Outline

1. Multiple Linear Regression
   - The model matrix
   - Least squares
   - Problem of confounded variables
2. Multiple Linear Regression: Case Study
   - Birds and Bats
   - Interpreting coefficients

Multiple Linear Regression

- Most interesting questions in biology involve relationships between multiple variables.
- Interactions between variables can be important in understanding a process.
- Example to show
  - how we can combine 2 quantitative and 1 categorical explanatory variables
  - what is the ‘model’ or ‘design’ matrix

An Example

- Small subset of the FEV data set: $n = 6$ children

fev	age	ht	sex
1.72	7	54.5	female
1.74	8	54.0	male
2.09	9	59.5	male
3.13	10	62.0	male
2.87	11	60.5	female
2.57	12	63.0	female

- Goal: develop a linear model to predict forced expiratory volume on the basis of age, height, and sex

Model Matrix

- The inputs are represented by a matrix of predictors.
- The intercept corresponds to a vector of ones.
- Each quantitative variable is a single columns.
- Each categorical variable with m levels is represented by $m - 1$ columns.
- The ith row has all information about the ith individual in the sample.

$$y = \begin{pmatrix} 1.72 \\ 1.74 \\ 2.09 \\ 3.13 \\ 2.87 \\ 2.57 \end{pmatrix}$$

$$X = \begin{pmatrix} 1 & 7 & 54.5 & 0 \\ 1 & 8 & 54.0 & 1 \\ 1 & 9 & 59.5 & 1 \\ 1 & 10 & 62.0 & 1 \\ 1 & 11 & 60.5 & 0 \\ 1 & 12 & 63.0 & 0 \end{pmatrix}$$

Model Coefficients

- The parameters of a model are written as a vector β of size k.
- The ith row of X is denoted X_i.
- The model

$$y_i = \beta_1 X_{i1} + \cdots + \beta_k X_{ik} + e_i$$

$$\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{pmatrix}$$

is written in matrix form as

$$y_i = X_i \beta + e_i.$$

where the $e_i \sim \mathrm{N}(0, \sigma^2)$.

Model Coefficients

- In matrix multiplication, the dot product of the *i*th row of the matrix on the left times the *j*th column of the matrix on the right is the *ij* element of the product.
- The number of columns of the left matrix must match the number of rows of the right matrix.
- For example,

$$X_2\beta = \begin{pmatrix} 1 & 8 & 54.0 & 1 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{pmatrix}$$

$$= 1 \cdot \beta_1 + 8 \cdot \beta_2 + 54.0 \cdot \beta_3 + 1 \cdot \beta_4$$

Shell thickness example

Mussels sampled from northern or southern New England. Reared in the lab for 3 months in the absence, or presence of Asian crabs or green crabs. Green crabs have invaded both locations for $>$ 50 years. Asian crabs invaded the southern location 15 years ago, not the northen location yet. Shell thickness was measured. (Freeman & Byers (2006) Science 313:831-833)

`STI`: adjusted final shell thickness index (g/mm^2)

Ex. of first mussels:

STI	pop	crab
0.67	North	Asian
0.69	North	Absent
0.71	North	Green
0.68	North	Asian
0.68	South	Asian
0.69	South	Asian
0.62	South	Absent
0.64	South	Green

Model equation, without interaction

Linear model to predict STI on the basis of pop and crab: first assume simple addition of `pop` and `crab` effects.

$$\text{mean}y = \text{mean STI} = \beta_1 + \beta_2 \mathbb{1}_{\text{north}} + \beta_3 \mathbb{1}_{\text{asian}} + \beta_4 \mathbb{1}_{\text{green}}$$

We will use $X_1 = 1$ for all mussels,
$X_2 = \mathbb{1}_{\text{north}} = 1$ if the mussel is from the North,
0 if not (i.e. if from the South)
$X_3 = \mathbb{1}_{\text{asian}} = 1$ if asian crab present, 0 if not
$X_4 = \mathbb{1}_{\text{asian}} = 1$ if green crab present, 0 if not

Model equation, without interaction

The equation reduces to $\mathbb{E}(y) =$

if Southern mussel in absence of crabs
if Northern mussel in absence of crabs
if Southern mussel in presence of Asian crabs
if Northern mussel in presence of Asian crabs
if Southern mussel in presence of Green crabs
if Northern mussel in presence of Green crabs

What does β_1 represent? β_2?

Shell thickness example

$$y = \text{STI} = \begin{pmatrix} 0.67 \\ 0.69 \\ 0.71 \\ 0.68 \\ 0.68 \\ 0.69 \\ 0.62 \\ 0.64 \end{pmatrix}$$

Model matrix:

$$X = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}$$

Shell thickness example

Coefficients could be:
$\beta_1 = 0.64$, $\beta_2 = 0.030$,$\beta_3 = 0.025$, $\beta_4 = 0.020$.

