1 Conventions for Scalar Functions

Let $\beta = (\beta_1, ..., \beta_k)'$ be a $k \times 1$ vector and let $f(\beta) = f(\beta_1, ..., \beta_k)$ be a real-valued function that depends on $\beta$, i.e. $f(\cdot): \mathbb{R}^k \rightarrow \mathbb{R}$ maps the vector $\beta$ into a single number, $f(\beta)$. Then the derivative of $f(\cdot)$ with respect to $\beta$ is defined as

$$
\frac{\partial f(\beta)}{\partial \beta} = \begin{pmatrix}
\frac{\partial f(\beta)}{\partial \beta_1} \\
\vdots \\
\frac{\partial f(\beta)}{\partial \beta_k}
\end{pmatrix}.
$$

This is a $k \times 1$ column vector with typical elements given by the partial derivative $\frac{\partial f(\beta)}{\partial \beta_i}$. Sometimes this vector is referred to as the gradient. It is useful to remember that the derivative of a scalar function with respect to a column vector gives a column vector as the result.$^1$

---

$^1$We can note that Wooldridge (2003, p.783) does not follow this convention, and let $\frac{\partial f(\beta)}{\partial \beta}$ be a $1 \times k$ row vector.
Similarly, the derivative of a scalar function with respect to a row vector yields the \(1 \times k\) row vector
\[
\frac{\partial f(\beta)}{\partial \beta} = \left( \frac{\partial f(\beta)}{\partial \beta_1} \cdots \frac{\partial f(\beta)}{\partial \beta_k} \right).
\]

2 Conventions for Vector Functions

Now let
\[g(\beta) = \left( \begin{array}{c} g_1(\beta) \\ \vdots \\ g_n(\beta) \end{array} \right)\]
be a vector function depending on \(\beta = (\beta_1, \ldots, \beta_k)^T\), i.e. \(g(\cdot) : \mathbb{R}^k \rightarrow \mathbb{R}^n\) maps the \(k \times 1\) vector into a \(n \times 1\) vector, where \(g_i(\beta) = g_i(\beta_1, \ldots, \beta_k)\), \(i = 1, 2, \ldots, n\), is a real-valued function.

Since \(g(\cdot)\) is a column vector it is natural to consider the derivatives with respect to a row vector, \(\beta^T\), i.e.
\[
\frac{\partial g(\beta)}{\partial \beta^T} = \left( \begin{array}{ccc} \frac{\partial g_1(\beta)}{\partial \beta_1} & \cdots & \frac{\partial g_1(\beta)}{\partial \beta_k} \\ \vdots & \ddots & \vdots \\ \frac{\partial g_n(\beta)}{\partial \beta_1} & \cdots & \frac{\partial g_n(\beta)}{\partial \beta_k} \end{array} \right),
\]
where each row, \(i = 1, 2, \ldots, n\), contains the derivative of the scalar function \(g_i(\cdot)\) with respect to the elements in \(\beta\). The result is therefore a \(n \times k\) matrix of derivatives with typical element \((i, j)\) given by \(\frac{\partial g_i(\beta)}{\partial \beta_j}\). If the vector function is defined as a row vector, it is natural to take the derivative with respect to the column vector, \(\beta\).

We can note that it holds in general that
\[
\frac{\partial (g(\beta)^T)}{\partial \beta} = \left( \frac{\partial g(\beta)}{\partial \beta^T} \right)^T,
\]
which in the case above is a \(k \times n\) matrix.

Applying the conventions in (1) and (2) we can define the Hessian matrix of second derivatives of a scalar function \(f(\beta)\) as
\[
\frac{\partial^2 f(\beta)}{\partial \beta \partial \beta^T} = \frac{\partial^2 f(\beta)}{\partial \beta^T \partial \beta} = \left( \begin{array}{ccc} \frac{\partial^2 f(\beta)}{\partial \beta_1 \partial \beta_1} & \cdots & \frac{\partial^2 f(\beta)}{\partial \beta_1 \partial \beta_k} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f(\beta)}{\partial \beta_k \partial \beta_1} & \cdots & \frac{\partial^2 f(\beta)}{\partial \beta_k \partial \beta_k} \end{array} \right),
\]
which is a \(k \times k\) matrix with typical elements \((i, j)\) given by the second derivative \(\frac{\partial^2 f(\beta)}{\partial \beta_i \partial \beta_j}\). Note that it does not matter if we first take the derivative with respect to the column or the row.
3 SOME SPECIAL FUNCTIONS

