAGE2  -5.0151   0.0727   -69.0242   0.0000
     sGROUP  -2.0020   0.2314    -8.6518   0.0000
lAGE:sGROUP  -0.9397   0.1125    -8.3551   0.0000

Residual standard error: 0.3408 on 94 degrees of freedom
Multiple R-Squared: 0.9925
F-statistic: 2489 on 5 and 94 degrees of freedom, the p-value is 0
```

In the plots of the residuals we see now, that there are two outliers (Fig. 7). And going back to the "original data" we realize, that we made a mistake in transfering the data, in that we have included the tenfould values for the number of calls that we should have. After correction of these values we run the model once more:

```
Coefficients:
                Value Std. Error   t value  Pr(>|t|)
(Intercept)   -5.2000    0.1796   -28.9468    0.0000
        SEX    1.9718    0.0185   106.7298    0.0000
       lAGE   15.0766    0.1032   146.1353    0.0000
      lAGE2   -5.0061    0.0196  -255.5806    0.0000
     sGROUP   -1.9236    0.0624   -30.8368    0.0000
lAGE:sGROUP   -1.0198    0.0303   -33.6335    0.0000

Residual standard error: 0.09188 on 94 degrees of freedom
Multiple R-Squared: 0.9995
F-statistic: 34520 on 5 and 94 degrees of freedom, the p-value is 0
```

Now even the plots of the residuals are satisfying as they do not largely deviate from what we expect them to look (Fig. 8). The points in the normal–plot lie on a line, they are randomly

scattered to both sides of zero along the axis of the fitted values and the variables, and there are no points in the leverage–plot with large residual and large leverage at the same time.

In Fig. 9 we see an example of how the residual pattern looks like if would have neglected the inteaction between age and group. We can clearly see that neighbouring residuals are likely to be similar; a clear hint on that they are not independent and there is an underlying interaction of lAGE and sGROUP.

The interpretation of this model would then be that females are more likely to call as a warning to others (as in matrifocal species), that middle aged individuals are giving more calls than others (those could be the ones most likely to have young offspring), that the number of calls decreases if the group size increases, and that the older individuals adjust better to group size than younger individuals (the interaction).

Final remark

In conclusion, I hope that I could show what strategies to follow if one has to rely on parametric statistics in the study of animal behaviour. Unfortunately with these tests the work is not done when one has received a p-value with the help of some computer software. But often the detailed occupation with ones data, e. g. during analysis of residuals can lead to new ideas and insights in a data set. Another approach is to use so called robuste statistics, but those are still hardly implemented in computer programs thus I don't want to further comment on them for the moment, but they are an important tool in dealing with random distributions that are slightly disturbed.

Acknowledgement

I would like to thank ... for critical suggestions and ... for clarifying my English. The ideas in this paper have been largely influenced by the scientist working at the Seminar for Statistics (Federal Institute of Technology) and at the Department of Applied Mathematics (University of Zürich). Some of the functions used in Splus have been written by W. Stahel and M. Maechler from the Seminar for Statistics (Federal Institute of Technology). This work has been supported by grant No. ... from the Swiss National Science Foundation to Prof. A. D. Barbour.

Literature

More general descriptions of how to proceede in applied statistics can be found in Stahel (1995), somthing similar in English?; specific information on transformation of data in ?; and information on the analysis of residuals in Gunst and Mason (1980), ?.

Cox, D. R. and Snell, e. J. 1981. Applied statistics, Chapman and Hall, London.

Gunst, R. F. and Mason, R. L. 1980. Regression analysis and its application: a data–oriented approach. Marcel Dekker Inc., New York.

Lehmann, E. L. Nonparametrics: statistical methods based on ranks, Holden-Day.

Stahel, Werner A. 1995. Statistische Datenanalyse; Eine Einführung für Naturwissenschaftler. vieweg, Braunschweig/Wiesbaden.

Figures

Fig. 1 Comparison of different graphical ways to display the distribution of data. All samples have means equal to zero and variance equal to one (in this case also the standard deviation is one) as seen in (A) mean $\pm$ std. dev. and (B) mean and std. dev. (C) boxplot of normally distributed random numbers, (D) boxplot of a distribution with outliers, (E) boxplot of a skew distribution, (F) boxplot of (E) after taking the log (of the data plus a constant) and normalisation.

Fig. 2 Normal–plots of the distributions shown in Fig. 1 (C) - (F).

Fig. 3 Histograms of the distributions shown in Fig. 1 (C) - (F).

Fig. 4 Number of calls per unit time in dependence of sex (top left), age (bottom left) and number of animals present (bottom right) for the simulated data set. Young animals are given by open circles, old ones by black squares. There seems to be no correlation between the age of an individual and the group size it is found in (top right).

