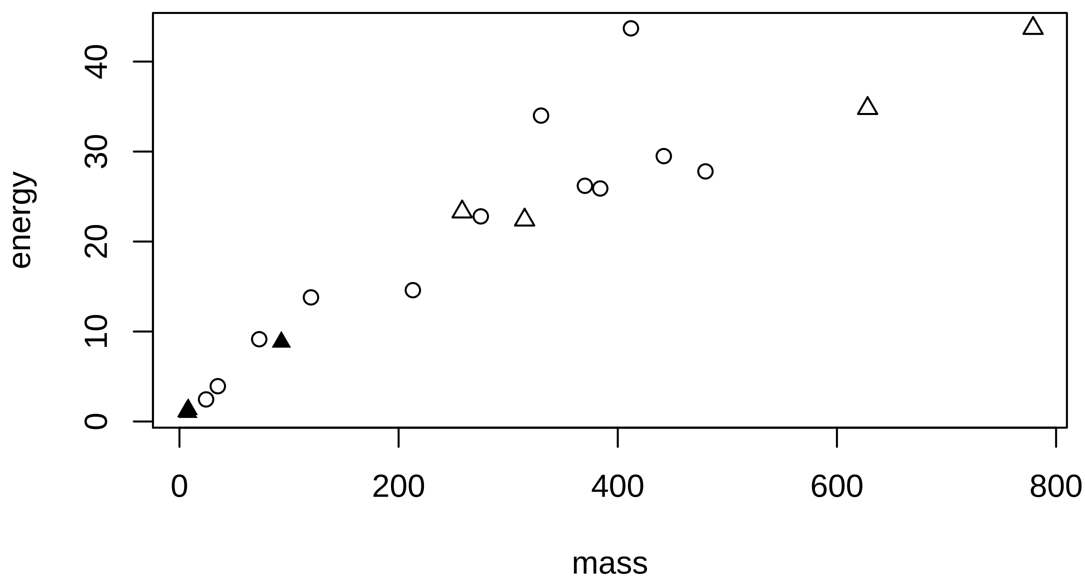
Here we look at the second case study in Chapter 10, involving bats, birds, and energy expenditure. The variables are *mass* (measured in grams), *type*, which is one of bird, echolocating bat, and non-echolocating bat, and *energy* expenditure in flight (measured in Watts). The data is on page 270 of the textbook (and on the previous hand out).

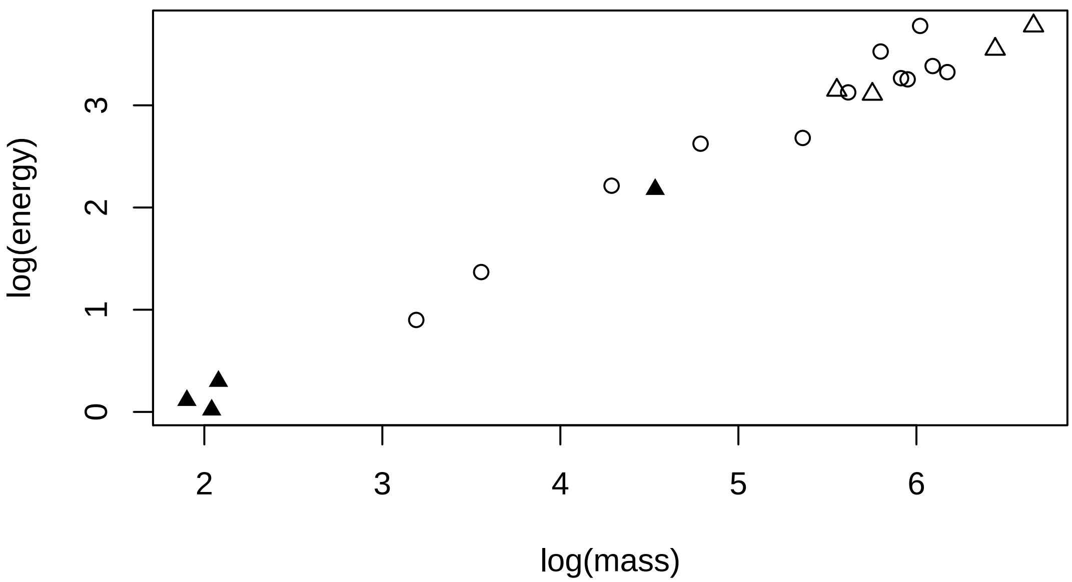
The first plot to examine is a scatterplot of *energy* versus *mass*, using different symbols for the different types. This R code does this.

```
> nBats <- type == "nBat"
> eBats <- type == "eBat"
> birds <- type == "bird"
> plot(mass, energy, type = "n")
> points(mass[nBats], energy[nBats], pch = 2)
> points(mass[eBats], energy[eBats], pch = 17)
> points(mass[birds], energy[birds], pch = 1)
```