First, let $c$ be a $k \times 1$ vector and let $\beta$ be a $k \times 1$ vector of parameters. Next define the scalar function $f(\beta) = c^T \beta$, which maps the $k$ parameters into a single number. It holds that

$$\frac{\partial (c^T \beta)}{\partial \beta} = c.$$  (*)

To see this, we can write the function as

$$f(\beta) = c^T \beta = c_1 \beta_1 + c_2 \beta_2 + \ldots + c_k \beta_k.$$

Taking the derivative with respect to $\beta$ yields

$$\frac{\partial f(\beta)}{\partial \beta} = \begin{pmatrix} \frac{\partial (c_1 \beta_1 + \ldots + c_k \beta_k)}{\partial \beta_1} \\ \vdots \\ \frac{\partial (c_1 \beta_1 + \ldots + c_k \beta_k)}{\partial \beta_k} \end{pmatrix} = \begin{pmatrix} c_1 \\ \vdots \\ c_k \end{pmatrix} = c,$$

which is a $k \times 1$ vector as expected. Also note that since $\beta^T c = c^T \beta$, it holds that

$$\frac{\partial (\beta^T c)}{\partial \beta} = c.$$  (*)

Now, let $A$ be an $n \times k$ matrix and let $\beta$ be a $k \times 1$ vector of parameters. Furthermore define the vector function $g(\beta) = A\beta$, which maps the $k$ parameters into $n$ function values. $g(\beta)$ is an $n \times 1$ vector and the derivative with respect to $\beta'$ is a $n \times k$ matrix given by

$$\frac{\partial (A\beta)}{\partial \beta'} = A.$$  (*)

To see this, write the function as

$$g(\beta) = A\beta = \begin{pmatrix} A_{11} \beta_1 + A_{12} \beta_2 + \ldots + A_{1k} \beta_k \\ \vdots \\ A_{n1} \beta_1 + A_{n2} \beta_2 + \ldots + A_{nk} \beta_k \end{pmatrix},$$

and find the derivative

$$\frac{\partial g(\beta)}{\partial \beta'} = \begin{pmatrix} \frac{\partial (A_{11} \beta_1 + \ldots + A_{1k} \beta_k)}{\partial \beta_1} \\ \vdots \\ \frac{\partial (A_{n1} \beta_1 + \ldots + A_{nk} \beta_k)}{\partial \beta_k} \end{pmatrix} = \begin{pmatrix} A_{11} & \cdots & A_{1k} \\ \vdots & \ddots & \vdots \\ A_{n1} & \cdots & A_{nk} \end{pmatrix} = A.$$

Similarly, if we consider the transposed function, $g(\beta) = \beta' A'$, which is a $1 \times n$ row vector, we can find the $k \times n$ matrix of derivatives as

$$\frac{\partial (\beta' A')}{\partial \beta} = A'.$$  (*)