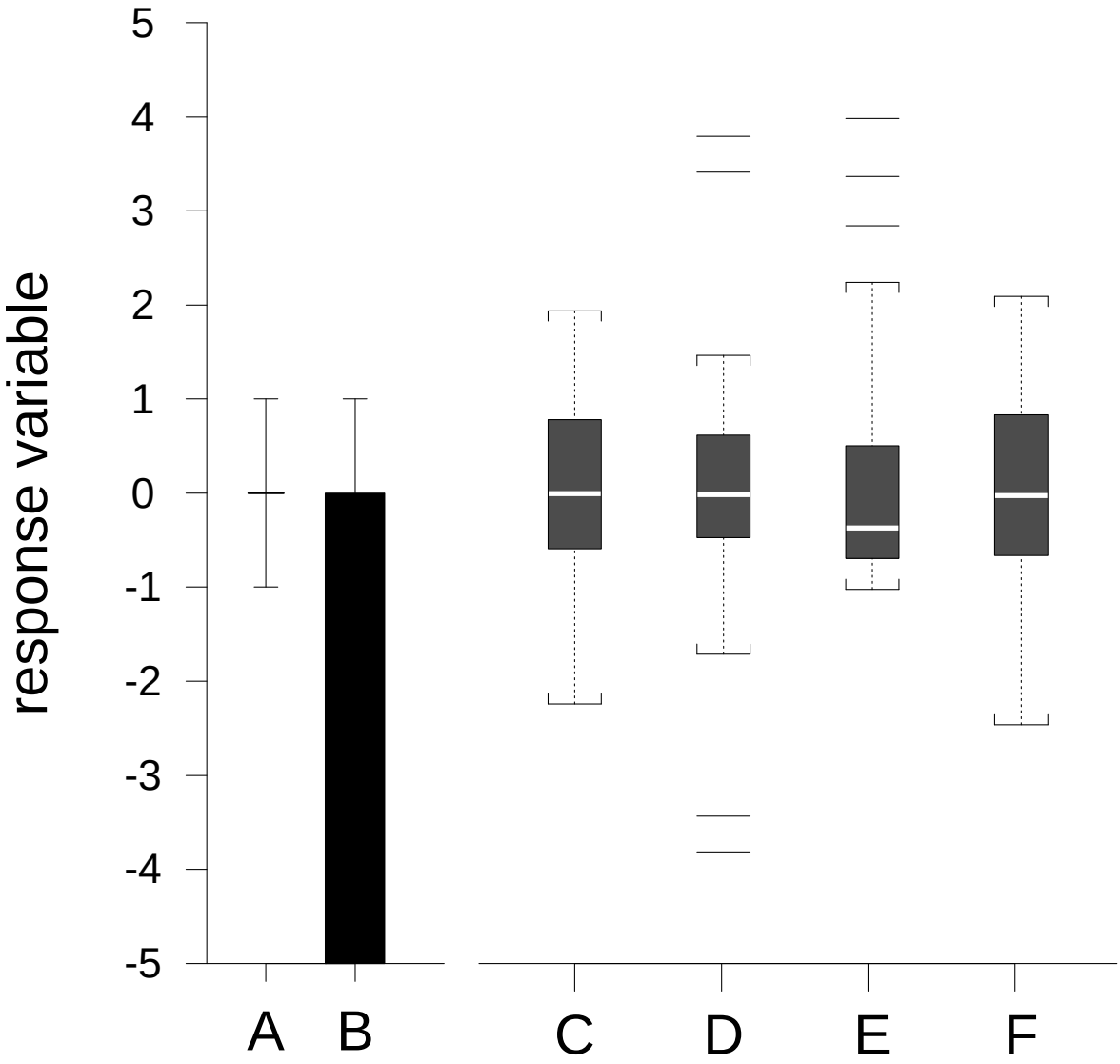
Fig. 5 Residual analysis for the untransformed data. From top left: normal–plot, Tukey–Anscombe–plot, leverage–plot, residuals against the variables (sex, age, number of animals in group).

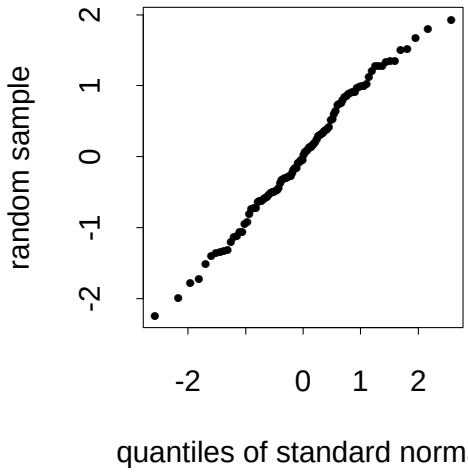
Fig. 6 Residual analysis for the data when Tukey–first–aid transformation is applied (types of plots as in Fig. 5).

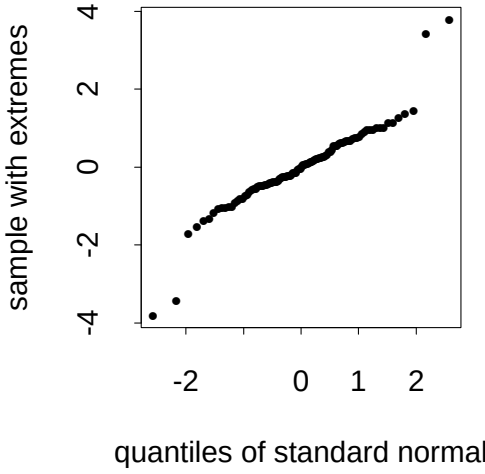
Fig. 7 Residual analysis for the data after inlcuding of age to the power of two (types of plots as in Fig. 5).

