MA 575 Linear Models:

Cedric E. Ginestet, Boston University

Multiple Linear Regression

Week 4, Lecture 2

1 Multiple Regression

1.1 The Data

The simple linear regression setting can be extended to the case of p independent variables, such that we may now have the following array of data points,

$$(y_i, x_{i1}, \ldots, x_{ip}), \qquad i = 1, \ldots, n.$$

In addition, and for notational convenience, we also include a dummy variable, which will be used to compute the y-intercept, β_0. Therefore, when the model includes such an intercept, we add the dummy variable $x_{i0} := 1$, for every i, and obtain the full data set,

$$(y_i, x_{i0}, x_{i1}, \ldots, x_{ip}), \qquad i = 1, \ldots, n.$$

Therefore,

1. the **number of covariates** will be denoted by p,
2. whereas the **number of parameters** will be denoted by $p^* := p + 1$.

Naturally, the IVs, x_{ij}, can take either *discrete* or *continuous* values, without affecting the estimation procedure. In the sequel, we will systematically assume that the model contains an intercept, except when specified otherwise. In multiple regression, it is assumed that $p < n$, and more generally, we mainly consider data sets, for which $p \ll n$.

1.2 The Model

Multiple linear regression (MLR) is defined in the following manner,

$$y_i = \sum_{j=0}^{p} x_{ij}\beta_j + e_i, \qquad \forall\ i = 1, \ldots, n,$$

which may then be reformulated, using linear algebra,

$$y_i = \mathbf{x}_i^T \boldsymbol{\beta} + e_i, \qquad \forall\ i = 1, \ldots, n,$$

where $\boldsymbol{\beta}$ and the $\mathbf{x}_i$'s are $(p^* \times 1)$ column vectors. Altogether, we can write the entire system of linear equations in matrix format,

$$\begin{vmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{vmatrix} = \begin{vmatrix} 1 & x_{11} & \dots & x_{1p} \\ 1 & x_{21} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \dots & x_{np} \end{vmatrix} \begin{vmatrix} \beta_0 \\ \vdots \\ \beta_p \end{vmatrix} + \begin{vmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{vmatrix}.$$

Alternatively, one could also re-express this as a single equation, alongside the assumption on the error terms,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e},$$

where $\mathbf{y}$ and $\mathbf{e}$ are $(n \times 1)$ vectors, $\mathbf{X}$ is an $(n \times p^*)$ matrix, and $\boldsymbol{\beta}$ is a $(p^* \times 1)$ vector.

1.3 Example: One-way ANOVA

The matrix $\mathbf{X}$ is usually referred to as the **design matrix**, because it specifies the experimental design. For instance, if considering a one-way analysis of variance (ANOVA) over three different groups, where we have 2 subjects in each group. We may select one of the following two design matrices,

$$\mathbf{X}_1 := \begin{vmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{vmatrix}, \qquad \text{and} \qquad \mathbf{X}_2 := \begin{vmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{vmatrix};$$

where in the former case, $\mathbf{X}_1$ is called a **cell means design**, whereas in the latter case, $\mathbf{X}_2$ is referred to as a **reference group design**, where the mean value in the two remaining groups are expressed as *offsets* from the value attained in the first group.

1.4 Geometrical Perspective

Just as the **mean function** of a simple regression determines a one-dimensional **line** in the two-dimensional Euclidean space, $\mathbb{R}^2$, the defining equation of multiple regression determines a **p-dimensional hyperplane** embedded into the p^*-dimensonal Euclidean space, $\mathbb{R}^{p^*}$.

The goal of OLS estimation is to identify the optimal hyperplane **minimizing** our target statistical criterion with respect to all the points in the sample, i.e. the n points of the form $(y_i, x_{i1}, \dots, x_{ip})$, positioned in p^*-dimensional Euclidean space, $\mathbb{R}^{p^*}$.

