Multiple Linear Regression

Hong Kong Baptist University

Fall 2021

Multiple Linear Regression

Multiple linear regression model:

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki} + \epsilon_i,\ i = 1, \ldots, n, \tag{1}$$

where $n \geq k + 1$, $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ are i.i.d, and x_{ji} is the i^{th} observation of the j^{th} independent variable.

General linear model is linear in its parameters. The independent variables can be higher order terms like x^2 or $\log x$. That is, the model

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 \log x + \varepsilon$$

is still a linear model; it is linear in its parameters. An example of nonlinear model is

$$Y = \beta_0 + \beta_1 x^{\beta_2} + \varepsilon,$$

which is not linear in its parameters.

The model given in (1) is in fact a system of linear equations, and can be expressed in matrix terms as follows:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{2}$$

where

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \ \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{21} & \cdots & x_{k1} \\ 1 & x_{12} & x_{22} & \cdots & x_{k2} \\ \vdots & \vdots & & & \vdots \\ 1 & x_{1n} & x_{2n} & \cdots & x_{kn} \end{bmatrix}, \ \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \ \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}.$$

Note that x_{ij} is the i^{th} row j^{th} column element of the design matrix $\mathbf{X}$. We have a column of 1's in $\mathbf{X}$ because we include β_0 in $\boldsymbol{\beta}$ so that in equation (2) we do not need a separate intercept term.

Revision of matrix algebra

Let us denote a matrix by

$$\boldsymbol{A}_{n\times m} = \boldsymbol{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nm} \end{bmatrix} = [a_{ij}]_{n\times m}.$$

Matrix addition/subtraction is done elementwise:

$$\boldsymbol{A} \pm \boldsymbol{B} = [a_{ij} \pm b_{ij}],$$

and so is scalar multiplication:

$$k\boldsymbol{A} = [ka_{ij}],$$

but matrix multiplication is more complicated:

$$\boldsymbol{A}_{n\times m}\boldsymbol{B}_{m\times p} = \left[\sum_{k=1}^{m} a_{ik}b_{kj}\right]_{n\times p},$$

and so in general $\boldsymbol{AB} \neq \boldsymbol{BA}$.

The transpose of a matrix $\boldsymbol{A} = [a_{ij}]_{n\times m}$ is denoted by $\boldsymbol{A}'$, which is defined by

$$\boldsymbol{A}' = [a_{ji}]_{m\times n},$$

and $(\boldsymbol{AB})' = \boldsymbol{B}'\boldsymbol{A}'$. Obviously $(\boldsymbol{A}')' = \boldsymbol{A}$.

If $\boldsymbol{A}' = \boldsymbol{A}$, then $\boldsymbol{A}$ is symmetric; it is of course a square matrix, in which $n = m$. A simple example of a symmetric matrix is $\boldsymbol{A}'\boldsymbol{A}$ because $(\boldsymbol{A}'\boldsymbol{A})' = \boldsymbol{A}'\boldsymbol{A}$.

The identity matrix $\boldsymbol{I}$ is the square matrix that the diagonal elements are all 1, whilst all others are zero, and so $\boldsymbol{I}\boldsymbol{A} = \boldsymbol{A}\boldsymbol{I} = \boldsymbol{A}$.

The inverse of a square matrix , denoted by $\boldsymbol{A}^{-1}$, if exists, is the unique matrix satisfying $\boldsymbol{A}\boldsymbol{A}^{-1} = \boldsymbol{A}^{-1}\boldsymbol{A} = \boldsymbol{I}$. If $\boldsymbol{A}^{-1}$ exists, then we say that $\boldsymbol{A}$ is non-singular.

A system of linear equations, expressed in matrix terms, is

$$\boldsymbol{A}\boldsymbol{x} = \hat{\boldsymbol{\beta}},$$

and the solution is simply

$$\boldsymbol{x} = \boldsymbol{A}^{-1}\hat{\boldsymbol{\beta}}.$$

Let $\boldsymbol{x} = [x_1, x_2, \ldots, x_n]'$ (a column vector) and y be a real-valued function of x_1, x_2, ..., x_n, then define the derivative of y with respect to the column vector $\boldsymbol{x}$ to be the column vector of partial derivatives:

$$\frac{\partial y}{\partial \boldsymbol{x}} = \begin{bmatrix} \frac{\partial y}{\partial x_1} \\ \frac{\partial y}{\partial x_2} \\ \vdots \\ \frac{\partial y}{\partial x_n} \end{bmatrix}_{n\times 1}.$$

If $\boldsymbol{a} = [a_1, a_2, \ldots, a_n]'$ (a vector of constants; constants mean they are not functions of $\boldsymbol{x}$) and $y = \boldsymbol{a}'\boldsymbol{x} = a_1x_1 + a_2x_2 + \cdots + a_nx_n$, then we have

$$\frac{\partial \boldsymbol{a}'\boldsymbol{x}}{\partial \boldsymbol{x}} = \frac{\partial y}{\partial \boldsymbol{x}} = \begin{bmatrix} \frac{\partial y}{\partial x_1} \\ \frac{\partial y}{\partial x_2} \\ \\ \vdots \\ \\ \frac{\partial y}{\partial x_n} \end{bmatrix} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix} = \boldsymbol{a}.$$

Suppose $\boldsymbol{A}$ is a matrix of constants:

$$\boldsymbol{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nm} \end{bmatrix}_{n \times m},$$

and

$$\boldsymbol{Y} = \boldsymbol{x}'\boldsymbol{A} = [x_1, x_2, \ldots, x_n] \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nm} \end{bmatrix}$$

$$= [(a_{11}x_1 + a_{21}x_2 + \cdots + a_{n1}x_n), \ \ldots, \ (a_{1m}x_1 + a_{2m}x_2 + \cdots + a_{nm}x_n)]$$

Then for the row vector $\boldsymbol{Y} = \boldsymbol{x}'\boldsymbol{A}$, its partial derivative with respect to the column vector x will be a matrix:

$$\begin{aligned}
\frac{\partial \boldsymbol{x}'\boldsymbol{A}}{\partial \boldsymbol{x}} &= \frac{\partial \boldsymbol{Y}}{\partial \boldsymbol{x}} = \left[\frac{\partial y_1}{\partial \boldsymbol{x}}, \frac{\partial y_2}{\partial \boldsymbol{x}}, \ldots, \frac{\partial y_m}{\partial \boldsymbol{x}}\right] \\
&= \begin{bmatrix} \frac{\partial \boldsymbol{Y}}{\partial x_1} \\ \frac{\partial \boldsymbol{Y}}{\partial x_2} \\ \vdots \\ \frac{\partial \boldsymbol{Y}}{\partial x_n} \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nm} \end{bmatrix} = \boldsymbol{A}. \qquad (3)
\end{aligned}$$