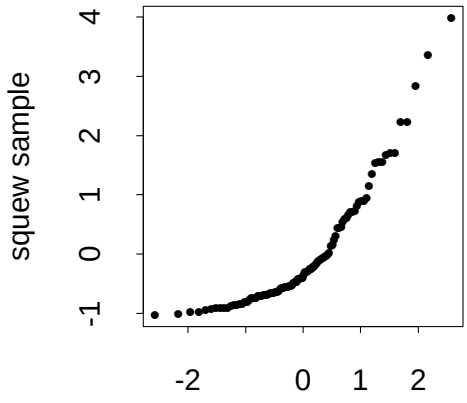
Fig. 8 Residual analysis for the data after correction of the extreme values: the final model (types of plots as in Fig. 5).

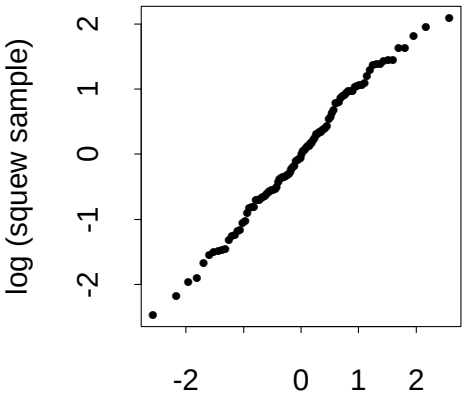
Fig. 9 Residuals (slope given by their sign, length reflects the size) if the significant interaction is not included in the model.

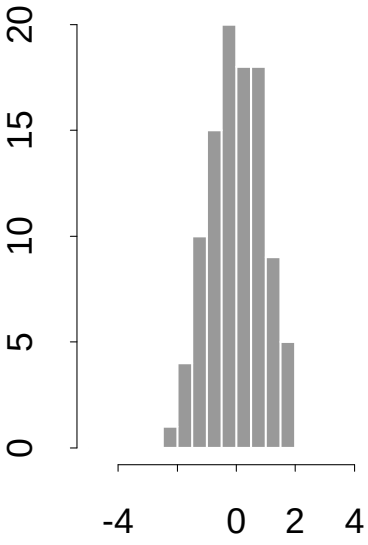

response variable
5
4
3
2
1
0
-1
-2
-3
-4
-5
A
B
C
D
E
F


random sample
-2
-1
0
1
2
-2
0
1
2
quantiles of standard normal

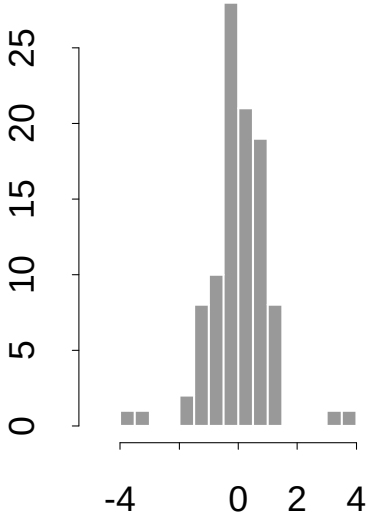

sample with extremes
-4
-2
0
2
4
-2
0
1
2