Then, predicted mean STI for a northern mussel in presence of asian crab:

$$
\begin{aligned}
X_2\beta &= \begin{pmatrix} 1 & 1 & 1 & 0 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{pmatrix} \\
&= 1 \cdot \beta_1 + 1 \cdot \beta_2 + 1 \cdot \beta_3 + 0 \cdot \beta_4 \\
&= 0.64 + 0.030 + 0.025 = 0.695
\end{aligned}
$$

Plot all predictions ...
Next: how to find the ‘best’ coefficients.

Least squares

residual: $y_i - X_i\beta =$ the distance between the ith outcome and its prediction.
We get all residuals as: $y - X\beta$. FEV example:

$$y - X\beta = \begin{pmatrix} 1.72 - (1\beta_1 + 7\beta_2 + 54.5\beta_3 + 0\beta_4) \\ 1.74 - (1\beta_1 + 8\beta_2 + 54.0\beta_3 + 1\beta_4) \\ 2.09 - (1\beta_1 + 9\beta_2 + 59.5\beta_3 + 1\beta_4) \\ 3.13 - (1\beta_1 + 10\beta_2 + 62.0\beta_3 + 1\beta_4) \\ 2.87 - (1\beta_1 + 11\beta_2 + 60.5\beta_3 + 0\beta_4) \\ 2.57 - (1\beta_1 + 12\beta_2 + 63.0\beta_3 + 0\beta_4) \end{pmatrix}$$

Least Squares

- SSR: the sum of squared residuals:

$$\sum_{i=1}^{n}(y_i - X_i\beta)^2 = (y - X\beta)^t(y - X\beta).$$

- The t stands for transpose where the rows and columns of a matrix are swapped.
- X is an $n \times k$ matrix and β is a $k \times 1$ matrix (or vector) so the product is an $n \times 1$ matrix.
- The transpose turns an $n \times 1$ matrix into a $1 \times n$ matrix.
- The product of a $1 \times n$ matrix and a $n \times 1$ matrix is a 1×1 matrix, or a single number.

$$\begin{aligned}(y - X\beta)^t(y - X\beta) &= (1.72 - (1\beta_1 + 7\beta_2 + 54.5\beta_3 + 0\beta_4))^2 \\ &+ \cdots \\ &+ (2.57 - (1\beta_1 + 12\beta_2 + 63.0\beta_3 + 0\beta_4))^2\end{aligned}$$

Least Squares

- Least squares criterion: the best choice for β is the one that minimizes the sum of squared residuals, $\text{SSR} = (y - X\beta)^t(y - X\beta)$.
- In theory, to find this we could take derivatives of the SSR with respect to β_j, for $j = 1, \ldots, k$, set each of these k equations to 0 and solve.
- In matrix notation, taking derivatives and doing some matrix algebra leads to

$$X^t X\beta = X^t y$$

  each side is a $k \times 1$ vector: k equations.
  $X^t X$ is a $k \times k$ matrix.
- If the square matrix $X^t X$ has an inverse, then the equation is solved for β by

$$\hat{\beta} = (X^t X)^{-1} X^t y$$

Example

In our example,

$$X^tX = \begin{pmatrix} 6 & 57 & 354 & 3 \\ 57 & 559 & 3390 & 27 \\ 354 & 3390 & 20900 & 176 \\ 3 & 27 & 176 & 3 \end{pmatrix}$$

$$(X^tX)^{-1} = \begin{pmatrix} 144.178 & 6.047 & -3.443 & 2.844 \\ 6.047 & 0.386 & -0.167 & 0.247 \\ -3.443 & -0.167 & 0.086 & -0.095 \\ 2.844 & 0.247 & -0.095 & 0.834 \end{pmatrix}$$

Least Squares Solution

With

$$\hat{\beta} = (X^t X)^{-1} X^t y$$

we get

$$\hat{\beta} = \begin{pmatrix} -5.285 \\ 0.022 \\ 0.126 \\ 0.060 \end{pmatrix}, \quad y = \begin{pmatrix} 1.72 \\ 1.74 \\ 2.09 \\ 3.13 \\ 2.87 \\ 2.57 \end{pmatrix}, \quad \text{and} \quad \hat{y} = X\hat{\beta} = \begin{pmatrix} 1.72 \\ 1.73 \\ 2.45 \\ 2.78 \\ 2.56 \\ 2.89 \end{pmatrix}$$

Problem of confounded variables

What if $(X^t X)$ does not have an inverse? Examples:

- We add a smoking effect in the model, but all 6 sampled children are in non-smoking families.

