

Bret Larget

Departments of Botany and of Statistics
University of Wisconsin—Madison

February 12, 2008

- Linear transformations *do not change the quality of a fit*.
- Linear transformations *can make parameters easier to interpret*.
- birds are the birds from the birds and bats example.

```
> bats = read.table("bats.txt", header = T)
> birds = with(bats, bats[type == "bird", ])
> birds0.lm = lm(energy ~ mass, data = birds)
> birds.mean = with(birds, mean(mass))
> birds.mean

[1] 263.175

> birds1.lm = lm(energy ~ I(mass - birds.mean), data = birds)
```

1 / 11

Transformations

Linear Transformations

2 / 11

Compare Output

- The R-Squared and residual sd are the same for both fits.
- The intercept is different.

```
> display(birds0.lm, digits = 5)

lm(formula = energy ~ mass, data = birds)
   coef.est coef.se
(Intercept) 3.31674 3.29882
mass        0.06777 0.01074
---
n = 12, k = 2
residual sd = 5.88610, R-Squared = 0.80

> display(birds1.lm, digits = 5)

lm(formula = energy ~ I(mass - birds.mean), data = birds)
   coef.est coef.se
(Intercept) 21.15333 1.69917
I(mass - birds.mean) 0.06777 0.01074
---
n = 12, k = 2
residual sd = 5.88610, R-Squared = 0.80
```

Predictions

- Use the predict() function to make predictions.
- The predicted values are the same.

```
> newbirds = data.frame(mass = c(200, 250, 300))
> predict(birds0.lm, newbirds)

      1      2      3
16.87167 20.26040 23.64913

> predict(birds1.lm, newbirds)

      1      2      3
16.87167 20.26040 23.64913
```

Transformations

Linear Transformations

3 / 11

Transformations

Linear Transformations

4 / 11

- Transformations by centering are especially helpful for coefficient interpretation when there are interactions in the model.

```
> mass.mean = with(bats, mean(mass))
> mass.mean
[1] 262.675

> bats1.lm = lm(energy ~ I(mass - mass.mean) * type, data = bats)
> display(bats1.lm, digits = 4)

lm(formula = energy ~ I(mass - mass.mean) * type, data = bats)
      (Intercept)      coef.est coef.se
I(mass - mass.mean)  21.1194  1.4552
typeeBat            0.0678  0.0092
typenBat            2.9198  16.1746
I(mass - mass.mean):typeeBat 0.6269  3.9758
I(mass - mass.mean):typenBat 0.0219  0.0687
I(mass - mass.mean):typenBat -0.0277  0.0149
---
n = 20, k = 6
residual sd = 5.0408, R-Squared = 0.90
```

```
> coef(bats1.lm)
      (Intercept)      I(mass - mass.mean)
21.11944601      0.06777464
      typeeBat      typenBat
2.91984047      0.62693918
I(mass - mass.mean):typeeBat I(mass - mass.mean):typenBat
0.02186199      -0.02772895
```

- 1 Interpret each intercept coefficient.
- 2 Interpret each slope coefficient.
- 3 Predict the energy for a 20 g echo-locating bat.
- 4 Predict the energy for a 220 g bird.
- 5 Predict the energy for a 420 g non-echo-locating bat.

- Log transformations:
 - ▶ *do change* the goodness of fit;
 - ▶ can make a linear model fit better.
 - ▶ require variables with positive numerical values;
 - ▶ are especially useful large values are more variable than small values
- The function `log()` takes the *natural* logarithm.
- Use the `exp()` function to go back to the original units.
- Since $\exp(1 + x) \approx x$ for x near 0, interpretations of coefficients after a log transformation of the outcome variable are easier to interpret.

```
> lmass.mean = with(bats, mean(log(mass)))
> bats2.lm = lm(log(energy) ~ I(log(mass) - lmass.mean) + type,
+ data = bats)
> display(bats2.lm, digits = 4)

lm(formula = log(energy) ~ I(log(mass) - lmass.mean) + type,
    data = bats)
      (Intercept)      coef.est coef.se
I(log(mass) - lmass.mean)  2.5074  0.0558
typeeBat                 -0.0236  0.1576
typenBat                 -0.1023  0.1142
---
n = 20, k = 4
residual sd = 0.1860, R-Squared = 0.98
```

- In a comparison of an echo-locating bat to a bird of equal mass, we expect the energy use to be about 2% less.
- In a comparison of a non-echo-locating bat to a bird of equal mass, we expect the energy use to be about 10% less.

```
> newData = data.frame(mass = c(20, 20, 400, 400), type = factor(c("bird",  
+ "eBat", "bird", "nBat")))  
> exp(predict(bats2.lm, newData))  
  
      1      2      3      4  
2.630822 2.569466 30.225643 27.287500
```

General Principles:

- 1 Include all input variables expected to be important on scientific grounds.
- 2 Consider including interactions for inputs of large effect.
- 3 It is okay to use judgment in deciding whether or not to include variables.
- 4 (We will revisit more formal methods of variable selection later.)

FEV Data Revisited

If time permits, do live in R.

- 654 kids with ages ranging from 3 to 19
- Age is measured in years.
- FEV is forced expiratory volume, a measure of lung capacity, and is measured in liters.
- Height is measured in inches.
- Sex has two levels (female, male)
- Smoking has two levels (nonsmoker, smoker)