If $y = \boldsymbol{x}'\boldsymbol{A}\boldsymbol{x} = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}x_i x_j$, then applying the product rule:

$$\frac{\partial \boldsymbol{u}'\boldsymbol{A}\boldsymbol{v}}{\partial \boldsymbol{x}} = \frac{\partial \boldsymbol{u}'}{\partial \boldsymbol{x}}\boldsymbol{A}\boldsymbol{v} + \frac{\partial \boldsymbol{v}'}{\partial \boldsymbol{x}}\boldsymbol{A}'\boldsymbol{u},$$

for vector-valued functions $\boldsymbol{u}$, $\boldsymbol{v}$ of $\boldsymbol{x}$, we have

$$\frac{\partial \boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}}{\partial \boldsymbol{x}} = \frac{\partial y}{\partial \boldsymbol{x}} = \begin{bmatrix} \frac{\partial \sum\sum a_{ij}x_i x_j}{\partial x_1} \\ \vdots \\ \frac{\partial \sum\sum a_{ij}x_i x_j}{\partial x_n} \end{bmatrix} = \frac{\partial \boldsymbol{x}'}{\partial \boldsymbol{x}}\boldsymbol{A}\boldsymbol{x} + \frac{\partial \boldsymbol{x}'}{\partial \boldsymbol{x}}\boldsymbol{A}'\boldsymbol{x} = (\boldsymbol{A} + \boldsymbol{A}')\,\boldsymbol{x}.$$

In particular, if $\boldsymbol{A}$ is symmetric so that $\boldsymbol{A} = \boldsymbol{A}'$, then we have

$$\frac{\partial \boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}}{\partial \boldsymbol{x}} = \frac{\partial y}{\partial \boldsymbol{x}} = 2\boldsymbol{A}\boldsymbol{x}. \tag{4}$$

Least Squares Estimation

Now, the least squares estimation for the model parameter $\boldsymbol{\beta}$ of the multiple regression model

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

can be obtained straightforwardly. Consider

$$\boldsymbol{\varepsilon} = \boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}.$$

The sum of squares of errors (which is a scalar) is equal to

$$Q = \boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} = (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) = \boldsymbol{Y}'\boldsymbol{Y} - 2\boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{Y} + \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta}$$

(notice that $\boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{Y}$ and $\boldsymbol{Y}'\boldsymbol{X}\boldsymbol{\beta}$ are the same because they are scalars).

To minimize Q we take the partial derivative with respect to $\boldsymbol{\beta}$. Using equations (3) and (4) above, we get

$$\frac{\partial Q}{\partial \boldsymbol{\beta}} = -2\boldsymbol{X}'\boldsymbol{Y} + 2\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta}.$$

The least squares estimator $\widehat{\boldsymbol{\beta}}$ for the parameter $\boldsymbol{\beta}$ should satisfy the *normal equation*

$$\left.\frac{\partial Q}{\partial \boldsymbol{\beta}}\right|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}} = \mathbf{0} \Rightarrow \boxed{\boldsymbol{X}'\boldsymbol{Y} = \boldsymbol{X}'\boldsymbol{X}\widehat{\boldsymbol{\beta}}},$$

and its solution is

$$\boxed{\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}}. \tag{5}$$

From this elegant expression, we can see immediately that if $\boldsymbol{\varepsilon}$ has a normal distribution, then $\boldsymbol{Y}$ has a normal distribution, and so does $\widehat{\boldsymbol{\beta}}$ [which is a non-random matrix $(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ times the normally distributed $\boldsymbol{Y}$].

Multivariate Normal Distribution

We say a column vector has a multivariate normal distribution, denoted by

$$\boldsymbol{X}_{n\times 1} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}),$$

if and only if the joint probability density function $f_{\boldsymbol{X}}$ is given by

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \left(\frac{1}{\sqrt{2\pi}}\right)^n \frac{1}{\sqrt{|\boldsymbol{\Sigma}|}} \mathrm{e}^{-\frac{(\boldsymbol{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}{2}}, \qquad \text{for all } \boldsymbol{x} \in \mathbb{R}^n,$$

where $|\boldsymbol{\Sigma}|$ denotes the determinant of $\boldsymbol{\Sigma}$.

Maximum Likelihood Estimation

Now, if $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{I})$ (i.e. ε_i are i.i.d. normal with mean zero and variance σ^2; the independence between ε_i results in a diagonal covariance matrix because the covariance between ε_i and ε_j is the (i, j)-element in $\boldsymbol{\Sigma}$; the identical distribution assumption results in one common value σ^2 for all diagonal elements in $\boldsymbol{\Sigma}$), then the likelihood is just

$$\ell(\boldsymbol{\beta}) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \mathrm{e}^{-\frac{\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}}{2\sigma^2}} = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \mathrm{e}^{-\frac{(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})}{2\sigma^2}}.$$

So maximizing ℓ is equivalent to minimizing $(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})$, i.e.

$$\text{MLE} = \text{LSE}. \tag{6}$$

Note that the equivalence given in (6) is true if we have normally distributed errors. If the errors follow another distribution, it may not be true. For example, if the errors are i.i.d., following the Laplace distribution, i.e.

$$f(\varepsilon) = \frac{1}{2\sigma} \mathrm{e}^{-\frac{|\varepsilon|}{\sigma}}, \qquad \text{for all } \varepsilon \in \mathbb{R},$$

then the likelihood is given by

$$\ell(\boldsymbol{\beta}) = \left(\frac{1}{2\sigma}\right)^n \mathrm{e}^{-\frac{\sum_i |\varepsilon_i|}{\sigma}},$$

and so maximizing the likelihood ℓ is equivalent to minimizing $\sum_i |\varepsilon_i|$ (least absolute deviations, or LAD for short).

Under LSE, each deviation is squared, i.e. each of them is weighted by itself, while under LAD, each deviation carries the same weight. Therefore, an outlier will have a stronger influence on the LSE than on the LAD, as a larger deviation will play a more dominant role in LSE than in LAD.

Geometry of the LSE

Recall that two vectors $\boldsymbol{p}$ and $\boldsymbol{q}$ are *orthogonal* if $\boldsymbol{p}'\boldsymbol{q} = \boldsymbol{q}'\boldsymbol{p} = 0$. Now, consider the fitted value $\hat{\boldsymbol{Y}}$ and residual $\boldsymbol{e}$:

$$\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} \equiv \boldsymbol{H}\boldsymbol{Y}, \ \boldsymbol{e} = \boldsymbol{Y} - \hat{\boldsymbol{Y}} = \boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y},$$

where $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ is the hat matrix. The actual observation $\boldsymbol{Y}$ is in $\mathbb{R}^n$. Because

$$\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \hat{\beta}_0 \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} + \hat{\beta}_1 \begin{bmatrix} x_{11} \\ x_{12} \\ \vdots \\ x_{1n} \end{bmatrix} + \cdots + \hat{\beta}_k \begin{bmatrix} x_{k1} \\ x_{k2} \\ \vdots \\ x_{kn} \end{bmatrix}$$

the fitted value $\hat{\boldsymbol{Y}}$ is lying in the so-called *estimation space*, which is the space generated by the $p = k + 1$ columns of $\boldsymbol{X}$ (we use p to represent the number of parameters, including the intercept β_0), whilst the residual $\boldsymbol{e}$ is lying in the *error space*, which has $n - p = n - k - 1$ dimensions.