  Two columns in X are equal

- Among sampled children, all females are in smoking families and all males are in non-smoking families.

  Two columns in X add up to a column of zeros

- Among sampled children, height = 7 * age.

  One column in X proportional to another.

Take-home message

- Quantitative and categorical explanatory variables all make it to the model matrix. Sometimes called the design matrix.
- I do not expect you to build and manipulate this matrix: it's R's business.
- Issues when this matrix does not have an inverse: completely correlated variables, confounded variables, etc. That can explain why R (or other sofware) decides that some effects are not estimable, and drop them.

Outline

1. Multiple Linear Regression
   - The model matrix
   - Least squares
   - Problem of confounded variables
2. Multiple Linear Regression: Case Study
   - Birds and Bats
   - Interpreting coefficients

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: 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`), 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
```

- 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.

Some plots

Box-and-whiskers plots:

```
> library(lattice)
> bwplot(mass   ~ type,data=bats)
> bwplot(energy ~ type,data=bats)
```

Scatterplot:

```
> plot(energy ~ mass, data = bats, col = type)
> legend("topleft", pch=1,col=1:3,
+         legend=levels(bats$type))
```

Observations

- 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 and show how these potential problems can be identified with residual plots.

Fitting Models

```
> fit0 = lm(energy ~ mass,         data = bats)
> fit1 = lm(energy ~ mass + type, data = bats)
> fit2 = lm(energy ~ mass * type, data = bats)
```

- `fit0` is a simple linear regression of energy on mass
- `fit1` adds `type` as an explanatory 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.

Plots of fitted models

```
> fit0.coef = coef(fit0)
> plot(energy~mass, data=bats, col=type)
> abline(fit0.coef[1], fit0.coef[2], col="grey")

> fit1.coef = coef(fit1)
> plot(energy~mass, data=bats, col=type)
> abline(fit1.coef[1],                fit1.coef[2],col=1)
> abline(fit1.coef[1]+fit1.coef[3],fit1.coef[2],col=2)
> abline(fit1.coef[1]+fit1.coef[4],fit1.coef[2],col=3)
> legend("topleft", pch=1,lty=1,col=1:3,legend=levels(bats$type))

> fit2.coef = coef(fit2)
> plot(energy~mass, data=bats, col=type)
> abline(fit2.coef[1],                fit2.coef[2],col=1)
> abline(fit2.coef[1]+fit2.coef[3],fit2.coef[2]+fit2.coef[5],col=2)
> abline(fit2.coef[1]+fit2.coef[4],fit2.coef[2]+fit2.coef[6],col=3)
> legend("topleft", pch=1,lty=1,col=1:3,legend=levels(bats$type))
```

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`.
- birds: intercept $= 6.02$.
  echolocating bats: intercept $= 6.02 + (-4.6) = 1.42$.
  non-echolocating bats: intercept $= 6.02 + (-3.43) = 2.59$.
- The three lines are parallel: common slope $= 0.0575$.

Estimated Coefficients

```
> coef(fit2)
  (Intercept)          mass      typeeBat      typenBat
   3.31674159    0.06777464   -2.82275855    7.91064213
mass:typeeBat mass:typenBat
   0.02186199   -0.02772895
```

Estimated coefficients: intercept and slope for `bird`, and then adjustments to each of these for the other types.

- For birds: intercept $= 3.32$ and slope $= 0.0678$.
- Echolocating bats: intercept $= 3.32 + (-2.82) = 0.494$ and slope $= 0.0678 + (0.0219) = 0.0896$
- Non-echolocating bats: intercept $= 3.32 + (7.91) = 11.2$ and slope $= 0.0678 + (-0.0277) = 0.04$

Interpretation of Coefficients

```
> coef(fit2)
  (Intercept)          mass      typeeBat      typenBat
   3.31674159    0.06777464   -2.82275855    7.91064213
mass:typeeBat mass:typenBat
   0.02186199   -0.02772895
```

- Intercept: predicted energy of a bird at mass 0 — no biological relevance.
- 3d coefficient: estimated difference between the predicted energies for echolocating bats and birds at mass 0 — no biological relevance.
- *The predicted difference is not the same at all masses.* Ex: 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

Residual plot from last fit:

```
> plot(residuals(fit2) ~ fitted(fit2))
```

- Fan-shaped pattern: residuals are larger for large mass.
- A transformation may help.

Log transformed data

```
> plot(log(energy) ~ log(mass), data = bats, col=type)
> legend("topleft", pch=1,col=1:3,
+                                legend=levels(bats$type))
```

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