Analysis of Covariance

Bret Larget

Departments of Botany and of Statistics
University of Wisconsin—Madison

February 27, 2007

The Big Picture

- *Analysis of covariance* is the term given to the special case of a linear model where there are a mix of categorical and quantitative explanatory variables.
- With a single categorical and a single quantitative explanatory variable, the analysis consists of fitting separate lines to each group.
- A model fit without an *interaction term* assumes that the slopes for all groups are identical, but that the intercepts are potentially different.
- A model with an interaction term allows for both different slopes and intercepts for each group.
- The fitted model when interactions are included is identical to fitting separate regression lines.
- However, inference can differ when fitting a single model as compared to fitting separate regression models for each group because error estimates are shared across models.

Birds and Bats

- Birds and bats must expend considerable energy to fly.
- Some bats use echolocation in flight which also requires energy.
- Other bats eat fruit and do not have the ability to echolocate.
- Scientists studied energy use of several species of birds and bats to examine the relationship between mass and energy expenditure during flight to see if echolocating bats had a higher cost.
- Variables are `mass` (grams), `type` (factor with levels `bird`, `eBat`, and `nBat`, latter two for echolocating and non-echolocating), and the response `energy` (Watts).

Data

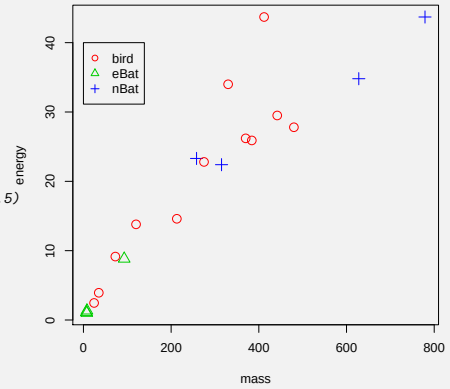
```
> bats = read.table("bats.txt", header = T)
> bats
```

	species	mass	type	energy
1	PteropusGouldi	779.0	nBat	43.70
2	PteropusPoliocephalus	628.0	nBat	34.80
3	HypsignathusMonstrosus	258.0	nBat	23.30
4	EidolonHelvum	315.0	nBat	22.40
5	MeliphagaVirescens	24.3	bird	2.46
6	MelipsittacusUndulatus	35.0	bird	3.93
7	SturmisVulgaris	72.8	bird	9.15
8	FalcoSpaverius	120.0	bird	13.80
9	FalcoTinnunculus	213.0	bird	14.60
10	CorvusOssifragus	275.0	bird	22.80
11	LarusAtricilla	370.0	bird	26.20
12	ColumbaLivia	384.0	bird	25.90
13	ColumbaLivia	442.0	bird	29.50
14	ColumbaLivia	412.0	bird	43.70
15	ColumbaLivia	330.0	bird	34.00
16	CorvusCrytoleucos	480.0	bird	27.80
17	PhyllostomasHastatus	93.0	eBat	8.83
18	PlecotusAuritus	8.0	eBat	1.35
19	PipistrellusPipistrellus	6.7	eBat	1.12
20	PlecotusAuritus	7.7	eBat	1.02

- Notice that both `mass` and `energy` span different orders of magnitude.
- The two bat types are quite different in mass.
- Birds fill the gap.
- Each observation corresponds to a single study.
- Some studies are on the same species.

Scatterplot

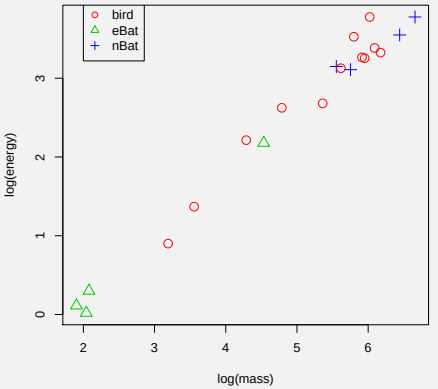
```
> attach(bats)
> pch.b = unclass(type)
> col.b = unclass(type) + 1
> plot(energy ~ mass, pch = pch.b, col = col.b, cex = 1.5)
> legend(0, 40, levels(type), pch = 1:3, col = 2:4)
```



Transformed Data

```
> bats0.form = formula(log(energy) ~ log(mass))
> plot(bats0.form, pch = pch.b, col = col.b, cex = 1.5)
> legend(2, 4, levels(type), pch = 1:3, col = 2:4)
```

- Log transformation of both variables leads to data that better fitslinear model assumptions.