Since

$$\hat{\boldsymbol{Y}}'\boldsymbol{e} = (\boldsymbol{X}\hat{\boldsymbol{\beta}})'\boldsymbol{e} = \hat{\boldsymbol{\beta}}'\boldsymbol{X}'(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) = \hat{\boldsymbol{\beta}}'(\boldsymbol{X}'\boldsymbol{Y} - \boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}}) = 0,$$

(note that $\boldsymbol{X}'\boldsymbol{Y} = \boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}}$ is the normal equation), we can see that the fitted value $\hat{\boldsymbol{Y}}$ and residual $\boldsymbol{e}$ are orthogonal, and hence the name *normal* equation, because under which the residual $\boldsymbol{e}$ is a normal vector (i.e. a vector perpendicular) to the estimation space.

Therefore, the least squares fitting procedure splits the space $\mathbb{R}^n$ into two orthogonal spaces; every vector in the estimation space is orthogonal to every vector in the error space, and the fitted value $\hat{\boldsymbol{Y}}$ is the orthogonal projection of the observed $\boldsymbol{Y}$ to the estimation space ($\hat{\boldsymbol{Y}} = \boldsymbol{H}\boldsymbol{Y}$, $\boldsymbol{H}\hat{\boldsymbol{Y}} = \boldsymbol{H}\boldsymbol{H}\boldsymbol{Y} = \boldsymbol{H}\boldsymbol{Y} = \hat{\boldsymbol{Y}}$, $\boldsymbol{e}'\hat{\boldsymbol{Y}} = 0$), so that no other vector in the estimation space is closer (measured in the Euclidean distance) to $\boldsymbol{Y}$ than $\hat{\boldsymbol{Y}}$ is.

The unbiasedness of $\hat{\boldsymbol{\beta}}$ (which will be shown below) implies

$$\mathbb{E}(\boldsymbol{e}) = \mathbb{E}(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) = \mathbb{E}(\boldsymbol{X}\boldsymbol{\beta}) + \mathbb{E}(\boldsymbol{\varepsilon}) - \mathbb{E}(\boldsymbol{X}\hat{\boldsymbol{\beta}}) = \boldsymbol{X}\boldsymbol{\beta} + \mathbf{0}_{n\times 1} - \boldsymbol{X}\boldsymbol{\beta} = \mathbf{0}_{n\times 1},$$

where $\mathbf{0}_{l\times m}$ is the $l \times m$ matrix of zeros. And we can show that these two orthogonal random vectors $\hat{\boldsymbol{Y}}$ and $\boldsymbol{e}$ are uncorrelated:

$$\begin{aligned}
\mathbf{cov}(\hat{\boldsymbol{Y}}, \boldsymbol{e}) &= \mathbf{cov}\{\boldsymbol{H}\boldsymbol{Y}, (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}\} \\
&= \mathbb{E}[\boldsymbol{H}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{I} - \boldsymbol{H})'] = \boldsymbol{H}\mathbf{cov}(\boldsymbol{Y})(\boldsymbol{I} - \boldsymbol{H}) \\
&= \boldsymbol{H}(\sigma^2\boldsymbol{I})(\boldsymbol{I} - \boldsymbol{H}) = \sigma^2\boldsymbol{H}(\boldsymbol{I} - \boldsymbol{H}) = \sigma^2(\boldsymbol{H} - \boldsymbol{H}) = \mathbf{0}_{n\times n}\,.
\end{aligned}$$

The expression given in equation (5) allow us to derive the statistical properties of LSE. The mean and the covariance matrix of $\hat{\boldsymbol{\beta}}$ are

$$\begin{aligned}
\mathbb{E}(\hat{\boldsymbol{\beta}}) &= \mathbb{E}\{(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}\} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\mathbb{E}(\boldsymbol{Y}) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{\beta} \\
\mathbf{cov}(\hat{\boldsymbol{\beta}}) &= \mathbf{cov}\{(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}\} =: \mathbf{cov}(\boldsymbol{A}\boldsymbol{Y}) = \boldsymbol{A}\,\mathbf{cov}(\boldsymbol{Y})\,\boldsymbol{A}' \\
&= \boldsymbol{A}\,(\sigma^2\boldsymbol{I})\,\boldsymbol{A}' = \sigma^2\boldsymbol{A}\boldsymbol{A}' = \sigma^2\{(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\}\{\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1'}\} \\
&= \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{X})(\boldsymbol{X}'\boldsymbol{X})^{-1} = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1},
\end{aligned}$$

where for ease of presentation we wrote $\boldsymbol{A} := (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$.

Finally, to work out the distribution of the estimator $\hat{\boldsymbol{\beta}}$, we just have to note that $\hat{\boldsymbol{\beta}} = \boldsymbol{A}\boldsymbol{Y}$ (where $\boldsymbol{A} := (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$) is random because $\boldsymbol{Y}$ is random, and $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ is random because $\boldsymbol{\varepsilon}$ is random. Thus, to know the distribution of $\hat{\boldsymbol{\beta}}$, we have to know the distribution of $\boldsymbol{\varepsilon}$. If $\boldsymbol{\varepsilon} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I})$, then $\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I})$ and consequently, we can see

$$\hat{\boldsymbol{\beta}} = \boldsymbol{A}\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{A}\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{A}\boldsymbol{A}') = \mathcal{N}(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}).$$

Let me put the formula and its distribution together in one line for your quick reference:

$$\boxed{\boxed{\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1})}}.$$

Consequently, because $\hat{\boldsymbol{Y}}$ is simply the product of a constant matrix and $\hat{\boldsymbol{\beta}}$, it also has the normal distribution:

$$\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} \sim \mathcal{N}(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}').$$

For $\hat{\boldsymbol{Y}}_0$ at some $\boldsymbol{x}_0$ for a given fixed vector $\boldsymbol{x}_0$ containing specified values of the independent variables

$$\boldsymbol{x}_0 = (1, x_{10}, x_{20}, \ldots, x_{k0})',$$

using the same argument, we have

$$\hat{Y}_0 = \boldsymbol{x}_0'\hat{\boldsymbol{\beta}} \sim \mathcal{N}(\boldsymbol{x}_0'\boldsymbol{\beta}, \sigma^2\boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_0).$$

Mathematically, we can see that if $\boldsymbol{X}'\boldsymbol{X}$ is singular, then $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$ is not well defined, because $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ does not exist. The implication of the singularity of $\boldsymbol{X}'\boldsymbol{X}$ is that we cannot estimate $\boldsymbol{\beta}$ uniquely.