quantiles of standard normal


squew sample
-1
0
1
2
3
4
-2
0
1
2
quantiles of standard normal

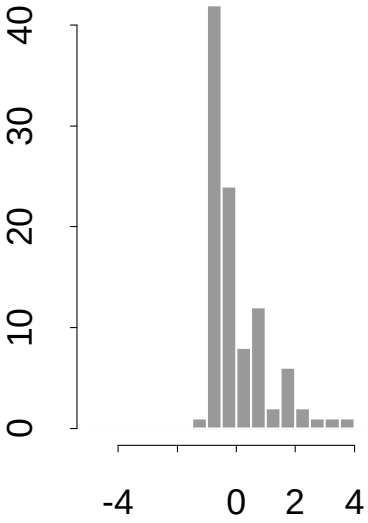

log (squew sample)
-2
-1
0
1
2
-2
0
1
2
quantiles of standard normal



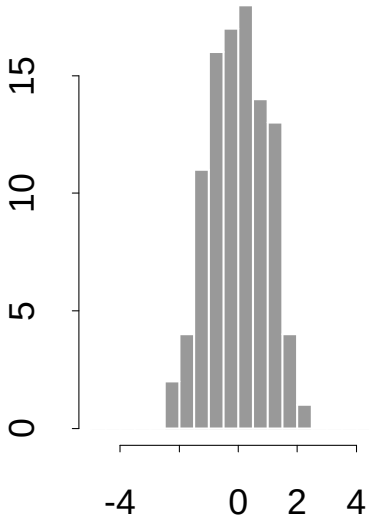
random sample



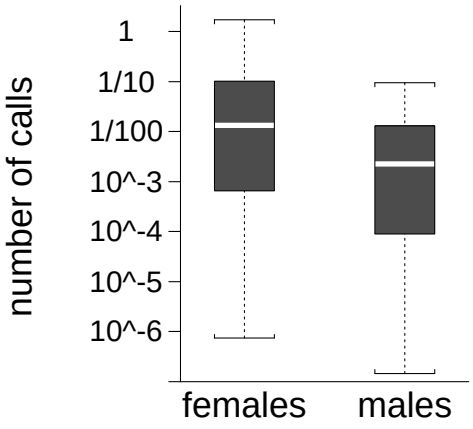
sample with extremes

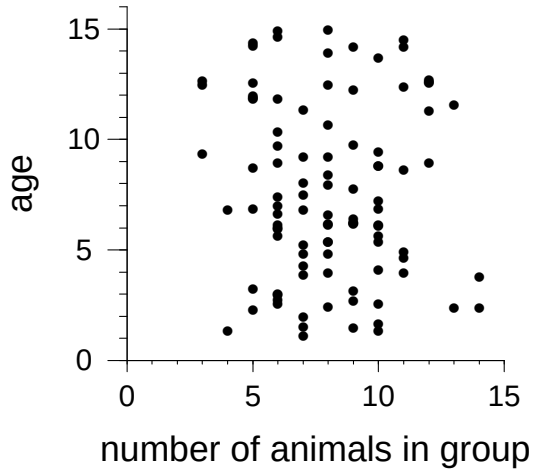


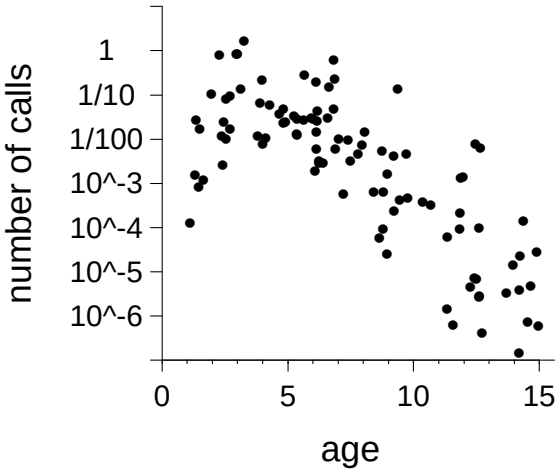
squew sample

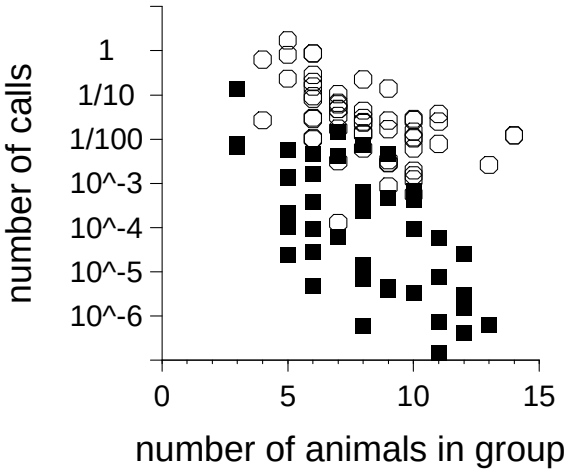


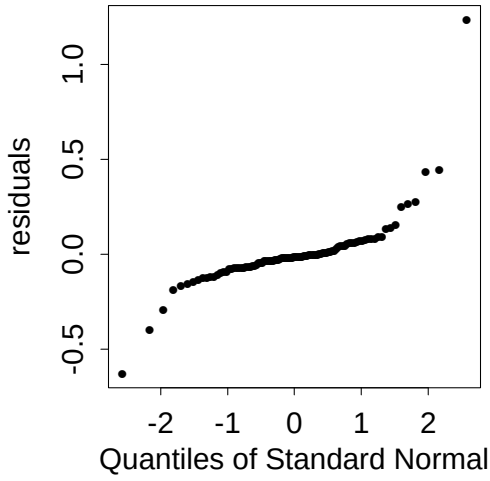
log (squew sample)

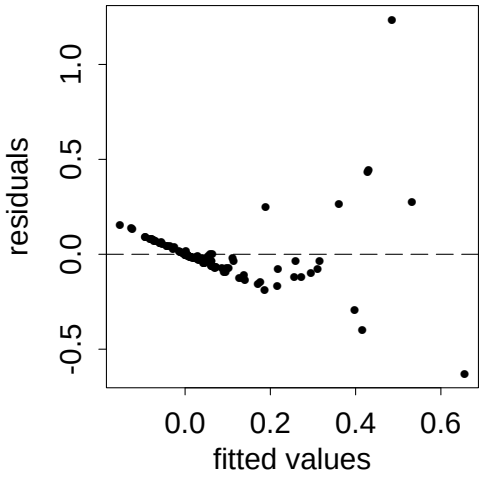

number of calls
1
1/10
1/100
10^-3
10^-4