Null Model (without type)

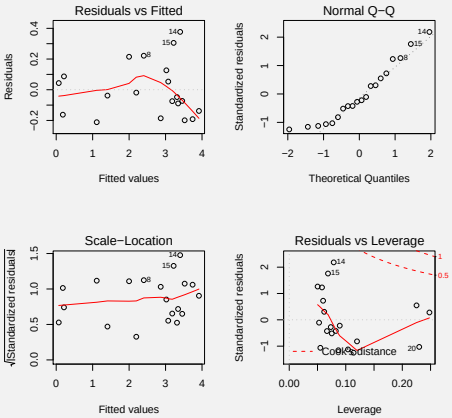
```
> bats0.lm = lm(log(energy) ~ log(mass))
> summary(bats0.lm)$coefficients

             Estimate Std. Error   t value     Pr(>|t|)
(Intercept) -1.4682584  0.1371618 -10.70457 3.101143e-09
log(mass)    0.8086098  0.0268400  30.12704 7.440291e-17

> anova(bats0.lm)

Analysis of Variance Table

Response: log(energy)
          Df  Sum Sq Mean Sq F value    Pr(>F)
log(mass)  1 29.3919 29.3919  907.64 < 2.2e-16 ***
Residuals 18  0.5829  0.0324
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Fitted model has a single slope and intercept.

Null Model Plots

```
> par(mfrow = c(2, 2))
> plot(bats0.lm)
```

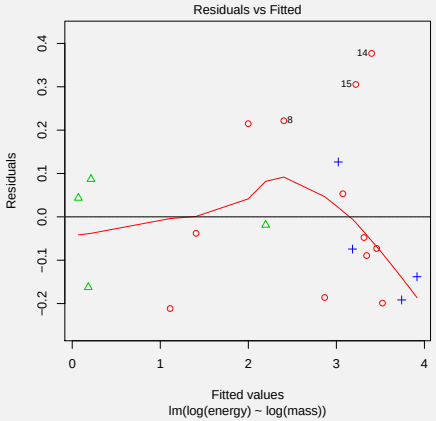
- The residual plot indicates potential minor heterskedasticity and non-linearity, primarily due to the two bird studies with highest energy.
- No highly influential points.
- Adequate fit.



Null Model Plots

```
> plot(bats0.lm, which = 1, pch = pch.b, col = col.b)
> abline(h = 0)
```

- Here is how to add plotting characters and color to a residual plot.



Model with type

```
> bats1.lm = lm(log(energy) ~ log(mass) + type)
> summary(bats1.lm)$coefficients

               Estimate Std. Error    t value     Pr(>|t|)
(Intercept) -1.47409828 0.23901543 -6.1673771 1.352481e-05
log(mass)    0.81495749 0.04454143 18.2966182 3.757576e-12
typeeBat    -0.02359824 0.15760050 -0.1497345 8.828453e-01
typenBat    -0.10226192 0.11418264 -0.8955995 3.837430e-01

> anova(bats1.lm)

Analysis of Variance Table

Response: log(energy)
          Df  Sum Sq Mean Sq  F value    Pr(>F)
log(mass)  1 29.3919 29.3919 849.9108 2.691e-15 ***
type       2  0.0296  0.0148   0.4276    0.6593
Residuals 16  0.5533  0.0346
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> cf1 = coef(bats1.lm)
> int1.bird = cf1[1]
> int1.eBat = cf1[1] + cf1[3]
> int1.nBat = cf1[1] + cf1[4]
```

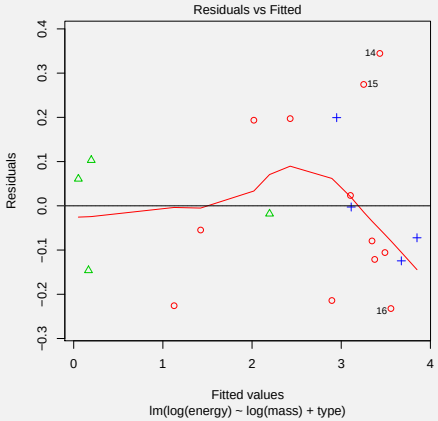
- Fitted model has a single slope, but different intercept for each type.
- Adding multiple intercepts does not improve fit significantly.

Type	Intercept	Slope
bird	-1.474	0.815
eBat	-1.498	0.815
nBat	-1.576	0.815

Residual Plot

```
> plot(bats1.lm, which = 1, pch = pch.b, col = col.b)
> abline(h = 0)
```

- Birds are slightly more variable than bats.



Model with Interaction

```
> bats2.lm = lm(log(energy) ~ log(mass) * type)
> round(summary(bats2.lm)$coefficients, 4)

                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        -1.5808     0.2944 -5.3692   0.0001
log(mass)           0.8354     0.0553 15.1004   0.0000
typeeBat            0.1103     0.3847  0.2867   0.7785
typenBat            1.3784     1.2952  1.0642   0.3053
log(mass):typeeBat -0.0307     0.1028 -0.2987   0.7696
log(mass):typenBat -0.2456     0.2134 -1.1507   0.2691

> anova(bats2.lm)