In the simple linear regression, if all observed x_i are the same, there will be infinitely many regression lines that pass through $(\overline{x}, \overline{Y})$.

In the multiple regression, if the matrix $\boldsymbol{X}$ is such that any of its columns can be explained as a linear combination of some other columns, this dependency will be transferred to $\boldsymbol{X}'\boldsymbol{X}$ and so $\boldsymbol{X}'\boldsymbol{X}$ will have a zero determinant and be singular.

The linear dependence arises often because the data are inadequate for fitting the model or, what is the same thing, the model is too complex for the available data. We need either more data or a simpler model for the available data.

ANOVA F-Test

The ANOVA F-test is a global test, testing the null hypothesis that $\beta_1 = \cdots = \beta_k = 0$ (note that the intercept β_0 is NOT included in the null hypothesis).

ANOVA table

Source of variation	Sum of Squares	degrees of freedom	Mean square	F-ratio	p-value
Model	SSR	k	$MSR = \frac{SSR}{k}$	$\frac{MSR}{MSE}$	$\Pr(F_{k,n-k-1} \geq F\text{-ratio})$
Error	SSE	$n-k-1$	$MSE = \frac{SSE}{n-k-1}$		
Total	SST	$n-1$			

Here

$$
\begin{aligned}
SST &= \sum (Y_i - \overline{Y})^2 = \sum Y_i^2 - \frac{1}{n}(\sum Y_i)^2 = \boldsymbol{Y}'\Big[\boldsymbol{I} - \frac{1}{n}\boldsymbol{1}\boldsymbol{1}'\Big]\boldsymbol{Y}\,, \\
SSE &= \sum (Y_i - \hat{Y}_i)^2 = \boldsymbol{e}'\boldsymbol{e} = (\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})'(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) \\
&= \boldsymbol{Y}'\boldsymbol{Y} - \hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{Y}\,, \\
SSR &= \sum (\hat{Y}_i - \overline{Y})^2 = (\boldsymbol{H}\boldsymbol{Y} - \frac{1}{n}\boldsymbol{1}\boldsymbol{1}'\boldsymbol{Y})'(\boldsymbol{H}\boldsymbol{Y} - \frac{1}{n}\boldsymbol{1}\boldsymbol{1}'\boldsymbol{Y}) \\
&= \boldsymbol{Y}'\boldsymbol{H}\boldsymbol{H}\boldsymbol{Y} - \frac{1}{n}\boldsymbol{Y}'\boldsymbol{H}\boldsymbol{1}\boldsymbol{1}'\boldsymbol{Y} - \frac{1}{n}\boldsymbol{Y}'\boldsymbol{1}\boldsymbol{1}'\boldsymbol{H}\boldsymbol{Y} + \frac{1}{n^2}\boldsymbol{Y}'\boldsymbol{1}\boldsymbol{1}'\boldsymbol{1}\boldsymbol{1}'\boldsymbol{Y} \\
&= \boldsymbol{Y}'\boldsymbol{H}\boldsymbol{Y} - \frac{1}{n}\boldsymbol{Y}'\boldsymbol{1}\boldsymbol{1}'\boldsymbol{Y} - \frac{1}{n}\boldsymbol{Y}'\boldsymbol{1}\boldsymbol{1}'\boldsymbol{Y} + \frac{1}{n}\boldsymbol{Y}'\boldsymbol{1}\boldsymbol{1}'\boldsymbol{Y} \\
&= \hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{Y} - \frac{1}{n}\boldsymbol{Y}'\boldsymbol{1}\boldsymbol{1}'\boldsymbol{Y} = SST - SSE,
\end{aligned}
$$

in which $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ and $\boldsymbol{1} = \boldsymbol{1}_{n\times 1} = [1, \ldots, 1]'$.

T-test

For each ℓ, where $0 \leq \ell \leq k$, to test the null hypothesis that the individual parameter $\beta_\ell = 0$, we use t-test:

$$T = \frac{\hat{\beta}_\ell}{\hat{\sigma}\sqrt{c_{\ell+1,\ell+1}}} \sim t_{n-k-1} \text{ when } \beta_\ell = 0,$$

where $c_{\ell+1,\ell+1}$ is the $(\ell+1)^{\text{st}}$ diagonal element of $(\boldsymbol{X}'\boldsymbol{X})^{-1}$. The $\sigma^2 c_{\ell+1,\ell+1}$ is the variance of $\hat{\beta}_\ell$, and the variance σ^2 of ε, when unknown, is estimated by sample variance of e_i:

$$\begin{aligned} \hat{\sigma}^2 &= \frac{\sum(Y_i - \hat{Y}_i)^2}{n-k-1} = \frac{(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})'(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})}{n-k-1} \\ &= \frac{\boldsymbol{e}'\boldsymbol{e}}{n-k-1} = \frac{SSE}{n-k-1} = MSE. \end{aligned}$$

(Recall that the value of the degrees of freedom of the t-distribution in the t-test is the same as the degrees of freedom of $(n-k-1)\hat{\sigma}^2/\sigma^2 \sim \chi^2_{n-k-1}$. It is the same as the dimension of the error space, in which the residual vector $\boldsymbol{e}$ is lying.)

Partial F-test

The ANOVA F-test tests whether all parameters except the intercept are zero, and the t-test tests whether an individual parameter is zero. Can we test whether some (more than one, but less than k) parameters are zero?

If we want to test whether a subset of parameters are all zero, then we are in fact comparing two models.

Two models are *nested* if one model contains all the terms of the second model and at least one additional term. *Reduced* (or *restricted*) model is a special case of (nested within) the *complete* (or *full*) model.

Suppose we suspect some independent variables are insignificant and we arrange the order of $x_1, \ldots, x_k$ so that the "suspicious" variables are labelled as the $(g+1)^{\text{st}}, \ldots, k^{\text{th}}$ independent variables.

We separate the independent variables into two groups and write:

$$\boldsymbol{Y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon},$$

where $\boldsymbol{\beta}_1 = [\beta_0, \beta_1, \ldots, \beta_g]'$ and $\boldsymbol{\beta}_2 = [\beta_{g+1}, \beta_{g+2}, \ldots, \beta_k]'$, where β_0 still denotes the intercept term. (Why β_0 must belong to the first group? Can we move β_0 to the second group?)

To test the null hypothesis that $\boldsymbol{\beta}_2 = \mathbf{0}$ against the alternative $\boldsymbol{\beta}_2 \neq \mathbf{0}$, we use the **partial F-test**:

$$\begin{aligned} F &= \frac{\text{extra SS}/\#\text{paramters being tested}}{MSE_{\text{complete}}} \\ &= \frac{(SSE_{\text{reduced}} - SSE_{\text{complete}})/(k-g)}{SSE_{\text{complete}}/(n-k-1)} \sim F_{k-g,n-k-1}, \end{aligned}$$

where the extra SS is the extra sum of squares of errors (extra unexplainable variation in Y_i) that we will have if we use the reduced model. The idea is that if using the reduced model, we will of course get a larger SSE (i.e. a smaller SSR [a smaller explainable variation]). But if the increase in SSE (i.e. the loss in SSR [loosely speaking, the loss in the information by the reduced model]) is not too large, then we can use the reduced model. This is exactly the **principle of parsimony**, which requires that in situations where two competing models have essentially the same predictive power, we choose the more parsimonious of the two; a parsimonious model is a model with a small number of parameters.