1.5 Model Assumptions

Multiple regression is based on the following assumptions:

1. **Linearity** in the parameters, such that the *mean function* is defined as

$$\mathbb{E}[Y_i|\mathbf{X}] = \mathbf{x}_i^T\boldsymbol{\beta}, \qquad \forall\ i = 1, \dots, n.$$

2. **Independence and homoscedasticity** of the error terms, such that

$$\mathbb{V}\text{ar}[\mathbf{e}|\mathbf{X}] = \sigma^2\mathbf{I}_n.$$

Equivalently, the *variance function* may be assumed to satisfy,

$$\mathbb{V}\text{ar}[\mathbf{y}|\mathbf{X}] = \sigma^2\mathbf{I}_n,$$

since the variance operator is *invariant* under translation, i.e. $\mathbb{V}\text{ar}[\mathbf{a}+\mathbf{Y}] = \mathbb{V}\text{ar}[\mathbf{Y}]$.

3. In addition to the standard OLS assumptions for simple linear regression, we will also assume that $\mathbf{X}$ has **full rank**. That is,
$$\text{rank}(\mathbf{X}) = p^*.$$

2 Minimizing the Residual Sum of Squares

2.1 Matrix Formulation for RSS

Since $p^* < n$, we have here a *non-homogeneous over-determined* system of linear equations, in the parameters β_j's. So, as before, we define, for computational convenience, a statistical criterion, which we wish to minimize. The RSS for this model is given by

$$\text{RSS}(\boldsymbol{\beta}) := \sum_{i=1}^{n} \left(y_i - \mathbf{x}_i^T\boldsymbol{\beta}\right)^2,$$

which can be re-expressed more concisely as follows,

$$\text{RSS}(\boldsymbol{\beta}) := (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}).$$

Observe that this criterion is a **scalar**, and not a vector or a matrix. This is a dot product. In particular, this should be contrasted with the variance of a vector, such as for instance,

$$\mathbb{V}\text{ar}[\mathbf{y}|\mathbf{X}] := \mathbb{E}\Big[(\mathbf{y} - \mathbb{E}[\mathbf{y}|\mathbf{X}])(\mathbf{y} - \mathbb{E}[\mathbf{y}|\mathbf{X}])^T\Big],$$

which is an $n \times n$ matrix.

2.2 Some Notions of Matrix Calculus

Consider a vector-valued function $\mathbf{F}(\mathbf{x}) : \mathbb{R}^n \mapsto \mathbb{R}^m$ of order $m \times 1$, such that

$$\mathbf{F}(\mathbf{x}) = \Big[f_1(\mathbf{x}), \ldots, f_m(\mathbf{x})\Big]^T.$$

The differentiation of such a vector-valued function $\mathbf{F}(\mathbf{x})$ by another vector $\mathbf{x}$ of order $n \times 1$ is **ambiguous** in the sense that the derivative

$$\frac{\partial \mathbf{F}(\mathbf{x})}{\partial \mathbf{x}}$$

can either be expressed as an $m \times n$ matrix (**numerator** layout convention), or as an $n \times m$ matrix (**denominator** layout convention), such that we have

$$\frac{\partial \mathbf{F}(\mathbf{x})}{\partial \mathbf{x}} = \begin{vmatrix} \frac{\partial f_1(\mathbf{x})}{\partial x_1} & \cdots & \frac{\partial f_1(\mathbf{x})}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m(\mathbf{x})}{\partial x_1} & \cdots & \frac{\partial f_m(\mathbf{x})}{\partial x_n} \end{vmatrix}, \qquad \text{or} \qquad \frac{\partial \mathbf{F}(\mathbf{x})}{\partial \mathbf{x}} = \begin{vmatrix} \frac{\partial f_1(\mathbf{x})}{\partial x_1} & \cdots & \frac{\partial f_m(\mathbf{x})}{\partial x_1} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_1(\mathbf{x})}{\partial x_n} & \cdots & \frac{\partial f_m(\mathbf{x})}{\partial x_n} \end{vmatrix}.$$