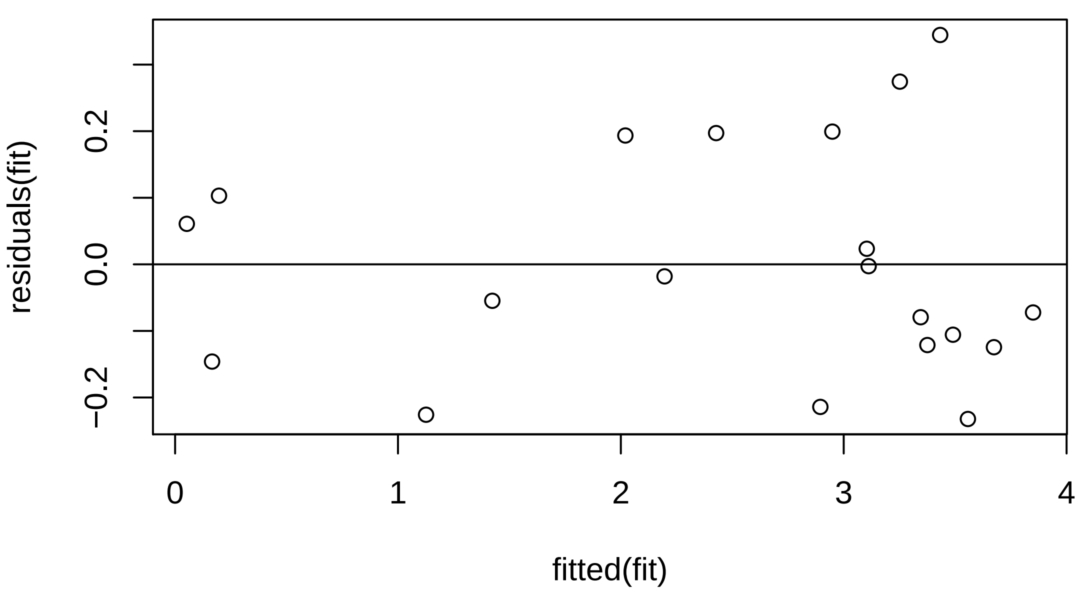
Now use the log scale.

```
> plot(log(mass), log(energy), type = "n")
> points(log(mass[nBats]), log(energy[nBats]), pch = 2)
> points(log(mass[eBats]), log(energy[eBats]), pch = 17)
> points(log(mass[birds]), log(energy[birds]), pch = 1)
```



Next, fit a linear model for the log transformed variables and plot the residuals.

```
> fit <- lm(log(energy) ~ log(mass) + type)
> plot(fitted(fit), residuals(fit))
> abline(h = 0)
```



Compare the summary from R to that found in the textbook on page 273.

```
> summary(fit)

Call:
lm(formula = log(energy) ~ log(mass) + type)

Residuals:
     Min       1Q   Median       3Q      Max
-0.23224 -0.12199 -0.03637  0.12574  0.34457

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.47410    0.23902  -6.167 1.35e-05 ***
log(mass)    0.81496    0.04454  18.297 3.76e-12 ***
typeeBat    -0.02360    0.15760  -0.150    0.883
typenBat    -0.10226    0.11418  -0.896    0.384
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.186 on 16 degrees of freedom
Multiple R-Squared: 0.9815,       Adjusted R-squared: 0.9781
F-statistic: 283.6 on 3 and 16 DF,  p-value: 4.464e-14
```

Summary of fit from the textbook.

Variable	Coefficient	SE	t	p-value
Constant	−1.5764	0.2872	−5.4880	< 0.0001
log(*mass*)	0.8150	0.0445	18.2966	< 0.0001
bird	0.1023	0.1142	0.8956	0.3837
ebat	0.0787	0.2027	0.3881	0.7030

Consider, for example, the equation to predict log(*energy*) from log(*mass*) for birds. From R, the intercept is −1.4741. In the textbook, it is $-1.5764 + 0.1023 = -1.4741$. The numerical summaries differ because R and the textbook made different choice as to which two (of three) levels of the factor *type* should be represented with indicator variables, a variable that is 1 when for a given level and 0 for all other levels. Notice that the fitted values are identical, but that the parameterization is different.