In fact, this partial F-test can test not only the reduced model with fewer parameters (i.e. some parameters are zero) but also reduced models with constraints on the parameters, e.g. $\beta_2 = \beta_4 = 0.5$ or $\beta_3 + \beta_6 = 2.3$, etc., and in such cases the value of the numerator degrees of freedom F-statistic is the number of constraints being tested, i.e.

$$F = \frac{\text{extra SS}/\#\text{constraints being tested}}{MSE_{\text{complete}}}.$$

However, note that the partial F-test cannot be used to compare two different models in general; the extra SS is nonnegative here because we know that SSE_{reduced} cannot be smaller than SSE_{complete}, as the reduced model is only a special case of the complete model and of course cannot be better than the complete model.

Coefficient of multiple determination

The **coefficient of multiple determination**

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

should be interpreted as, the same as that for the simple linear regression, how many percent of the variation in Y can be explained by the multiple regression model.

However, a larger value of R^2 computed from the sample data does not necessarily mean that the model provides a better fit of all of the data points in the whole population. You will always obtain a perfect fit $R^2 = 1$ for a sample of n points if the model contains exactly n parameters.

Adjusted coefficient of multiple determination

Thus, we introduce the **adjusted coefficient of multiple determination**:

$$\begin{aligned} R_a^2 &:= 1 - \frac{n-1}{n-k-1}\frac{SSE}{SST} = 1 - \frac{n-1}{n-k-1}(1-R^2) \\ &\leq 1 - \frac{n-1}{n-1}(1-R^2) = R^2. \end{aligned}$$

The idea is that R_a^2 takes into account (adjusted for) both the sample size and the number of parameters such that a model of more parameter will have a heavier penalty so that R_a^2 cannot be forced to 1 by simply adding more and more parameters.

One obvious disadvantage of R_a^2 is that its numerical value does not have a nice and easy to understand interpretation, while R^2 does.

Both R^2 and R_a^2 may be misleading when there are repeat observations. For example, if we have 100 observations in 5 groups (observations in each group have the same set of values for the independent variables), each of 20 repeats.

- A 5-parameter model will provide a perfect fit to the 5 sample means of Y and may give a very large value of R^2 or R_a^2, especially if σ^2 is small compared with the spread of the 5 means.
- However, such a model, which passes through the 5 sample means of Y at the 5 observed values of x actually may be a very bad model if we could observe more distinct values of x.
- Even if we do not have repeat observations at the same value of x, we may still have 5 groups such that in each group the within-group variation is very small, whilst the between-group variation is large, leading again to a high R^2 or R_a^2.

Therefore, we must make a scatter plot of the data before we do any analysis.

Coefficient of partial determination

The **coefficient of partial determination** is defined as

$$R^2_{Y,B|A} = \frac{SSE_A - SSE_{A,B}}{SSE_A} = \frac{SSE_{\text{reduced}} - SSE_{\text{complete}}}{SSE_{\text{reduced}}},$$

where tells us the percentage of variation that cannot be explained in the reduced model but can be explained in the complete model.

For example, $R^2_{Y,x_2,x_3|x_1}$ will be the percentage of variation that cannot be explained in the reduced model $Y = \beta_0 + \beta_1 x_1 + \varepsilon$ but can be explained in the complete model $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$, i.e. it measures the contribution of x_2 and x_3, when added to the model already containing x_1, in terms of the model's explanatory power for the variation in Y.

Partial coefficient

Typically, we will consider the case that B contains only one independent variable, i.e. $R^2_{Y,x_i|A}$ (adding x_i to the model already containing the variables listed in A).

The square root of $R^2_{Y,x_i|A}$, with the sign (i.e. positive or negative) the same as the sign of $\hat{\beta}_i$ (the estimate of the coefficient of x_i in the model), is called the **partial correlation** between x_i and Y, given A.

Confidence Intervals

Next, consider confidence intervals. We know that

$$\hat{\boldsymbol{\beta}} \sim \mathcal{N}(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}) \text{ and } \hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} \sim \mathcal{N}(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}').$$

These normal distributions would allow us, if we knew σ^2, to construct confidence intervals for each individual β_j and for the estimation of the mean response and the prediction of an individual response.
In practice we would not know σ^2, and thus we have to use the sample variance to estimate it:

$$\hat{\sigma}^2 = \frac{\sum(Y_i - \hat{Y}_i)^2}{n-k-1} = \frac{(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})'(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})}{n-k-1} = \frac{SSE}{n-k-1} = MSE.$$

These upper and lower confidence limits are reported by SAS if we add the `clb clm cli` option after the `model` statement.

Note that if the model is underspecified (i.e. some true independent variables are missing), then $\hat{\sigma}^2$ is, on the average, an overestimate of σ^2, because the variation in the calculated SSE comes from not only the errors (of variance σ^2) but also from the variation due to the missing independent variables which have not been accounted for in the regression model.

Because

$$\hat{\boldsymbol{\beta}} \sim \mathcal{N}(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}),$$

for a given fixed vector $\boldsymbol{x}_0$ containing specified values of the independent variables

$$\boldsymbol{x}_0 = (1, x_{10}, x_{20}, \ldots, x_{k0})',$$

the point estimation of the mean response and the prediction of an individual response at $\boldsymbol{x} = \boldsymbol{x}_0$ are

$$\hat{Y}_0 = \boldsymbol{x}_0'\hat{\boldsymbol{\beta}} \sim \mathcal{N}(\boldsymbol{x}_0'\boldsymbol{\beta}, \sigma^2\boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_0).$$

and the unknown σ^2 again is estimated by $\hat{\sigma}^2 = MSE = SSE/(n-k-1)$. Thus, we have the following $100(1-\alpha)\%$ confidence intervals for estimation of the mean response:

$$\hat{Y}_0 \pm t_{n-k-1;\alpha/2}\,\hat{\sigma}\,\sqrt{\boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_0},$$

and for prediction of a future response:

$$\hat{Y}_0 \pm t_{n-k-1;\alpha/2}\,\hat{\sigma}\,\sqrt{1 + \boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_0}.$$

Confidence Regions

We may want to construct a joint confidence region for more than one parameters, i.e. a region which, with $100(1-\alpha)\%$ confidence, contains the true parameter vector.

Such a joint confidence region is a kind of *simultaneous inference* for several unknown parameters.