10^-5
10^-6
females
males

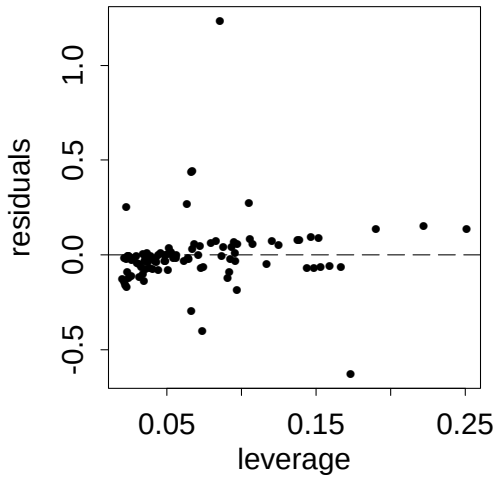

age
15
10
5
0
0
5
10
15
number of animals in group

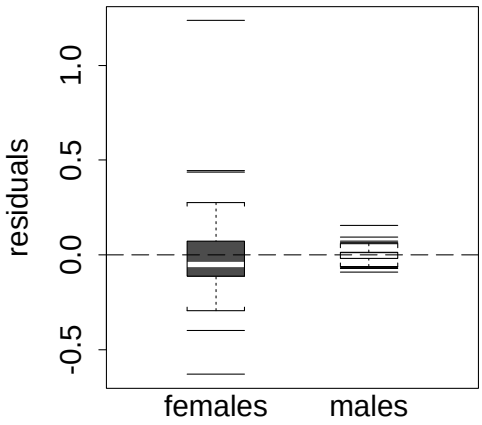

number of calls
1
1/10
1/100
10^-3
10^-4
10^-5
10^-6
0
5
10
15
age

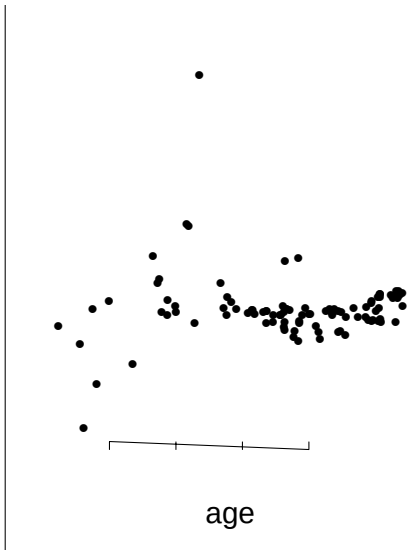

number of calls
1
1/10
1/100
10^-3
10^-4
10^-5
10^-6
0
5
10
15
number of animals in group

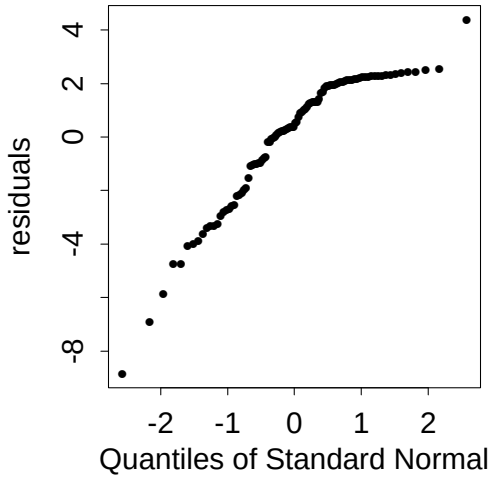

residuals
1.0
0.5
0.0
-0.5
-2
-1
0
1
2
Quantiles of Standard Normal

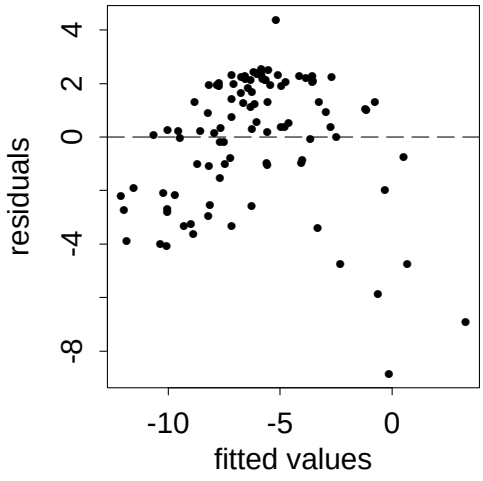

residuals
1.0
0.5
0.0
-0.5
0.0
0.2
0.4
0.6
fitted values

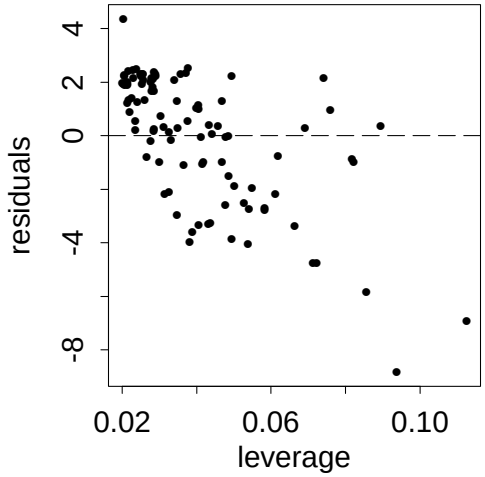

