Multiple Linear Regression

Bret Larget
Departments of Botany and of Statistics
University of Wisconsin—Madison
February 6, 2007

The Big Picture

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 \( e_i \sim \text{iid } N(0, \sigma^2) \) 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;
- \( \beta_0 \) 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
  \[
  \sum_{i=1}^{n} (y_i - (b_0 + b_1x_{i1} + b_2x_{i2} + \cdots + b_kx_{ki}))^2
  \]
- The solution is the set of estimated coefficients $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$.
- The $i$th fitted value is $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1x_{i1} + \cdots + \hat{\beta}_kx_{ki}$.
- The $i$th 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 $\sigma^2$ is the mean squared error (MSErr) or $\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-(k+1)}$.

$k = 2$ Case

- The model is
  \[
  y_i = \beta_0 + \beta_1x_{i1} + \beta_2x_{i2} + e_i
  \]
- where $e_i \sim \text{iid } N(0, \sigma^2)$ for $i = 1, \ldots, n$.
- Intercept $\beta_0$: expected $y$ when $x_1 = 0, x_2 = 0$.
- Slope $\beta_1$: expected change in $y$ for 1 unit increase in $x_1$ with $x_2$ held constant.
- Slope $\beta_2$: expected change in $y$ for 1 unit increase in $x_2$ with $x_1$ held constant.

Matrix Notation for Estimates

- There are no simple expressions for the estimated coefficients.
- The matrix notation solution is concise.
- \[
  \begin{pmatrix}
  y_1 \\
  y_2 \\
  \vdots \\
  y_n
  \end{pmatrix} = \begin{pmatrix}
  1 & x_{11} & \cdots & x_{1k} \\
  1 & x_{21} & \cdots & x_{2k} \\
  \vdots & \vdots & \ddots & \vdots \\
  1 & x_{n1} & \cdots & x_{nk}
  \end{pmatrix} \begin{pmatrix}
  \beta_0 \\
  \beta_1 \\
  \vdots \\
  \beta_k
  \end{pmatrix}
  \]
- $\hat{\beta} = (X^TX)^{-1}X^Ty$
- $\hat{y} = X\hat{\beta} = X(X^TX)^{-1}X^Ty = Hy$
- The matrix $H$ is called the hat matrix. The diagonal entries are the leverages.

Formula

- $\hat{\beta}_1 = \frac{\sum(y_i - \bar{y})(x_{i1} - \bar{x_1}) - \sum(x_{i1} - \bar{x_1})(y_i - \bar{y})(x_{i2} - \bar{x_2})}{\sum(x_{i1} - \bar{x_1})^2 - \sum(x_{i2} - \bar{x_2})^2}$
- $\hat{\beta}_2 = \frac{\sum(y_i - \bar{y})(x_{i2} - \bar{x_2}) - \sum(y_i - \bar{y})(x_{i1} - \bar{x_1})(x_{i2} - \bar{x_2})}{\sum(x_{i2} - \bar{x_2})^2 - \sum(x_{i1} - \bar{x_1})(x_{i2} - \bar{x_2})}$
- $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x_1} - \hat{\beta}_2\bar{x_2}$. 
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?).

```r
> 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 0.865 0.723 ...
$ weight: num 0.321 0.354 0.134 0.184 0.304 0.208 0.367 0.406 0.490 0.223 ...
$ effect: num 0.324 0.367 0.321 0.375 0.345 0.341 0.327 0.256 0.214 0.501 ...
```

**Analysis**

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

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