For the parameter vector $\boldsymbol{\beta}$ in the multiple regression model with normally distributed errors, because the LSE (also MLE) estimator $\hat{\boldsymbol{\beta}} \sim \mathcal{N}(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1})$, if normalising this general multivariate normal to the standard multivariate normal (by subtracting the mean and "dividing" by the standard deviation) and then taking square, we will have the χ^2-distribution:

$$\frac{1}{\sigma^2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\boldsymbol{X}'\boldsymbol{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \sim \chi^2_{k+1}.$$

Combining with the facts that SSE/σ^2 is also χ^2-distributed and that they are independent, we have

$$\frac{\left\{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\boldsymbol{X}'\boldsymbol{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\right\} \Big/ \sigma^2 \Big/ (k+1)}{\frac{SSE}{\sigma^2} \Big/ (n-k-1)} \sim F_{k+1,n-k-1},$$

where $SSE/(n-k-1) = MSE = \hat{\sigma}^2$.

Hence a $100(1-\alpha)\%$ *joint confidence region* for $\boldsymbol{\beta}$ can be obtained from solving

$$(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})'\boldsymbol{X}'\boldsymbol{X}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}) \leq (k+1)\hat{\sigma}^2 F_{k+1,n-k-1;\alpha}, \tag{7}$$

where $F_{k+1,n-k-1;\alpha}$ denotes the value such that with $100(1-\alpha)\%$ probability, a random variable following the $F_{k+1,n-k-1}$ distribution is not larger than $F_{k+1,n-k-1;\alpha}$.

The equality given in (7) is the equation of the boundary of an elliptically (or ellipsoidally, when $k>1$) shaped contour in $\mathbb{R}^{k+1}$, and the strict inequality is the interior of the ellipse (ellipsoid).

The joint confidence region takes into account the correlation between the estimators $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$.

The individual confidence intervals are only appropriate for specifying the ranges for the individual parameters irrespective of the value of the other parameters.

If we (inappropriately) interpret these intervals for individual parameters simultaneously, i.e. wrongly regard the (hyper-)rectangle that they define as a joint confidence region, then e.g. it may be thought that a point lying outside the ellipse but inside the rectangle provide a reasonable value for β, but it is in fact not reasonable because it is not in the 95% joint confidence region.

When only two parameters are involved, construction of the confidence ellipse is not difficult. In practice, however, even for two parameters, it is rarely drawn.

Bonferroni Correction

What is more popular in simultaneous inference is the *Bonferroni* correction (also known as the Bonferroni adjustment, the Bonferroni procedure, etc.). It is based on the Bonferroni inequality, which is described as follows.

Let A_i denote the event that the confidence interval I_i of β_i does not cover the true β_i and suppose the confidence level is $100(1-\gamma)\%$, i.e.

$$\Pr(\beta_i \notin I_i) = \Pr(A_i) = \gamma.$$

We might take the hyper-rectangle $I_0 \times I_1 \times \cdots \times I_k$ as a $100(1-\alpha)\%$ joint confidence region if the following were true:

$$\Pr(\boldsymbol{\beta} \in I_0 \times I_1 \times \cdots \times I_k) = \Pr(\overline{A}_0 \cap \overline{A}_1 \cap \cdots \cap \overline{A}_k) = 1-\alpha,$$

where $\overline{A}_i$ denotes the complement of A_i, i.e. $\overline{A}_i = \{\beta_i \in I_i\}$.

However, because these $\overline{A}_i$ are not independent and the dependence structure between $\overline{A}_i$ and $\overline{A}_j$ is complicated, the best we can do is to be conservative, i.e. choose a γ such that

$$\Pr(\overline{A}_0 \cap \overline{A}_1 \cap \cdots \cap \overline{A}_k) \geq 1 - \alpha. \tag{8}$$

Now,

$$\begin{aligned}
&\Pr(\overline{A}_0 \cap \overline{A}_1 \cap \cdots \cap \overline{A}_k) \\
&= 1 - \Pr(A_0 \cup A_1 \cup \cdots \cup A_k) \\
&= 1 - \{\Pr(A_0) + \cdots + \Pr(A_k) - \Pr(\text{parts counted more than once})\} \\
&= 1 - \{\Pr(A_0) + \cdots + \Pr(A_k)\} + \Pr(\text{parts counted more than once}) \\
&\geq 1 - \{\Pr(A_0) + \cdots + \Pr(A_k)\} = 1 - (k+1)\gamma.
\end{aligned}$$

Therefore, in order to achieve (8), we require that

$$\gamma = \frac{\alpha}{k+1} = \frac{\alpha}{\text{\# parameters}}.$$

That is to say, using the Bonferroni adjustment, the rectangular region formed by the Cartesian product of $k+1$ intervals, each of them is a $100(1-\frac{\alpha}{k+1})\%$ confidence interval of an individual parameter β_ℓ:

$$\hat{\beta}_\ell \pm t_{n-k-1;\frac{\alpha}{2(k+1)}} \cdot \hat{\sigma} \cdot \sqrt{c_{\ell+1,\ell+1}}, \qquad \ell = 0, 1, \ldots, k,$$

where $c_{\ell+1,\ell+1}$ is the $(\ell+1)^{\text{st}}$ diagonal element of $(\boldsymbol{X'X})^{-1}$, is a joint confidence region for the $k+1$ parameters in the vector $\boldsymbol{\beta}$, the confidence level of this rectangle is <u>at least</u> $100(1-\alpha)\%$.

This is more popular than the exact $100(1-\alpha)\%$ confidence ellipsoid because a hyper-rectangle is easier for understanding and interpretation for individual parameters than an ellipsoid.

Of course the Bonferroni inequality can also be applied to multiple testing problems so that if we want to perform, say, individual t-test for each of m parameters, in order to have the overall significance level not higher than α, each t-test should use α/m as its significance level.

Note that whilst using Bonferroni adjustment we can control the overall significance of multiple t-tests to be bounded above by α, but using ANOVA we can control the significance level of the F-test at exactly α.

Higher-order models

Suppose we have two independent variables x_1 and x_2. The model containing only the *first-order terms* is

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon.$$

Under this model, for each fixed x_2, the relationship between x_1 and Y is a straight line with slope β_1 and intercept $\beta_0 + \beta_2 x_2$.

That is to say, for different values of x_2, we get a set of parallel lines with different intercepts; the value of x_2 does not affect the relationship (i.e. the slope) between x_1 and Y. In such a situation we say that x_1 and x_2 have no *interaction*.

However, in real applications, it is common that the independent variables interact, i.e. the relationship between x_1 and Y depends on the value of x_2.

Imagine a situation in which for $x_2 = 10$, the straight line relationship between x_1 and Y have a positive slope, whilst that for $x_2 = 50$, x_1 and Y have a negative slope; we have nonparallel lines.

In such a situation, we say x_1 and x_2 have interaction and the model that allows interaction between two independent variables will include the cross-product term of these two independent variables.