In the sequel, we will adopt the **denominator** layout convention, such that the resulting vector of partial derivatives will be of the dimension of the vector that we are conducting the differentiation with respect to. Moreover, we will be mainly concerned with differentiating scalars. However, observe that the choice of layout convention remains important, even though we are only considering the case of scalar-valued functions. Indeed, if we were to adopt the numerator convention, the derivative $\frac{\partial}{\partial \mathbf{x}} f(\mathbf{x})$ would produce a **row vector**, whereas the denominator convention would yield a **column vector**. In a sense, the choice of a particular layout convention is equivalent to the question of treating all vectors as either row or column vectors in linear algebra. Here, for consistency, all vectors are treated as column vectors, and therefore we also select the denominator layout convention.

For instance, given any *column vectors* $\mathbf{a}, \mathbf{x} \in \mathbb{R}^n$, the scalar-valued function $f(\mathbf{x}) := \mathbf{a}^T\mathbf{x}$ is differentiated as follows,

$$\frac{\partial}{\partial \mathbf{x}}(\mathbf{a}^T\mathbf{x}) = \begin{vmatrix} \frac{\partial \mathbf{a}^T\mathbf{x}}{\partial x_1} \\ \vdots \\ \frac{\partial \mathbf{a}^T\mathbf{x}}{\partial x_n} \end{vmatrix} = \begin{vmatrix} \frac{\partial}{\partial x_1} \sum a_i x_i \\ \vdots \\ \frac{\partial}{\partial x_n} \sum a_i x_i \end{vmatrix} = \mathbf{a}.$$

Moreover, it immediately follows that differentiating a scalar is invariant to **transposition**,

$$\frac{\partial}{\partial \mathbf{x}}(\mathbf{a}^T\mathbf{x}) = \frac{\partial}{\partial \mathbf{x}}(\mathbf{a}^T\mathbf{x})^T = \frac{\partial}{\partial \mathbf{x}}(\mathbf{x}^T\mathbf{a}) = \mathbf{a}. \tag{1}$$

For a **quadratic form**, however, things become slightly more cumbersome. Here, we are considering the function of a column vector $\mathbf{x} \in \mathbb{R}^n$, for some square matrix $\mathbf{A}$ of order $n \times n$,

$$\mathbf{x}^T\mathbf{A}\mathbf{x} = \begin{vmatrix} x_1 & \dots & x_n \end{vmatrix} \begin{vmatrix} a_{11} & \dots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \dots & a_{nn} \end{vmatrix} \begin{vmatrix} x_1 \\ \vdots \\ x_n \end{vmatrix}.$$

Naturally, this quantity is also a **scalar**. Differentiation with respect to $\mathbf{x}$, adopting the denominator convention gives

$$\frac{\partial}{\partial \mathbf{x}}(\mathbf{x}^T\mathbf{A}\mathbf{x}) = \begin{vmatrix} \frac{\partial}{\partial x_1} \sum_{i,j} a_{ij} x_i x_j \\ \vdots \\ \frac{\partial}{\partial x_n} \sum_{i,j} a_{ij} x_i x_j \end{vmatrix}, \tag{2}$$

where observe that the double summation in each element of this vector can be simplified as follows,

$$\begin{aligned} \left[\frac{\partial}{\partial \mathbf{x}}(\mathbf{x}^T\mathbf{A}\mathbf{x})\right]_k &= \frac{\partial}{\partial x_k} \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} x_i x_j \\ &= \sum_{j=1}^{n} a_{kj} x_j + \sum_{j=1}^{n} a_{jk} x_j = \mathbf{A}_{k\cdot}\mathbf{x} + \mathbf{A}_{\cdot k}^T\mathbf{x}, \end{aligned}$$

for every $k = 1, \dots, n$, and where $\mathbf{A}_{k\cdot}$ and $\mathbf{A}_{\cdot k}$ denote the k^{th} row and the k^{th} column of $\mathbf{A}$, respectively. For the entire vector, the above expression therefore gives

