Multiple Linear Regression Case Study

Bret Larget

Departments of Botany and of Statistics
University of Wisconsin—Madison

February 5, 2008

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

<table>
<thead>
<tr>
<th></th>
<th>species</th>
<th>mass</th>
<th>type</th>
<th>energy</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>PteropusGouldi</td>
<td>779.0</td>
<td>nBat</td>
<td>43.7</td>
</tr>
<tr>
<td>2</td>
<td>PteropusPoliocephalus</td>
<td>628.0</td>
<td>nBat</td>
<td>34.8</td>
</tr>
<tr>
<td>3</td>
<td>HypsignathusMonotremus</td>
<td>258.0</td>
<td>nBat</td>
<td>23.3</td>
</tr>
<tr>
<td>4</td>
<td>EidolonHelvum</td>
<td>315.0</td>
<td>nBat</td>
<td>22.4</td>
</tr>
<tr>
<td>5</td>
<td>NemaphasVirens</td>
<td>24.3</td>
<td>bird</td>
<td>2.4</td>
</tr>
<tr>
<td>6</td>
<td>MelopsittacusUndulatus</td>
<td>35.0</td>
<td>bird</td>
<td>3.8</td>
</tr>
<tr>
<td>7</td>
<td>SturnusVulgaris</td>
<td>72.8</td>
<td>bird</td>
<td>9.15</td>
</tr>
<tr>
<td>8</td>
<td>FalcoSpaverius</td>
<td>120.0</td>
<td>bird</td>
<td>13.8</td>
</tr>
<tr>
<td>9</td>
<td>FalcoTinnunculus</td>
<td>213.0</td>
<td>bird</td>
<td>14.6</td>
</tr>
<tr>
<td>10</td>
<td>CorvusOssifragus</td>
<td>275.0</td>
<td>bird</td>
<td>22.8</td>
</tr>
<tr>
<td>11</td>
<td>LarusAtricilla</td>
<td>370.0</td>
<td>bird</td>
<td>26.20</td>
</tr>
<tr>
<td>12</td>
<td>ColumbLivia</td>
<td>384.0</td>
<td>bird</td>
<td>26.90</td>
</tr>
<tr>
<td>13</td>
<td>ColumbLivia</td>
<td>442.0</td>
<td>bird</td>
<td>29.50</td>
</tr>
<tr>
<td>14</td>
<td>ColumbLivia</td>
<td>412.0</td>
<td>bird</td>
<td>43.70</td>
</tr>
<tr>
<td>15</td>
<td>ColumbLivia</td>
<td>330.0</td>
<td>bird</td>
<td>34.00</td>
</tr>
<tr>
<td>16</td>
<td>CorvusCryptoleucus</td>
<td>480.0</td>
<td>bird</td>
<td>27.80</td>
</tr>
<tr>
<td>17</td>
<td>PsephotostomaHastatus</td>
<td>93.0</td>
<td>eBat</td>
<td>8.83</td>
</tr>
<tr>
<td>18</td>
<td>PlecotusAuritus</td>
<td>8.0</td>
<td>eBat</td>
<td>1.35</td>
</tr>
<tr>
<td>19</td>
<td>PipistrellusPipistrellus</td>
<td>6.7</td>
<td>eBat</td>
<td>1.12</td>
</tr>
<tr>
<td>20</td>
<td>PlecotusAuritus</td>
<td>7.7</td>
<td>eBat</td>
<td>1.02</td>
</tr>
</tbody>
</table>

Box-and-Whisker Plots

• 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.
The scatterplot reveals potential problems with fitting a standard regression model:
- Two bird observations appear to be potential outliers;
- There is some apparent curvature;
- Points with high mass have more variable energy measurements than points with low mass;

We will, however, fit a few models to illustrate the method and to show how these potential problems can be identified more readily with residual plots.

`fit0` is a simple linear regression of energy on mass

`fit1` adds type as an input variable. This has the effect of allowing the intercept to be different for each type.

`fit2` has mass and type and an interaction between them. This has allows each type to have its own slope and intercept.
Estimated Coefficients

> coef(fit0)

(Intercept) mass
4.09991727 0.05869642

- fit0 shows the intercept and parameter for mass which is the slope.

Estimated Coefficients

> coef(fit1)

(Intercept) mass typeeBat typenBat
6.02197707 0.05749542 -4.60071984 -3.43220829

- fit1 shows an intercept for all predictions, a parameter for mass which is the common slope, and then adjustments to be made if the type is eBat or nBat.
- In effect, these are estimated differences of the intercept relative to bird.
- For birds, the intercept is 6.02.
- For echolocating bats the intercept is $6.02 + (-4.6) = 1.42$.
- For non-echolocating bats the intercept is $6.02 + (-3.43) = 2.59$.
- The three lines are parallel and share the common slope 0.0575.

Estimated Coefficients

> coef(fit2)

(Intercept) mass typeeBat typenBat
3.31674159 0.06777464 -2.82275855 7.91064213

- fit2 shows six estimated coefficients, the intercept and slope for bird and then adjustments to each of these for the other types.
- For birds, the intercept is 3.32 and the slope is 0.0678.
- For echolocating bats the intercept is $3.32 + (-2.82) = 0.494$ and the slope is $0.0678 + (0.0219) = 0.0896$.
- For non-echolocating bats the intercept is $3.32 + (7.91) = 11.2$ and the slope is $0.0678 + (-0.0277) = 0.04$.

Interpretation of Coefficients

> coef(fit2)

(Intercept) mass typeeBat typenBat
3.31674159 0.06777464 -2.82275855 7.91064213

- The intercept is the predicted energy of a bird at mass 0 — no biological relevance.
- The third coefficient is the estimated difference between the predicted energies for echolocating bats and birds at mass 0.
- Notice that the predicted difference is not the same at all masses.
- This parameter has no biological significance also.
- Similar comments can be made about the non-echolocating bats — in particular, even though the intercept for non-echolocating bats is higher than for birds, at the range of mass where there are both birds and non-echolocating bats, the bird line is higher.
Residual Plot

> plot(xyplot(residuals(fit2) ~ fitted(fit2), pch = 16))

- Residual plot from last fit.
- Notice the fan-shaped pattern.
- Residuals are larger for large mass.
- A transformation may help.

Log Transformed Data

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

More Analysis

- Do remaining analysis live in R.