In particular, if we have only two independent variables, then the model

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon \tag{9}$$

is an interaction model, where $\beta_1 + \beta_3 x_2$ is the change in $\mathbb{E}(Y)$ for every 1-unit increase in x_1, holding x_2 fixed, and $\beta_2 + \beta_3 x_1$ is the change in $\mathbb{E}(Y)$ for every 1-unit increase in x_2, holding x_1 fixed.

The **interaction term** $x_1 x_2$ is a **second-order** term, whilst x_1 and x_2, called the **main effects**, are **first-order**.

When we use an interaction model, extreme care is needed in interpreting the signs and sizes of coefficients. A negative β_1 does not necessarily mean that an increase in x_1 is accompanied by a decrease in the mean of Y.

The most important parameter is the one associated with the highest order term in the model. Once an interaction is significant, do NOT test on the first-order terms of the interacting variables.

In models with more than two independent variables, there may be several higher order terms (e.g. third- or fourth-order interactions). The same principle applies to such cases: do NOT test on the lower order terms of the variables contained in the existing higher order terms in the model.

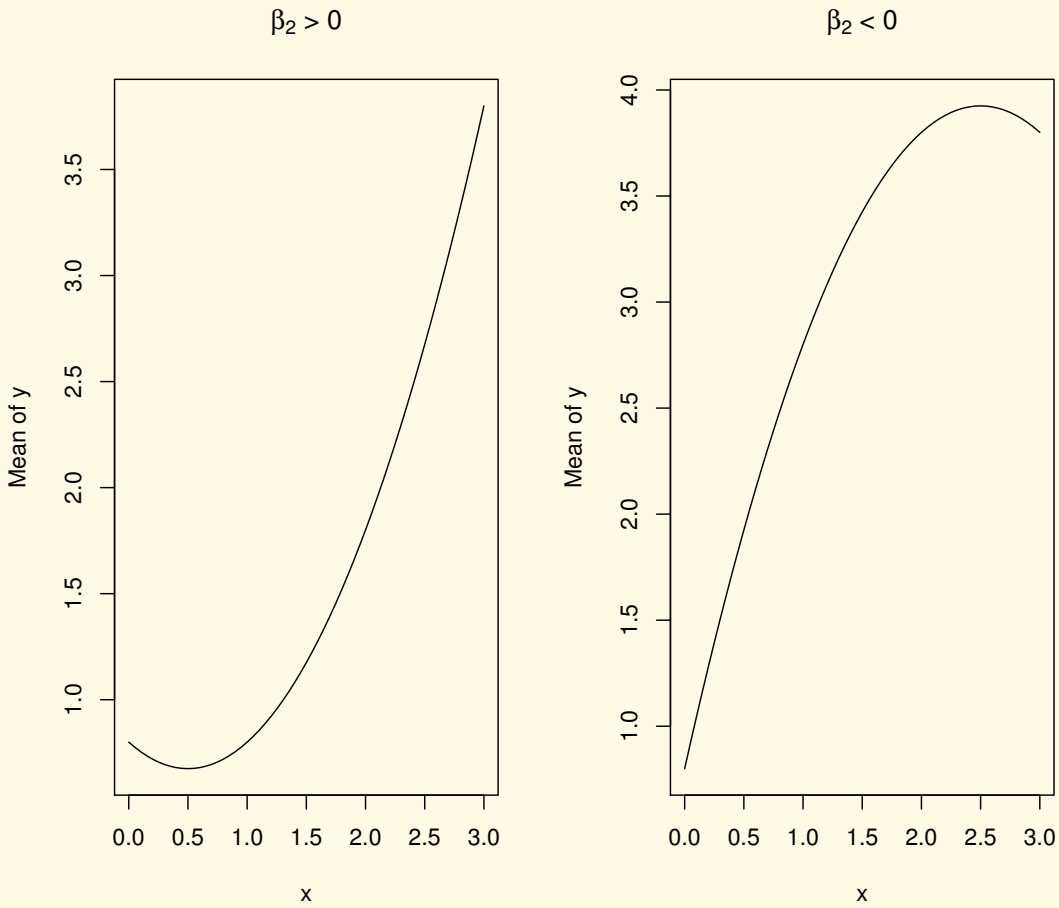
The presence of interaction implies that the lower order terms are important regardless of their p-values shown on the computer printout.

However, note that it is not a restriction required by mathematics; this is a principle that makes practical interpretation of the model easier and neater. The mathematics would not go wrong even if we excluded lower order terms but kept their higher order terms.

The model given in (9) is a second-order model. If we have only one independent variable, we still can form a second-order model, namely the **quadratic model**

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon,$$

where the parameter β_2 controls the rate of curvature so that if $\beta_2 > 0$, the curve is concave upward, whilst if $\beta_2 < 0$, the curve is concave downward.


$\beta_2 > 0$
Mean of y
1.0
1.5
2.0
2.5
3.0
3.5
0.0
0.5
1.0
1.5
2.0
2.5
3.0
x
$\beta_2 < 0$
Mean of y
1.0
1.5
2.0
2.5
3.0
3.5
4.0
0.0
0.5
1.0
1.5
2.0
2.5
3.0
x

In a quadratic model, model interpretations are not meaningful outside the range of the independent variables; in particular, $\hat{\beta}_0$ can be meaningfully interpreted only if the range of the predictor includes $x = 0$.

The parameter β_1, in general, will not have a meaningful interpretation at all, and so rather than interpreting the numerical value of the estimate $\hat{\beta}_1$ itself, we utilize a graphical representation of the model to describe the model, which can be done in the SAS procedure `gplot`:

```
proc gplot;
symbol v=plus i=rq;
plot y*x;
run;
```

in which `i=rq` means that the interpolation method is regression and the model is quadratic. An example of the output is given in Figure 3, which is the fitted model for Example 4.7.

