Matrix Differentiation

- Let $a$ and $x$ are $(k 	imes 1)$ vectors and $A$ is an $(k 	imes k)$ matrix.

\[
\frac{\partial (a'x)}{\partial x} = a \quad \frac{\partial (a'x)}{\partial x^r} = a' \\
\frac{\partial (x'Ax)}{\partial x} = (A + A')x \\
\frac{\partial (x'Ax)}{\partial x} = x'x
\]

- We don’t want to prove the claim rigorously. But

\[
a'x = \sum_{i=1}^{k} a_i x_i
\]

If you want to differentiate the function with respect to $x$, you have to differentiate the function with respect to each element of vector $x$ and form a vector-called gradient- with the result.

\[
\frac{\partial (a'x)}{\partial x} = \begin{bmatrix}
\frac{\partial (a'x)}{\partial x_1} \\
\frac{\partial (a'x)}{\partial x_2} \\
\vdots \\
\frac{\partial (a'x)}{\partial x_k}
\end{bmatrix} = \begin{bmatrix}
a_1 \\
a_2 \\
\vdots \\
a_k
\end{bmatrix} = a
\]

You can understand $\frac{\partial (a'x)}{\partial x^r}$ simply as the transpose of $\frac{\partial (a'x)}{\partial x}$. For the differentiation of the quadratic form, consider the summation expression;

\[
x'Ax = \sum_{i=1}^{k} \sum_{j=1}^{k} x_i a_{ij} x_j \\
= x_1 a_{11} x_1 + x_1 a_{12} x_2 + x_1 a_{13} x_3 + \cdots + x_1 a_{1k} x_k \\
+ x_2 a_{21} x_1 + x_2 a_{22} x_2 + x_2 a_{23} x_3 + \cdots + x_2 a_{2k} x_k \\
+ x_3 a_{31} x_1 + x_3 a_{32} x_2 + x_3 a_{33} x_3 + \cdots + x_3 a_{3k} x_k \\
+ \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \\
+ x_k a_{k1} x_1 + x_k a_{k2} x_2 + x_k a_{k3} x_3 + \cdots + x_k a_{kk} x_k
\]

Now, we have

\[
\frac{\partial (x'Ax)}{\partial x_1} = 2a_{11} x_1 + a_{12} x_2 + a_{13} x_3 + \cdots + a_{1k} x_k \\
+ x_2 a_{21} + x_3 a_{31} + \cdots + x_k a_{k1} \\
= a_{11} x_1 + a_{12} x_2 + a_{13} x_3 + \cdots + a_{1k} x_k \\
+ a_{11} x_1 + a_{21} x_2 + a_{31} x_3 + \cdots + a_{k1} x_k \\
= A_1 x + A^1 x = (A_1 + A^1) x
\]

where $A_1$ is the first row of the matrix $A$ and $A^1$ is the first column of the matrix $A$. Similarly,

\[
\frac{\partial (x'Ax)}{\partial x_2} = a_{21} x_1 + 2a_{22} x_2 + a_{23} x_3 + \cdots + a_{2k} x_k \\
+ x_1 a_{12} + x_3 a_{32} + \cdots + x_k a_{k2} \\
= a_{21} x_1 + a_{22} x_2 + a_{23} x_3 + \cdots + a_{2k} x_k \\
+ a_{12} x_1 + a_{22} x_2 + a_{32} x_3 + \cdots + a_{k2} x_k \\
= A_2 x + A^2 x = (A_2 + A^2) x
\]
You see the pattern emerging from the calculation. In general,

\[
\frac{\partial (x'Ax)}{\partial x_i} = (A_i + A') x \quad i = 1, 2, \ldots, k
\]

We stack the vectors to get:

\[
\frac{\partial (x'Ax)}{\partial x} = \begin{bmatrix}
\frac{\partial(x'Ax)}{\partial x_1} \\
\frac{\partial(x'Ax)}{\partial x_2} \\
\vdots \\
\frac{\partial(x'Ax)}{\partial x_k}
\end{bmatrix} = \begin{bmatrix}
A_1 \\
A_2 \\
\vdots \\
A_k
\end{bmatrix} + \begin{bmatrix}
A_1 \\
A_2 \\
\vdots \\
A_k
\end{bmatrix} = (A + A') x
\]

You can verify the result for \( \frac{\partial(x'Ax)}{\partial A} = xx' \) with a similar argument.