Hypothesis tests

Suppose that we wanted to test the hypothesis that the intercept for the echolocating bats and the non-echolocating bats was the same. From the textbook output, the implied model is

$$\log(energy) = \beta_0 + \beta_1(\log(mass)) + \beta_2(bird) + \beta_3(eBat)$$

The intercept for non-echolocating bats is β_0 and for echolocating bats is $\beta_0 + \beta_3$. Thus the null hypothesis is $H_0:\ \beta_3 = 0$ and the alternative is $H_a:\ \beta_3 \neq 0$. The test statistic is computed for us, $t = 0.3881$ with a p-value of 0.7030. There is little evidence that the intercepts differ for the two types of bats.

But can we do the same thing in R? Here, because the parameters were selected differently, there is a different implied model.

$$\log(energy) = \beta_0 + \beta_1(\log(mass)) + \beta_2(eBat) + \beta_3(nBat)$$

Now the intercepts are $\beta_0 + \beta_2$ and $\beta_0 + \beta_3$. If these are equal, then $\beta_2 = \beta_3$. But, unfortunately, we cannot read this directly off the output. One solution is to force R to pick different levels for the indicator variables so that this test would be based on a single coefficient. But there is another way.

The test statistic for the null hypothesis $H_0: \ \beta_2 - \beta_3 = 0$ is $t = (\hat\beta_2 - \hat\beta_3)/\mathrm{SE}(\hat\beta_2 - \hat\beta_3)$. The formula for $\mathrm{SE}(\hat\beta_2 - \hat\beta_3)$ comes from the general formula for the SE of $g = C_0\hat\beta_0 + C_1\hat\beta_1 + \cdots + C_p\hat\beta_p$, which is

$$\mathrm{SE}(g) = \sqrt{\sum_{i=0}^{p} C_i^2 \mathrm{Var}(\hat\beta_i) + 2C_iC_j \sum_{i=0}^{p-1} \sum_{j=i+1}^{p} \mathrm{Cov}(\hat\beta_i, \hat\beta_j)}$$

For out case, all of the C_i are 0 except $C_2 = 1$ and $C_3 = -1$, so

$$\mathrm{SE}(\hat\beta_2 - \hat\beta_3) = \sqrt{1^2\mathrm{Var}(\hat\beta_2) + (-1)^2\mathrm{Var}(\hat\beta_3) + 2(1)(-1)\mathrm{Cov}(\hat\beta_2, \hat\beta_3)}$$

In R, we can find the variance-covariance matrix using the function `vcov`.

```
> vc <- vcov(fit)
> vc

             (Intercept)    log(mass)     typeeBat     typenBat
(Intercept)  0.057128374 -0.010374091 -0.029752785  0.006169357
log(mass)   -0.010374091  0.001983939  0.005138789 -0.001730953
typeeBat    -0.029752785  0.005138789  0.024837918 -0.001601645
typenBat     0.006169357 -0.001730953 -0.001601645  0.013037676

> se <- sqrt(vc[3, 3] + vc[4, 4] - 2 * vc[3, 4])
> se

[1] 0.2026793

> tstat <- ((coef(fit))[3] - (coef(fit))[4])/se
> tstat

 typeeBat
0.3881190
```

Compare this to the figure above.

Here is the fit that forces the variable choices we made.

```
> fit2 <- lm(log(energy) ~ log(mass) + birds + eBats)
> summary(fit2)

Call:
lm(formula = log(energy) ~ log(mass) + birds + eBats)

Residuals:
     Min       1Q   Median       3Q      Max
-0.23224 -0.12199 -0.03637  0.12574  0.34457

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.57636    0.28724  -5.488 4.96e-05 ***
log(mass)    0.81496    0.04454  18.297 3.76e-12 ***
birdsTRUE    0.10226    0.11418   0.896    0.384
eBatsTRUE    0.07866    0.20268   0.388    0.703
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.186 on 16 degrees of freedom
Multiple R-Squared: 0.9815,       Adjusted R-squared: 0.9781
F-statistic: 283.6 on 3 and 16 DF,  p-value: 4.464e-14
```