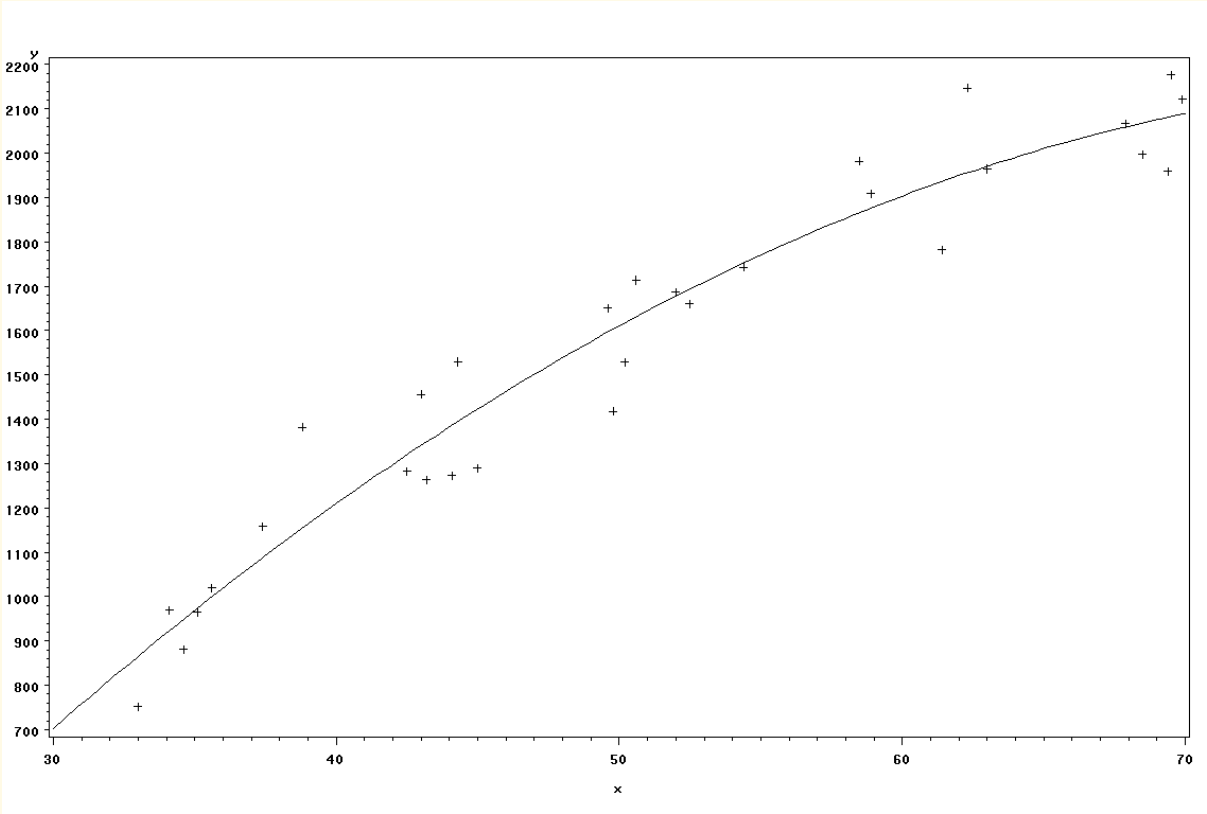


Figure 3: The scatterplot of the dataset aerobic in Example 4.7 on pages 203–207, with the fitted quadratic model.

If we have two independent variables, the *complete second-order model* is

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \beta_4 x_1^2 + \beta_5 x_2^2 + \varepsilon.$$

Since most relationships in the real world are curvilinear, at least to some extent, and so a good first choice would be the complete second-order model. [A curvilinear regression typically refers to a regression model containing quadratic, cubic, quartic, or higher order terms of one or more independent variables.]

If we have prior information that there is no curvature, then we may let $\beta_4 = \beta_5 = 0$, leading to the interaction model given in (9). If there is no interaction, then we may let $\beta_3 = 0$. If we do not have any prior information, we may still test whether a reduced model is as good as the complete model by the partial F-test. Which variables should be kept in the model and which should be removed from the model is the problem of *variable selection*, which will be discussed in details in Chapter 6.

Qualitative (categorical) independent variables

Up to here, we have considered only quantitative independent variables. Suppose we have qualitative (categorical) independent variables, then we have to introduce *dummy variables*. The different values of a qualitative independent variable are called its *levels*. For a two-level independent variable (say, level A and level B), we need one dummy variable:

$$x = \begin{cases} 1, & \text{if level A,} \\ 0, & \text{if level B,} \end{cases}$$

and write

$$Y = \beta_0 + \beta_1 x + \varepsilon,$$

so that

at level A: mean of the response is $\beta_0 + \beta_1$,

at level B: mean of the response is β_0.

In general, if we have an m-level qualitative independent variable, we need $m-1$ dummy variables

$$x_i = \begin{cases} 1, & \text{for the } i^{\text{th}} \text{ level,} \\ 0, & \text{otherwise,} \end{cases}$$

for $i = 1, \ldots, m-1$, and write

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_{m-1} x_{m-1} + \varepsilon,$$

so that the mean responses are

at the i^{th} level: mean of the response is $\beta_0 + \beta_i$, $i = 1, \ldots, m-1$,

at the m^{th} level: mean of the response is β_0.

But, wait, why don't we define a variable, say,

$$x = \begin{cases} 0, & \text{for the first level,} \\ 1, & \text{for the second level,} \\ 2, & \text{for the third level,} \\ \vdots & \vdots \\ m-1 & \text{for the } m\text{th level,} \end{cases}$$

instead of getting into the trouble of introducing $m-1$ dummy variables?

The answer is that if we do so, then we actually assume that changing x from level 1 to level 2, or changing x from level 4 to level 5, the change in the mean of Y is the same; this is because the change in the value of x is 1 in either case.

In other words, at each level of this categorical variable we get, say, a straight line and if we have no interaction, then such an artificial x will give us parallel lines that are equally apart, i.e. the vertical distance between two consecutive lines is always the same value (equal to the coefficient in front of x). This is not desirable unless there is a good reason to impose such a strong restriction by scientific theory.

To allow parallel lines that are not equally apart, each level should contribute a different β_i to the intercept, and this has to be done by using dummy variables. If using the artificial x, it means we require artificially $\beta_2 = 2\beta_1$ and $\beta_3 = \beta_2 + \beta_1 = 3\beta_1$, and so on.

Let us consider $m-1$ dummy variables $x_1, \ldots, x_{m-1}$ for an m-level categorical independent variable, i.e.

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_{m-1} x_{m-1} + \varepsilon. \tag{10}$$

Testing the null hypothesis that $\beta_1 = \cdots = \beta_{m-1} = 0$ by the F-test in the ANOVA table of the Regression Analysis studied in this course is equivalent to testing the null hypothesis that all m means are the same $\mu_1 = \cdots = \mu_m = 0$, where μ_i is the mean response at level i. The latter is the ANOVA that we have learnt in MATH2206 Prob. Stat. These two ANOVA are of course the same.

A neater expression for the model in (10) is:

$$Y_i = \beta_0 + \beta_i + \varepsilon, \qquad i = 1, \ldots, m,$$

where Y_i denotes the response at level i and $\beta_m = 0$ (so that the mean response at level m is just the intercept term β_0), or equivalently,

$$Y_i = \mu + \alpha_i + \varepsilon, \qquad i = 1, \ldots, m,$$

where μ is the overall mean and $\sum \alpha_i = 0$. The parameter α_i is interpreted as the effect of the level i on the mean, and so the mean response at level i will be the overall mean μ plus the effect α_i (some are positive and some are negative), i.e.

$$\mu_i = \mu + \alpha_i,$$

which immediately leads to the constraint we wrote above:

$$\mu = \frac{\sum \mu_i}{m} = \frac{\sum(\mu + \alpha_i)}{m} = \mu + \frac{\sum \alpha_i}{m} \Rightarrow \sum \alpha_i = 0.$$

This formulation is commonly used for ANOVA, and this will be the model for the simplest case in experimental design (MATH3815).