- Consider the least squares problem:

\[
S(b) = (y - Xb)' (y - Xb) = (y' - b'X')(y - Xb)
= y'y - y'Xb - b'X'y + b'X'Xb
= y'y - 2y'Xb + b'X'Xb
\]

Note that \( y'X \) is a vector, \( b \) is a vector and \( X'X \) is a matrix in the formula above. Hence,

\[
\frac{S(b)}{\partial b} = -2X'y + \left[(X'X) + (X'X)' \right] b
= -2X'y + 2X'Xb
\]

### Least Squares Estimator in Matrix Form

- The model is given by

\[
y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \cdots + \beta_k x_{ik} + \varepsilon_i
\]

\[
E(\varepsilon_i) = 0, E(\varepsilon_i^2) = \sigma^2, E(\varepsilon_i \varepsilon_j) = 0 \text{ when } i \neq j
\]

In matrix notation

\[
y = X\beta + \varepsilon
\]

\[
E(\varepsilon) = 0, E(\varepsilon \varepsilon') = \sigma^2 I
\]

- The least squares estimator is

\[
\hat{\beta} = (X'X)^{-1} X'y
\]

- Unbiasedness of \( \hat{\beta} \)

\[
E(\hat{\beta}) = E\left[(X'X)^{-1} X'y \right] = E\left[(X'X)^{-1} X' (X\beta + \varepsilon) \right]
= E\left[\beta + (X'X)^{-1} X' \varepsilon \right] = \beta + (X'X)^{-1} X' E(\varepsilon) = \beta
\]

- Variance of \( \hat{\beta} \)

\[
Var(\hat{\beta}) = E\left[\left(\hat{\beta} - E(\hat{\beta})\right) \left(\hat{\beta} - E(\hat{\beta})\right)'\right]
= E\left[\left(\hat{\beta} - \beta\right) \left(\hat{\beta} - \beta\right)\right]
= E\left[(X'X)^{-1} \varepsilon \varepsilon' X (X'X)^{-1} \right] = (X'X)^{-1} X' E(\varepsilon) E(\varepsilon') X (X'X)^{-1}
= \sigma^2 (X'X)^{-1} X' I X (X'X)^{-1} = \sigma^2 (X'X)^{-1}
\]
• Residual vector and $M$ matrix

\[ e = y - X\hat{\beta} = y - X(X'X)^{-1}X'y = \begin{bmatrix} I - X(X'X)^{-1}X' \end{bmatrix} y \]

\[ = My \]

The matrices $P = X(X'X)^{-1}X'$ and $M = (I - P)$ are called projection matrices. Especially, $P$ is the projection matrix onto space spanned by columns of $X$ and $M$ is the projection onto the space orthogonal to the space spanned by columns of $X$. When people simply say the projection matrix, they mean $P, P$ and $M$ have a nice interpretation in terms of geometry..

• Properties of $P$ and $M$ matrix

(i) Both $P$ and $M$ are symmetric and idempotent. - proof is easy.

(ii) $\rho(P) = k$ and $\rho(M) = N - k$.

\[ \rho(P) = \rho \left( X(X'X)^{-1}X' \right) = \min \left( \rho(X) , \rho \left( (X'X)^{-1} \right) , \rho(X') \right) = \min(k,k,k) = k \]

\[ \rho(M) = \text{tr}(M) = \text{tr}(I - P) = \text{tr}(I) - \text{tr}(P) = \text{tr}(I) - \rho(P) = N - k \]

Note that the rank of an idempotent matrix is its trace and both $P$ and $M$ are idempotent.

(iii) $MX = 0$ and $P + M = I$

\[ MX = \begin{bmatrix} I - X(X'X)^{-1}X' \end{bmatrix} X = X - X(X'X)^{-1}X'X = X - X = 0 \]

\[ P + M = X(X'X)^{-1}X' + \begin{bmatrix} I - X(X'X)^{-1}X' \end{bmatrix} = I \]

• Estimation of $\sigma^2$

Since $\varepsilon$ is unobservable by definition, we do not know its variance $\sigma^2$, either. However, we can estimate it using the sum of squared residuals.

\[ \sum_{i=1}^{N} \left( y_i - \hat{\beta}_1 - \hat{\beta}_2x_{i2} - \cdots - \hat{\beta}_kx_{ik} \right)^2 = \sum_{i=1}^{N} e_i^2 = e'e \]

Note that

\[ e = \left( y - X\hat{\beta} \right) = \left( y - X(X'X)^{-1}X' \right)y = \left( I - X(X'X)^{-1}X' \right)y = My \]

\[ = M(X\beta + \varepsilon) = MX\beta + M\varepsilon = M\varepsilon \]

Hence,

\[ e'e = (M\varepsilon)'(M\varepsilon) = \varepsilon'M'M\varepsilon = \varepsilon'MM\varepsilon = \varepsilon'M\varepsilon \]

