Estimation

Based on a book by Julian J. Faraway

University of Iceland

Setup

Suppose that Y is the fuel consumption of a particular model of a car. Suppose that the predictors are:

- X_1 - the weight of the car
- X_2 - the horse power
- X_3 - the no. of cylinders.

Typically the data will be available in the form of an array like this

$$\begin{array}{cccc} y_1 & x_{11} & x_{12} & x_{13} \\ y_2 & x_{21} & x_{22} & x_{23} \\ \vdots & \vdots & \vdots & \vdots \\ y_n & x_{n1} & x_{n2} & x_{n3} \end{array}$$

where n is the number of observations or cases in the dataset.

Linear model

One very general form for the model would be

$$Y = f(X_1, X_2, X_3) + \varepsilon$$

where f is some unknown function and ε is the error.

Since we usually don't have enough data to try to estimate f directly, we usually have to assume that it has some more restricted form, perhaps linear as in

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$$

where $\beta_i, i = 0, 1, 2, 3$ are unknown parameters. Thus the problem is reduced to the estimation of four values rather than the complicated infinite dimensional f.

In a linear model the parameters enter linearly - the predictors do not have to be linear.

Matrix Representation

Given actual data we might write:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i \quad i = 1, 2, ..., n$$

but it is simpler to use a matrix/vector representation:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

where $\boldsymbol{y} = (y_1, ..., y_n)^T$, $\boldsymbol{\varepsilon} = (\varepsilon_1, ..., \varepsilon_n)^T$, $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)^T$, $\text{var}[\boldsymbol{\varepsilon}] = \sigma^2 \boldsymbol{I}$ and

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & x_{13} \\ 1 & x_{21} & x_{22} & x_{23} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & x_{n3} \end{bmatrix}$$

X is a *matrix* of dimension $n \times p$ where $p - 1$ is the number of predictors.

Least squares estimation

The least squares estimate of β, often called $\hat{\boldsymbol{\beta}}$ minimizes

$$\sum \varepsilon_i^2 = \boldsymbol{\varepsilon}^{\boldsymbol{T}} \boldsymbol{\varepsilon} = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^{\boldsymbol{T}} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$$

expanding gives

$$\boldsymbol{y}^{\boldsymbol{T}}\boldsymbol{y} - 2\boldsymbol{\beta}^{\boldsymbol{T}}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{y} + \boldsymbol{\beta}^{\boldsymbol{T}}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X}\boldsymbol{\beta}$$

Differentiating with respect to $\boldsymbol{\beta}$ and setting to zero, we find that $\hat{\boldsymbol{\beta}}$ satisfies

$$\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{y}$$

These are called the normal equations. Now provided that $\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X}$ is invertible we get:

$$\begin{aligned} \hat{\boldsymbol{\beta}} = & \quad (\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{y} \\ \boldsymbol{X}\hat{\boldsymbol{\beta}} = & \quad \boldsymbol{X}(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{y} \\ = & \quad H\boldsymbol{y} \end{aligned}$$

where $H = \boldsymbol{X}(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\boldsymbol{T}}$ is the *hat matrix*.

Least squares estimation

The Gauss-Markov theorem shows that when the errors have expectation zero, are uncorrelated and have equal variances, the best (the lowest variance) linear unbiased estimator (BLUE) of the coefficients is given by the ordinary least squares (OLS) estimator.

Situations where estimators other than ordinary least squares should be considered are

1. When the errors are correlated or have unequal variance, generalized least squares should be used.
2. When the error distribution is long-tailed, then robust estimates might be used.
3. When the predictors are highly correlated (collinear), then biased estimators such as ridge regression might be preferable.

Mean and variance of $\hat{\beta}$

$$\begin{aligned} \mathrm{E}[\hat{\boldsymbol{\beta}}] &= \boldsymbol{\beta} \ \ (\text{unbiased}) \\ \mathrm{var}[\hat{\boldsymbol{\beta}}] &= (\boldsymbol{X^T X})^{-1}\boldsymbol{\sigma^2} \end{aligned}$$

Note that since $\hat{\boldsymbol{\beta}}$ is a vector, $(\boldsymbol{X^T X})^{-1}\boldsymbol{\sigma^2}$ is a variance-covariance matrix.

Often we want the standard error for a particular component which can be picked out as in

$$\mathrm{se}(\hat{\beta}_i) = \sqrt{(\boldsymbol{X^T X})^{-1}_{ii}}\hat{\sigma}$$

Estimating σ^2

An estimate of $\hat{\sigma}^2$ can be found using

$$\hat{\sigma}^2 = \frac{\boldsymbol{\varepsilon}^{\boldsymbol{T}}\boldsymbol{\varepsilon}}{n-p}$$

where p is the number of parameters in the model (including the intercept term).

Goodness of fit

One measure on how well our model fits the data is, R^2, the so called *coefficient of variation* or *coefficient of determination* or simply *percentage of variation explained*:

$$R^2 = 1 - \frac{\sum(\hat{y}_i - y_i)^2}{\sum(y_i - \bar{y})^2} = 1 - \frac{RSS}{TotalSS}$$

RSS is often denoted by SSE and $TotalSS$ with SST.

Matrices in R

- `as.matrix()` can be used to change data frames to matrices
- Matrix multiplication is done with %*%
- `t()` returns the transpose
- `solve()` is used to find an inverse of a matrix