residuals
1.0
0.5
0.0
-0.5
0.05
0.15
0.25
leverage

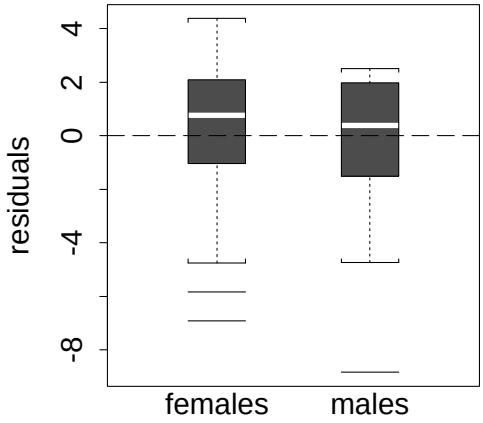

residuals
1.0
0.5
0.0
-0.5
females
males

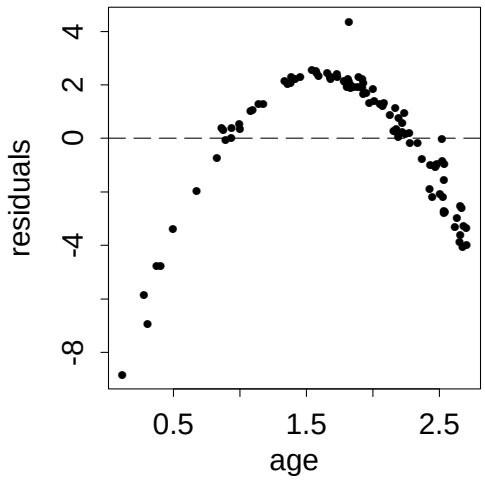

age

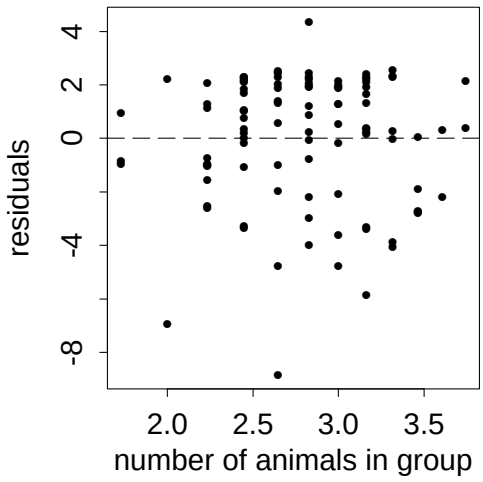

residuals
4
2
0
-4
-8
-2
-1
0
1
2
Quantiles of Standard Normal

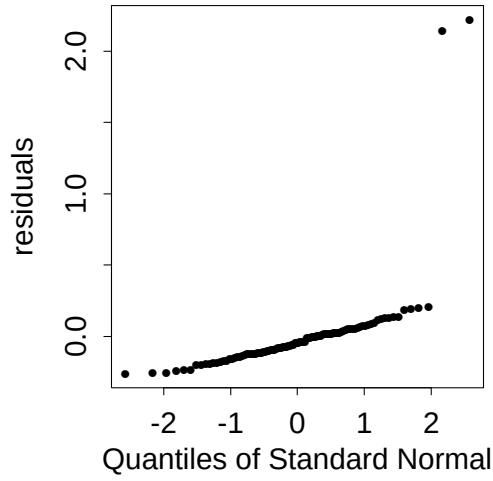

residuals
4
2
0
-4
-8
-10
-5
0
fitted values

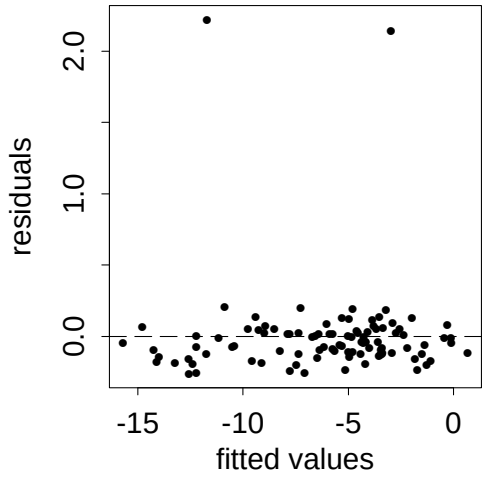

residuals
4
2
0
-4
-8
0.02
0.06
0.10
leverage

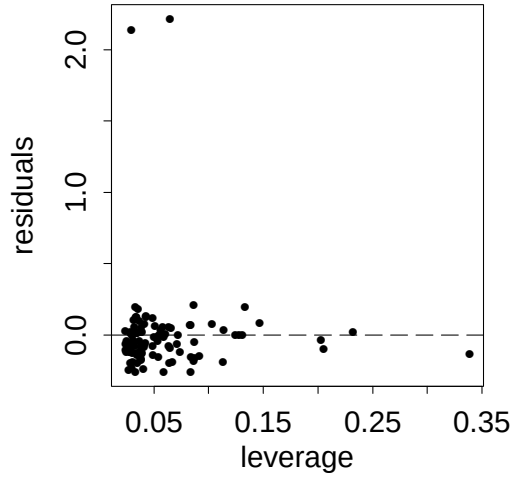

residuals
4
2
0
-4
-8
females
males

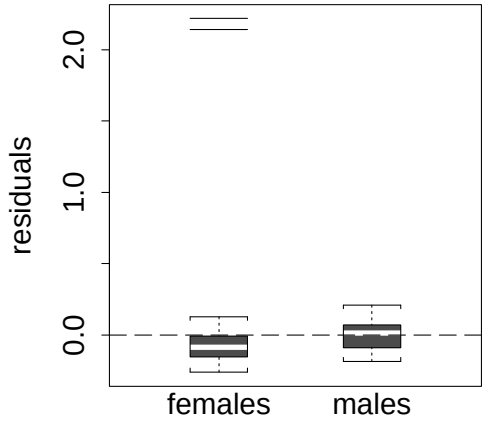