Now, taking expectation on both sides,

\[ E(e'e) = E(\varepsilon'M\varepsilon) \]

\[ = E[\text{tr}(\varepsilon'M\varepsilon)] \text{ since } \varepsilon'M\varepsilon \text{ is scalar} \]

\[ = E[\text{tr}(M\varepsilon\varepsilon')] \text{ since } \text{tr}(AB) = \text{tr}(BA) \]

\[ = \text{tr}[E(M\varepsilon\varepsilon')] \text{ since expectation is a linear operator} \]

\[ = \text{tr}[ME(\varepsilon\varepsilon')] \text{ since } M \text{ is non-stochastic} \]

\[ = \text{tr}[M\sigma^2I] = \sigma^2\text{tr}(M) \text{ since } \text{tr}(aA) = atr(A) \text{ when } a \text{ is a scalar} \]

\[ = \sigma^2\rho(M) \text{ since } M \text{ is idempotent} \]

\[ = \sigma^2(N - k) \text{ from the argument above} \]

Therefore, to get an unbiased estimator of $\sigma^2$, we propose;

\[ s^2 = \frac{e'e}{N - k} \]
Then, 
\[ E(s^2) = \frac{1}{(N-k)}E(e'e) = \frac{\sigma^2(N-k)}{(N-k)} = \sigma^2 \]

- Distribution of \( s^2 \)

Fact-you can actually prove this, try-.

\[ \frac{(N-k)s^2}{\sigma^2} = \frac{e'e}{\sigma^2} \sim \chi^2(N-k) \]

Then,

\[ E\left(\frac{e'e}{\sigma^2}\right) = (N-k) \Rightarrow E(e'e) = \sigma^2(N-k) \]

\[ \text{Var}\left(\frac{e'e}{\sigma^2}\right) = 2(N-k) \Rightarrow \text{Var}(e'e) = 2\sigma^4(N-k) \]

- \( A \) matrix

\[ A = I - \mathbf{1}(\mathbf{1}'\mathbf{1})^{-1}\mathbf{1}' \]

where \( \mathbf{1} \) is an \((N \times 1)\) vector whose elements are all 1.

If we postmultiply \( A \) matrix with a vector, say \( y \), it will results in a vector in mean deviation form:

\[ Ay = \left[ I - \mathbf{1}(\mathbf{1}'\mathbf{1})^{-1}\mathbf{1}' \right] y = y - \mathbf{1}(\mathbf{1}'\mathbf{1})^{-1}\mathbf{1}'y \]

Why do we introduce the matrix \( A \)? There is a good reason for it. Consider the classical multiple regression model in the following form:

\[ y = X\beta + \varepsilon = \begin{bmatrix} 1 & X_2 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + \varepsilon = \beta_1 \mathbf{1} + X_2 \beta_2 + \varepsilon \]

where we partitioned \( X \) matrix into the column corresponding to the constant term, \( \mathbf{1} \), and the columns corresponding to all the other regressors, \( X_2 \). Then,

\[ \hat{\beta} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = (X'X)^{-1}X'y = \left( \begin{bmatrix} 1' \\ X_2' \end{bmatrix} \begin{bmatrix} 1 & X_2 \end{bmatrix} \right)^{-1} \begin{bmatrix} 1' \\ X_2' \end{bmatrix} y \]

\[ = \begin{bmatrix} 1'X_2 \\ X_2' \end{bmatrix} ^{-1} \begin{bmatrix} 1'y \\ X_2'y \end{bmatrix} \]
What is the lower right block of the inverse matrix? From the formula for the inverse of the partitioned matrix,

\[
\hat{\beta}_2 = - \left( X'_2 X_2 - X'_2 1 (1' 1)^{-1} 1' X_2 \right)^{-1} X'_2 1 (1' 1)^{-1} 1'y \\
+ \left( X'_2 X_2 - X'_2 1 (1' 1)^{-1} 1' X_2 \right)^{-1} X'_2 y
\]

\[
= - X'_2 \left( I - 1 (1' 1)^{-1} 1' \right) X_2^{-1} X'_2 1 (1' 1)^{-1} 1'y \\
+ X'_2 \left( I - 1 (1' 1)^{-1} 1' \right) X_2^{-1} X'_2 y
\]

\[
= - [X'_2 AX_2]^{-1} X'_2 1 (1' 1)^{-1} 1'y + [X'_2 AX_2]^{-1} X'_2 y
\]

\[
= [X'_2 AX_2]^{-1} X'_2 \left[ I - 1 (1' 1)^{-1} 1' \right] y = [X'_2 AX_2]^{-1} [X'_2 Ay]
\]

\[
= [X'_2 A' AX_2]^{-1} [X'_2 A' Ay] = [(AX_2)' (AX_2)]^{-1} [(AX_2)' (Ay)]
\]