Analysis of Variance Table

Response: log(energy)
              Df  Sum Sq Mean Sq  F value    Pr(>F)
log(mass)      1 29.3919 29.3919 815.0382 8.265e-14 ***
type           2  0.0296  0.0148   0.4100    0.6713
log(mass):type 2  0.0484  0.0242   0.6718    0.5265
Residuals     14  0.5049  0.0361
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> cf2 = coef(bats2.lm)
> int2.bird = cf2[1]
> slp2.bird = cf2[2]
> int2.eBat = cf2[1] + cf2[3]
> slp2.eBat = cf2[2] + cf2[5]
> int2.nBat = cf2[1] + cf2[4]
> slp2.nBat = cf2[2] + cf2[6]
```

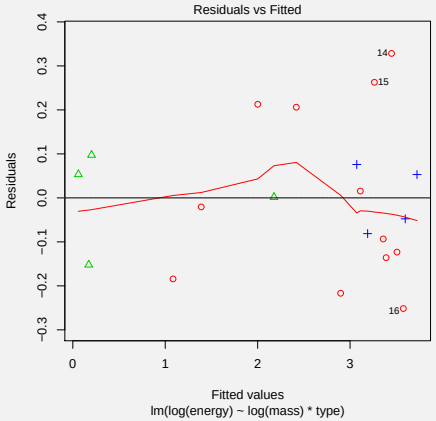
- Fitted model has different slopes and intercepts for each type.
- Differences are not statistically significant.

Type	Intercept	Slope
bird	-1.581	0.835
eBat	-1.471	0.805
nBat	-0.202	0.59

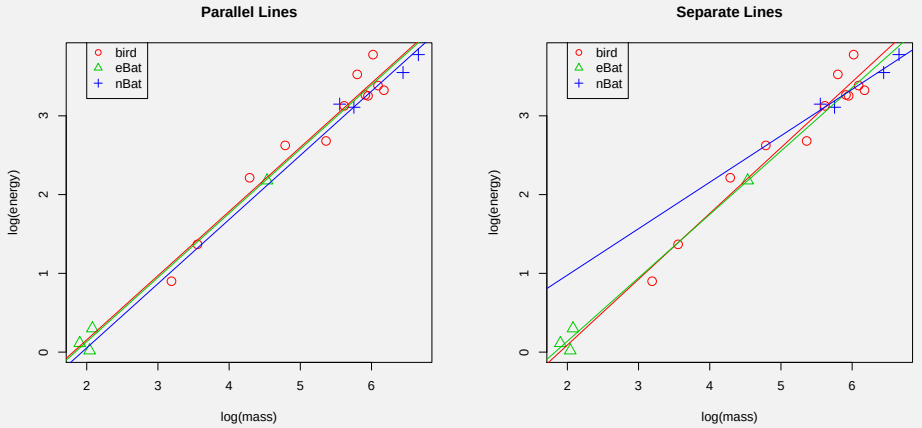
Residual Plot

```
> plot(bats2.lm, which = 1, pch = pch.b, col = col.b)
> abline(h = 0)
```

- Birds are slightly more variable than bats.



Plots of Fits



Reduced and Full Models

- Every hypothesis test we have seen up to this point is a comparison between *two nested models*.
- The *null model* is the *reduced model* which is a special case of the *alternative model* or the *full model*.
- The full model has more free parameters.
- In the reduced model, one or more parameters from the full model are constrained (often set to 0).
- Other possible constraints are equality constraints (such as $\beta_2 = \beta_3$).
- Stating which parameters are constrained in the reduced model is another way to specify the null hypothesis.

The F-Test Statistic

- Let the residual sums of squares from the reduced and full models be $\mathrm{RSS}_{(reduced)}$ and $\mathrm{RSS}_{(full)}$, respectively.
- Denote the number of unconstrained parameters for the mean in the models be $k_{(reduced)}$ and $k_{(full)}$, respectively.
- Suppose that there are n total observations.

It follows that:

1. $\mathbb{E}(\mathrm{RSS}_{(reduced)} - \mathrm{RSS}_{(full)}) = (k_{(full)} - k_{(reduced)})(\sigma^2 + C)$ so that $\mathbb{E}(\mathrm{RSS}_{(reduced)} - \mathrm{RSS}_{(full)})/(k_{(full)} - k_{(reduced)}) = \sigma^2 + C$ where $C = 0$ if the reduced model is true and $C > 0$ when the full model is true.
2. $\mathbb{E}(\mathrm{RSS}_{(full)}) = (n - k_{(full)})\sigma^2$ so that $\mathbb{E}(\mathrm{RSS}_{(full)})/(n - k_{(full)}) = \sigma^2$ when the reduced or full models are true.
3. $$F = \frac{\left(\mathrm{RSS}_{(reduced)} - \mathrm{RSS}_{(full)}\right)/(k_{(full)} - k_{(reduced)})}{\mathrm{RSS}_{(full)}/(n - k_{(full)})} \sim F_{k_{(full)}-k_{(reduced)}, n-k_{(full)}}$$
   when the reduced model is true and the usual assumptions hold.