residuals
4
2
0
-4
-8
0.5
1.5
2.5
age

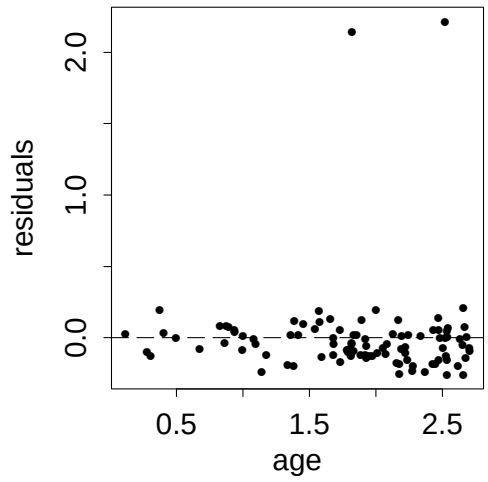

residuals
4
2
0
-4
-8
2.0
2.5
3.0
3.5
number of animals in group

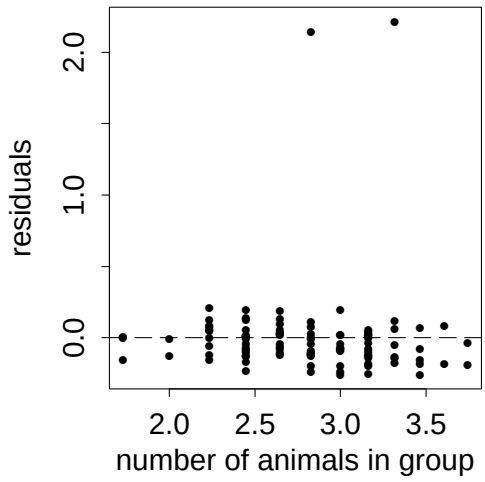

residuals
0.0
1.0
2.0
-2
-1
0
1
2
Quantiles of Standard Normal

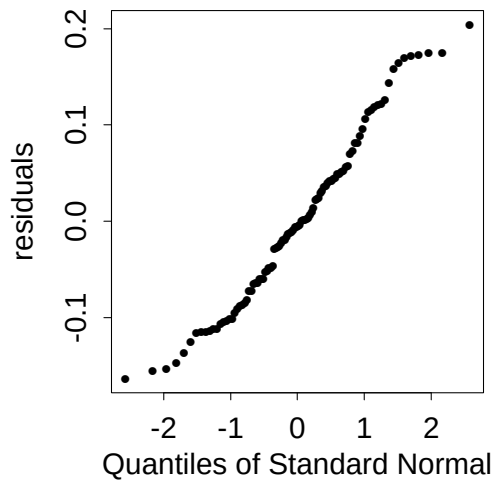

residuals
0.0
1.0
2.0
-15
-10
-5
0
fitted values

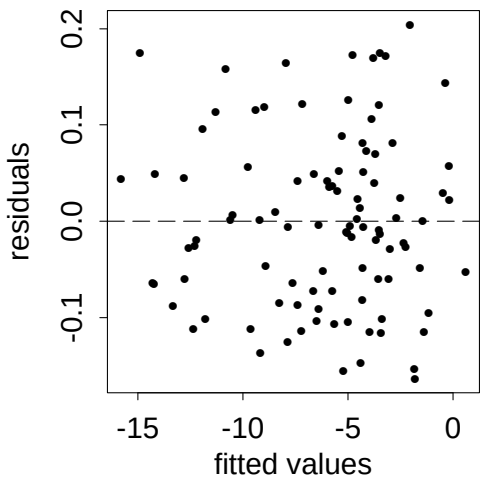

residuals
0.0
1.0
2.0
0.05
0.15
0.25
0.35
leverage

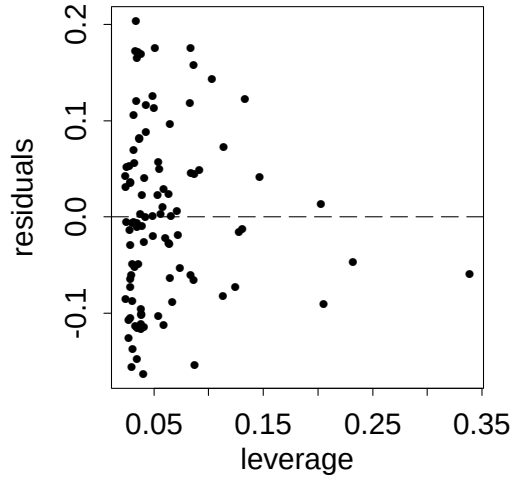

residuals
0.0
1.0
2.0
females
males

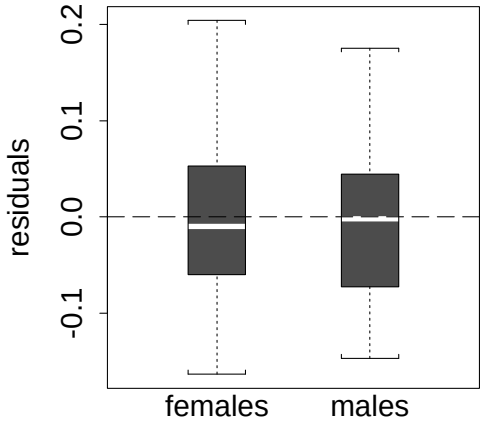

residuals
0.0
1.0
2.0
0.5
1.5
2.5
age