Now consider another approach to the estimation;

\[
y = X\beta + \varepsilon = \begin{bmatrix} 1 & X_2 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + \varepsilon = \beta_1 1 + X_2 \beta_2 + \varepsilon
\]

Premultiplying both sides with \( A \) gives;

\[
Ay = \beta_1 A1 + AX_2 \beta_2 + A\varepsilon
\]

\[
= AX_2 \beta_2 + A\varepsilon
\]

since

\[
A1 = \left[ I - 1 (1' 1)^{-1} 1' \right] 1 = 0
\]

Now, define \( Ay = y^* \), \( AX_2 = X_2^* \), and \( A\varepsilon = \varepsilon^* \) to get

\[
y^* = X_2^* \beta_2 + \varepsilon^*
\]

The least squares estimator is given by;

\[
\hat{\beta}_2 = (X_2'^* X_2^*)^{-1} X_2'^* y^* = [(AX_2)' (AX_2)]^{-1} [(AX_2)' Ay]
\]

\[
= [X'_2 A' AX_2]^{-1} [X'_2 A' Ay] = [X'_2 AX_2]^{-1} [X'_2 Ay]
\]

which is identical to the least squares estimator for \( \beta_2 \) in the original model. The transformed regression does not include a constant term and the data used in the transformed regression is in mean deviation forms as shown above- \( Ay \) and \( AX_2 \). In sum, the slope estimates from the original regression - one with a constant term and untransformed data- is identical to those from the transformed regression - one without a constant term and with data in mean deviation forms. Then, what about the constant term? The least squares estimator for the constant term is given by;

\[
\hat{\beta}_1 = \overline{y} - \hat{\beta}_2 \overline{x}_2 - \hat{\beta}_3 \overline{x}_3 - \cdots - \hat{\beta}_k \overline{x}_k
\]

which can be derived easily from the first order condition.

- Variance matrix from the two regressions

In model without transformation, we know that

\[
Var (\hat{\beta}) = \begin{bmatrix} Var (\hat{\beta}_1) & Cov (\hat{\beta}_1, \hat{\beta}_2) \\ Cov (\hat{\beta}_1, \hat{\beta}_2) & Var (\hat{\beta}_2) \end{bmatrix}
\]

\[
= \sigma^2 (X' X)^{-1} = \sigma^2 \begin{bmatrix} 1'1 & 1'X_2 \\ X'_2 1 & X'_2 X_2 \end{bmatrix}^{-1}
\]
Therefore,

\[
Var \left( \hat{\beta}_2 \right) = \sigma^2 \left( X_2' X_2 - X_2' I (1'1)^{-1} 1' X_2 \right)^{-1}
\]

\[
= \sigma^2 \left[ X_2' \left( I - 1 (1'1)^{-1} 1' \right) X_2 \right]^{-1} = \sigma^2 [X_2' AX_2]^{-1}
\]

The variance matrix of \( \hat{\beta}_2 \) is identical to that from the regression in mean deviation forms since

\[
Var \left( \hat{\beta}_2 \right) = \sigma^2 (X_2' X_2^*)^{-1} = \sigma^2 (X_2' AX_2)^{-1}
\]

Therefore, the two regressions result in the same estimates of the slope coefficients and variances of the estimates.

- \( R^2 \) in the multiple regression analysis;
  
  \( R^2 \) is defined as the ratio between the explained sum of squares and the total sum of squares;

\[
R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}
\]

TSS is the sum of squares of variations in the dependent variable around the mean;

\[
TSS = \sum_{i=1}^{N} (y_i - \bar{y})^2 = \sum_{i=1}^{N} (y_i - \bar{y}) (y_i - \bar{y}) = (Ay)' (Ay) = y'Ay
\]

On the other hand,

\[
y'Ay = (Ay)' (Ay) = (A\hat{y} + Ae)' (A\hat{y} + Ae) = (A\hat{y} + e)' (A\hat{y} + e)
\]

\[
= \hat{y}' A\hat{y} + e'e
\]

Hence,

\[
R^2 = \frac{\hat{y}' A\hat{y}}{y'Ay} = \frac{\left( X\hat{\beta} \right)' A \left( X\hat{\beta} \right)}{y'Ay} = \frac{\hat{\beta}' (X'AX) \hat{\beta}}{y'Ay}
\]

\[
= 1 - \frac{e'e}{y'Ay}
\]