$$\frac{\partial}{\partial \mathbf{x}}(\mathbf{x}^T\mathbf{A}\mathbf{x}) = \mathbf{A}\mathbf{x} + \mathbf{A}^T\mathbf{x} = \left(\mathbf{A} + \mathbf{A}^T\right)\mathbf{x},$$

where recall that $\mathbf{A}$ is a **square** matrix, which can hence be transposed. Now, if in addition, this matrix is also **symmetric**, such that $\mathbf{A} = \mathbf{A}^T$, then

$$\frac{\partial}{\partial \mathbf{x}}(\mathbf{x}^T\mathbf{A}\mathbf{x}) = 2\mathbf{A}\mathbf{x}, \tag{3}$$

which provides a natural matrix generalization of the classical *power rule* of differential calculus, $\frac{\partial}{\partial x}x^k = kx^{k-1}$, when $k = 2$. A useful mnemonic for recalling whether one should eliminate $\mathbf{x}^T$ or $\mathbf{x}$, is to remember that we must obtain a matrix, which is conformable to the order of the argument with respect to which we differentiate. In this case, this is $\boldsymbol{\beta}$, which is of order $(n \times 1)$, and therefore we know that we must obtain $\mathbf{Ax}$, which is also of order $(n \times 1)$.

2.3 Derivation of OLS Estimators

Now, the OLS estimators can be defined as the vector of β_j's that minimizes the RSS,

$$\widehat{\boldsymbol{\beta}} := \underset{\widetilde{\boldsymbol{\beta}} \in \mathbb{R}^{p^*}}{\operatorname{argmin}} \operatorname{RSS}(\widetilde{\boldsymbol{\beta}}).$$

This can be expanded in the following manner,

$$\begin{aligned}
\operatorname{RSS}(\boldsymbol{\beta}) &:= (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \\
&= \mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbf{X}\boldsymbol{\beta} - \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{y} + \boldsymbol{\beta}^T(\mathbf{X}^T\mathbf{X})\boldsymbol{\beta} \\
&= \mathbf{y}^T\mathbf{y} - 2\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{y} + \boldsymbol{\beta}^T(\mathbf{X}^T\mathbf{X})\boldsymbol{\beta},
\end{aligned}$$

where we have used the fact that any scalar is **invariant** under **transposition**, such that

$$\mathbf{y}^T\mathbf{X}\boldsymbol{\beta} = (\mathbf{y}^T\mathbf{X}\boldsymbol{\beta})^T = \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{y}.$$

Differentiating and setting to 0, we obtain

$$\frac{\partial}{\partial \boldsymbol{\beta}}\boldsymbol{\beta}^T(\mathbf{X}^T\mathbf{X})\boldsymbol{\beta} - 2\frac{\partial}{\partial \boldsymbol{\beta}}\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{y} = 0,$$

where using equations (1) and (3), we obtain

$$2(\mathbf{X}^T\mathbf{X})\boldsymbol{\beta} = 2\mathbf{X}^T\mathbf{y}$$

since $\mathbf{X}^T\mathbf{X}$ can be shown to be **symmetric**. This produces a system of linear equations, which are referred to as the **normal equations**, in statistics. These equations put p^* constraints on the random vector, $\mathbf{y}$, of observed values.

Finally, we have also assumed that $\mathbf{X}$ is **full rank**. Moreover, since $\mathbf{X}$ is a matrix with **real entries**, it is known that the rank of $\mathbf{X}$ is equal to the rank of its **Gram** matrix, defined as $\mathbf{X}^T\mathbf{X}$, such that

$$\operatorname{rank}(\mathbf{X}) = \operatorname{rank}(\mathbf{X}^T\mathbf{X}) = p^*.$$

Since $\mathbf{X}^T\mathbf{X}$ is a matrix of order $p^* \times p^*$, it follows that this matrix is therefore also of full rank, which is equivalent to that matrix being **invertible**. Thus, the minimizer of $\operatorname{RSS}(\boldsymbol{\beta})$ is given by

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