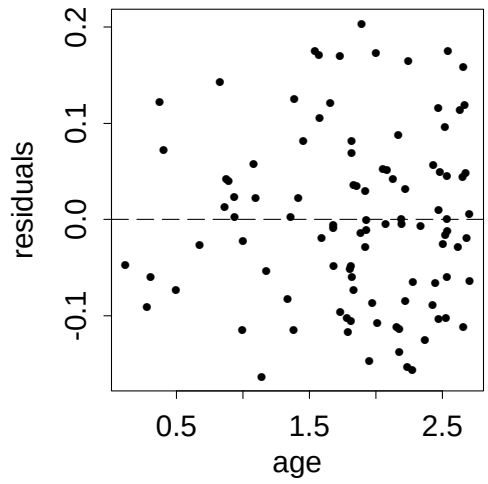

residuals
0.0
1.0
2.0
2.0
2.5
3.0
3.5
number of animals in group

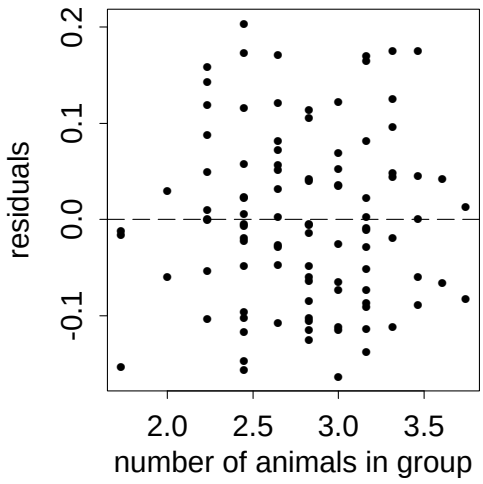

residuals
-0.1
0.0
0.1
0.2
-2
-1
0
1
2
Quantiles of Standard Normal

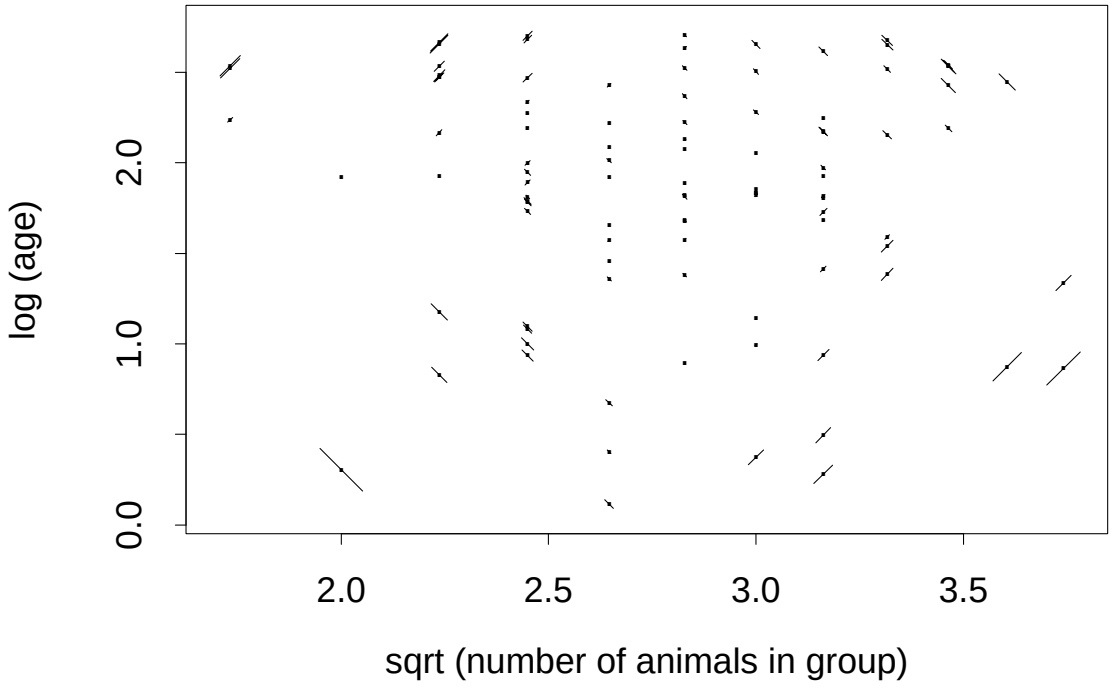

residuals
-0.1
0.0
0.1
0.2
-15
-10
-5
0
fitted values


residuals
-0.1
0.0
0.1
0.2
0.05
0.15
0.25
0.35
leverage


residuals
-0.1
0.0
0.1
0.2
females
males


residuals
-0.1
0.0
0.1
0.2
0.5
1.5
2.5
age


residuals
-0.1
0.0
0.1
0.2
2.0
2.5
3.0
3.5
number of animals in group


log (age)
0.0
1.0
2.0
2.0
2.5
3.0
3.5
sqrt (number of animals in group)