Multiple regression

Situation: more than one independent variable; want to predict Y from x_1, x_2, ..., x_k.

ex:
- IRS predicts the amount of money to be recovered in an audit using (among other variables) amt. of deduction for charitable gifts, amt. of real estate losses, etc.

- House appraiser predicts sale price of a house based on sq. ft., # bedrooms, ave. sale price in neighborhood, etc.

Idea: mean of Y is taken to be a linear function of the predictor variables:

$$\mathrm{E}(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \cdots + \beta_k x_k$$

With just two predictor variables (not functionally dependent), this equation is a **plane**.

Model:

$$Y \sim \text{normal}\big(\beta_0 + \beta_1 x_1 + \beta_2 x_2 \cdots + \beta_k x_k,\, \sigma^2\big)$$

Different types of predictor variables:

- ordinary quantitative variables
- indicator variables (AOV is a regression!)

3 treatments; means μ_1, μ_2, μ_3

$$x_1 = \begin{cases} 1 \text{ if observation is from trt 1} \\ 0 \text{ otherwise} \end{cases}$$

$$x_2 = \begin{cases} 1 \text{ if observation is from trt 2} \\ 0 \text{ otherwise} \end{cases}$$

$$\mathrm{E}(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

$$\mu_1 = \beta_0 + \beta_1$$

$$\mu_2 = \beta_0 + \beta_2$$

$$\mu_3 = \beta_0$$

- nonlinear terms, e.g.

$$\mathrm{E}(Y) = \beta_0 + \beta_1 x + \beta_2 x^2$$

(Note: this is still a **linear statistical model** because the parameters appear linearly)

- interactions, e.g.

$$\mathrm{E}(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$

Estimates of unknown parameters are conveniently represented with **matrix notation**

(your task: learn to multiply matrices; learn what an inverse matrix is)

Data:

$$\begin{array}{l} (y_1, x_{11}, x_{12}, ..., x_{1k}) \\ (y_2, x_{21}, x_{22}, ..., x_{2k}) \\ \qquad\qquad \vdots \\ (y_n, x_{n1}, x_{n2}, ..., x_{nk}) \end{array}$$

Matrices:

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}_{(n\times(k+1))}$$

(design matrix)

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}_{(n\times 1)} \qquad \widehat{\boldsymbol{\beta}} = \begin{bmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \\ \widehat{\beta}_2 \\ \vdots \\ \widehat{\beta}_k \end{bmatrix}_{((k+1)\times 1)} \qquad \text{ML, LS}$$

Matrix representations:

$$\widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}} \quad \text{predicted values}$$

$$\mathbf{e} = \left(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right) \quad \text{residuals}$$

$$\text{SS(residuals)} = \left(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right)'\left(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right)$$

sum of squared errors (minimized at $\widehat{\boldsymbol{\beta}}$)

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} \\ 1 & x_{21} & x_{22} \\ 1 & x_{31} & x_{32} \\ 1 & x_{41} & x_{42} \end{bmatrix} \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} \quad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix}$$

$$\mathbf{X}\boldsymbol{\beta} = \begin{bmatrix} \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} \\ \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} \\ \beta_0 + \beta_1 x_{31} + \beta_2 x_{32} \\ \beta_0 + \beta_1 x_{41} + \beta_2 x_{42} \end{bmatrix}$$

$$\mathbf{X'X} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ x_{11} & x_{21} & x_{31} & x_{41} \\ x_{12} & x_{22} & x_{32} & x_{42} \end{bmatrix} \begin{bmatrix} 1 & x_{11} & x_{12} \\ 1 & x_{21} & x_{22} \\ 1 & x_{31} & x_{32} \\ 1 & x_{41} & x_{42} \end{bmatrix}$$

$$= \begin{bmatrix} n & \Sigma x_{i1} & \Sigma x_{i2} \\ \Sigma x_{i1} & \Sigma x_{i1}^2 & \Sigma x_{i1} x_{i2} \\ \Sigma x_{i2} & \Sigma x_{i1} x_{i2} & \Sigma x_{i2}^2 \end{bmatrix}$$

$$(\mathbf{X'X})^{-1}(\mathbf{X'X}) = \mathbf{I} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$\mathbf{X}'\mathbf{X}$ matrix of sums of squares & crossproducts

of the predictors (symmetric, $(k+1) \times (k+1)$)

$$(\mathbf{X}'\mathbf{X})\widehat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y} \quad \text{“normal equations”}$$

(minimizing the sum of squared errors results in
a system of $k+1$ linear equations in $k+1$ unknowns)

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

ML & LS estimates of the parameters β_0, β_1, ..., β_k

$$\widehat{\sigma}^2 = \frac{1}{(n-k-1)}\left(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right)'\left(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right)$$

unbiased (adjusted ML) estimate of σ^2

Note: different ways of writing the sum of squared residuals result from matrix-algebraic rearrangements

$$\left(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right)'\left(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right) = \mathbf{y}'\mathbf{y} - \widehat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y}$$

$$= \mathbf{y}'[\mathbf{I} - \mathbf{H}]\mathbf{y}$$

where $\mathbf{I}$ = identity matrix, $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ “hat matrix”

Inferences for multiple regression

The "AOV table"

source of variation	SS	df	MS
regression	SS(regression)	k	$\frac{\text{SS(regression)}}{k}$
error	SS(residual)	$n-k-1$	$\frac{\text{SS(residual)}}{n-k-1}$
total	SS(total)	$n-1$	

$$\text{SS(regression)} = \left(\mathbf{X}\widehat{\boldsymbol{\beta}} - \tfrac{1}{n}\mathbf{J}\mathbf{y}\right)'\left(\mathbf{X}\widehat{\boldsymbol{\beta}} - \tfrac{1}{n}\mathbf{J}\mathbf{y}\right)$$
$$\text{SS(residual)} = \left(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right)'\left(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right)$$
$$\text{SS(total)} = \left(\mathbf{y} - \tfrac{1}{n}\mathbf{J}\mathbf{y}\right)'\left(\mathbf{y} - \tfrac{1}{n}\mathbf{J}\mathbf{y}\right)$$

where $\mathbf{J}$ is an $n \times n$ matrix of 1's

Test of the model vs the mean

H_0: $\beta_1 = \beta_2 = \cdots = \beta_k = 0$
(just the overall mean is used for predicting Y)

H_a: $\neq$
(all predictor variables are included in the model)

Test statistic (from LR): $f = \frac{\text{SS(regression)}/k}{\text{SS(residual)}/(n-k-1)}$

Rejection region: reject H_0 if $f \geq f_\alpha$ where the F distribution has k and $n-k-1$ df.

Coefficient of determination

$$r^2 = \frac{\text{SS(regression)}}{\text{SS(total)}} = 1 - \frac{\text{SS(residual)}}{\text{SS(total)}}$$

(proportional reduction in the prediction error attained by using the multiple regression model instead of the overall mean)