This is just an application of the result in (3).
Now consider a quadratic function \( f(\beta) = \beta' V \beta \) for some \( k \times k \) matrix \( V \). This function maps the \( k \) parameters into a single number. Here we find the derivatives as the \( k \times 1 \) column vector
\[
\frac{\partial (\beta' V \beta)}{\partial \beta} = (V + V') \beta,
\]
or the row variant
\[
\frac{\partial (\beta' V \beta)}{\partial \beta'} = \beta' (V + V').
\]
If \( V \) is symmetric this reduces to \( 2V \beta \) and \( 2\beta' V \), respectively. To see how this works, consider the simple case \( k = 3 \) and write the function as
\[
\beta' V \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} \begin{pmatrix} V_{11} & V_{12} & V_{13} \\ V_{21} & V_{22} & V_{23} \\ V_{31} & V_{32} & V_{33} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}
= V_{11} \beta_1^2 + V_{22} \beta_2^2 + V_{33} \beta_3^2 + (V_{12} + V_{21}) \beta_1 \beta_2 + (V_{13} + V_{31}) \beta_1 \beta_3 + (V_{23} + V_{32}) \beta_2 \beta_3.
\]
Taking the derivative with respect to \( \beta \), we get
\[
\frac{\partial (\beta' V \beta)}{\partial \beta} = \begin{pmatrix} \frac{\partial (\beta' V \beta)}{\partial \beta_1} \\ \frac{\partial (\beta' V \beta)}{\partial \beta_2} \\ \frac{\partial (\beta' V \beta)}{\partial \beta_3} \end{pmatrix}
= \begin{pmatrix} 2V_{11} \beta_1 + (V_{12} + V_{21}) \beta_2 + (V_{13} + V_{31}) \beta_3 \\ 2V_{22} \beta_2 + (V_{12} + V_{21}) \beta_1 + (V_{23} + V_{32}) \beta_3 \\ 2V_{33} \beta_3 + (V_{13} + V_{31}) \beta_1 + (V_{23} + V_{32}) \beta_2 \end{pmatrix}
= \begin{pmatrix} V_{11} + V_{12} + V_{13} \\ V_{21} + V_{22} + V_{23} \\ V_{31} + V_{32} + V_{33} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}
= (V + V') \beta.
\]

4 The Linear Regression Model

To illustrate the use of matrix differentiation consider the linear regression model in matrix notation,
\[
Y = X \beta + \epsilon,
\]
where \( Y \) is a \( T \times 1 \) vector of stacked left-hand-side variables, \( X \) is a \( T \times k \) matrix of explanatory variables, \( \beta \) is a \( k \times 1 \) vector of parameters to be estimated, and \( \epsilon \) is a \( T \times 1 \) vector of error terms. Here \( k \) is the number of explanatory variables and \( T \) is the number of observations.
One way to motivate the ordinary least squares (OLS) principle is to choose the estimator, $\hat{\beta}_{OLS}$ of $\beta$, as the value that minimizes the sum of squared residuals, i.e.

$$\hat{\beta}_{OLS} = \arg \min_{\beta} \sum_{t=1}^{T} \hat{\epsilon}_t^2 = \arg \min_{\beta} \hat{\epsilon}^2.$$

Looking at the function to be minimized, we find that

$$\hat{\epsilon}^2 = \left( Y' - X \hat{\beta} \right)' \left( Y - X \hat{\beta} \right)$$

$$= \left( Y' - \hat{\beta}' X' \right)' \left( Y - X \hat{\beta} \right)$$

$$= Y'Y - Y'X\hat{\beta} - \hat{\beta}'X'Y + \hat{\beta}'X'X\hat{\beta}$$

$$= Y'Y - 2Y'X\hat{\beta} + \hat{\beta}'X'X\hat{\beta},$$

where the last line uses the fact that $Y'X\hat{\beta}$ and $\hat{\beta}'X'Y$ are identical scalar variables.

Note that $\hat{\epsilon}^2$ is a scalar function and taking the first derivative with respect to $\hat{\beta}$ yields the $k \times 1$ vector

$$\frac{\partial (\hat{\epsilon}^2)}{\partial \hat{\beta}} = \frac{\partial \left( Y'Y - 2Y'X\hat{\beta} + \hat{\beta}'X'X\hat{\beta} \right)}{\partial \hat{\beta}} = -2X'Y + 2X'X\hat{\beta}.$$

Solving the $k$ equations, $\frac{\partial (\hat{\epsilon}^2)}{\partial \hat{\beta}} = 0$, yields the OLS estimator

$$\hat{\beta}_{OLS} = (X'X)^{-1} X'Y,$$

provided that $X'X$ is non-singular.

To make sure that $\hat{\beta}_{OLS}$ is a minimum of $\hat{\epsilon}^2$ and not a maximum, we should formally take the second derivative and make sure that it is positive definite. The $k \times k$ Hessian matrix of second derivatives is given by

$$\frac{\partial^2 (\hat{\epsilon}^2)}{\partial \hat{\beta} \partial \hat{\beta}'} = \frac{\partial \left( -2X'Y + 2X'X\hat{\beta} \right)}{\partial \hat{\beta}'} = 2X'X,$$

which is a positive definite matrix by construction.

**REFERENCES**
