Multiple Linear Regression

Bret Larget

Departments of Botany and of Statistics
University of Wisconsin—Madison

February 6, 2007

Multiple Linear Regression

- Most interesting questions in biology involve relationships between multiple variables.
- There are typically multiple explanatory variables.
- Interactions between variables can be important in understanding a process.
- We will now study statistical models for when there is a single continuous quantitative response variable and multiple explanatory variables.
- Explanatory variables may be quantitative or factors (categorical variables).

Model

- We extend simple linear regression to consider models of the following form:
$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + e_i$$
where $e_i \sim \mathrm{iid}\ \mathrm{N}(0, \sigma^2) \quad \text{for } i = 1, \ldots, n$.
- y is the *response variable*;
- $x_1, x_2, \ldots, x_k$ are the *explanatory variables*;
- Some people use the terms *dependent* and *independent* variables.
- I do not like this terminology because the x_i are often not independent.
- e_i are random errors;
- β_0 is an intercept and $\beta_1, \ldots, \beta_k$ are slopes.

Multiple Regression Objectives

- Inference (estimation and testing) on the model parameters;
- Estimation/prediction of y at $x_1^*, x_2^*, \ldots, x_k^*$
- Model selection: Select which explanatory variables are best to include in a model.

Estimation of Regression Coefficients

- We extend the least squares criterion from SLR.
- Seek the parameters $b_0, \ldots, b_k$ that minimize

$$\sum_{i=1}^{n}(y_i - (b_0 + b_1 x_{1i} + b_2 x_{2i} + \cdots + b_k x_{ki}))^2$$

- The solution is the set of estimated coefficients $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$.
- The ith fitted value is $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \cdots + \hat{\beta}_k x_{ki}$.
- The ith residual is $y_i - \hat{y}_i$.
- The least square criterion minimizes the sum of the squared residuals, also called the sum of squares for error (SSErr).
- The estimate of the variance σ^2 is the mean squared error (MSErr) or $\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-(k+1)}$.

Matrix Notation for Estimates

- There are no simple expressions for the estimated coefficients.
- The matrix notation solution is concise.

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \qquad X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{k1} \\ 1 & x_{12} & \cdots & x_{k2} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{1n} & \cdots & x_{kn} \end{pmatrix}, \qquad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}$$

- $\hat{\beta} = (X^T X)^{-1} X^T y$
- $\hat{y} = X\hat{\beta} = X(X^T X)^{-1} X^T y = Hy$
- The matrix H is called the *hat matrix*. The diagonal entries are the *leverages*.

$k = 2$ Case

- The model is

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + e_i$$

  where $e_i \sim \mathrm{iid}\ \mathrm{N}(0, \sigma^2)$ for $i = 1, \ldots, n$.
- Intercept β_0: expected y when $x_1 = 0, x_2 = 0$.
- Slope β_1: expected change in y for 1 unit increase in x_1 with x_2 held constant.
- Slope β_2: expected change in y for 1 unit increase in x_2 with x_1 held constant.

Formula

$$\begin{aligned}
\hat{\beta}_1 &= \frac{\sum(y_i - \bar{y})(x_{1i} - \bar{x}_1) - \frac{[\sum(y_i - \bar{y})(x_{2i} - \bar{x}_2)][\sum(x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2)]}{\sum(x_{2i} - \bar{x}_2)^2}}{\sum(x_{1i} - \bar{x}_1)^2 - \frac{[\sum(x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2)]^2}{\sum(x_{2i} - \bar{x}_2)^2}} \\
\hat{\beta}_2 &= \frac{\sum(y_i - \bar{y})(x_{2i} - \bar{x}_2) - \frac{[\sum(y_i - \bar{y})(x_{1i} - \bar{x}_1)][\sum(x_{2i} - \bar{x}_2)(x_{1i} - \bar{x}_1)]}{\sum(x_{1i} - \bar{x}_1)^2}}{\sum(x_{2i} - \bar{x}_2)^2 - \frac{[\sum(x_{2i} - \bar{x}_2)(x_{1i} - \bar{x}_1)]^2}{\sum(x_{1i} - \bar{x}_1)^2}} \\
\hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x}_1 - \hat{\beta}_2 \bar{x}_2.
\end{aligned}$$

Pesticide Example

- A study was conducted to assess the toxic effect of a pesticide on a given species of insect.
- The data consist of:
  - dose rate of the pesticide (x_1, units unknown)
  - body weight of an insect (x_2, grams, maybe?)
  - rate of toxic action (y, time to death in minutes, maybe?).

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

```
'data.frame': 19 obs. of  3 variables:
 $ dose  : num  0.696 0.729 0.509 0.559 0.679 0.583 0.742 0.781 0.8
 $ weight: num  0.321 0.354 0.134 0.184 0.304 0.208 0.367 0.406 0.4
 $ effect: num  0.324 0.367 0.321 0.375 0.345 0.341 0.327 0.256 0.2
```


Analysis

- Use R to show graphical analysis
- Use R to show differences in possible models to fit.

```
> attach(toxic)
> fit0 = lm(effect ~ 1)
> fit1 = lm(effect ~ dose)
> fit2 = lm(effect ~ weight)
> fit12 = lm(effect ~ dose + weight)
> fit21 = lm(effect ~ weight + dose)
```