Econ 620

Matrix Differentiation

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

$$\frac{\partial (a'x)}{\partial x} = a \qquad \frac{\partial (a'x)}{\partial x'} = a'$$
$$\frac{\partial (x'Ax)}{\partial x} = (A + A')\,x \qquad \frac{\partial (x'Ax)}{\partial x \partial x'} = (A + A')$$
$$\frac{\partial (x'Ax)}{\partial A} = xx'$$

- 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.

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

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

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

Now, we have

$$\begin{aligned} \frac{\partial (x'Ax)}{\partial x_1} &= 2a_{11} x_1 + a_{12} x_2 + a_{13} x_3 + \cdots + a_{1k} x_k \\ &+ x_2 a_{21} + x_3 a_{31} + \cdots + x_k a_{k1} \\ &= a_{11} x_1 + a_{12} x_2 + a_{13} x_3 + \cdots + a_{1k} x_k \\ &+ a_{11} x_1 + a_{21} x_2 + a_{31} x_3 + \cdots + a_{k1} x_k \\ &= A_1 x + A^1 x = \left(A_1 + A^1\right) x \end{aligned}$$

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

$$\begin{aligned} \frac{\partial (x'Ax)}{\partial x_2} &= a_{21} x_1 + 2a_{22} x_2 + a_{23} x_3 + \cdots + a_{2k} x_k \\ &+ x_1 a_{12} + x_3 a_{32} + \cdots + x_k a_{k2} \\ &= a_{21} x_1 + a_{22} x_2 + a_{23} x_3 + \cdots + a_{2k} x_k \\ &+ a_{12} x_1 + a_{22} x_2 + a_{32} x_3 + \cdots + a_{k2} x_k \\ &= A_2 x + A^2 x = \left(A_2 + A^2\right) x \end{aligned}$$

You see the pattern emerging from the calculation. In general,

$$\frac{\partial\left(x'Ax\right)}{\partial x_i} = \left(A_i + A^i\right)x \qquad i = 1, 2, \cdots, k$$

We stack the vectors to get;

$$\frac{\partial\left(x'Ax\right)}{\partial x} = \begin{bmatrix} \frac{\partial(x'Ax)}{\partial x_1} \\ \frac{\partial(x'Ax)}{\partial x_2} \\ \cdots \\ \frac{\partial(x'Ax)}{\partial x_k} \end{bmatrix} = \left[\begin{bmatrix} A_1 \\ A_2 \\ \cdots \\ A_k \end{bmatrix} + \begin{bmatrix} A^1 \\ A^2 \\ \cdots \\ A^k \end{bmatrix}\right] x = \left(A + A'\right)x$$

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

- Consider the least squares problem;

$$\begin{aligned} S(b) &= (y - Xb)'(y - Xb) = (y' - b'X')(y - Xb) \\ &= y'y - y'Xb - b'X'y + b'X'Xb \\ &= y'y - 2y'Xb + b'X'Xb \end{aligned}$$

Note that $y'X$ is a' vector, b is x vector and $X'X$ is A matrix in the formula above. Hence,

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

Least Squares Estimator in Matrix Form

- The model is given by

$$y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \cdots + \beta_k x_{ik} + \varepsilon_i$$
$$E(\varepsilon_i) = 0, E\left(\varepsilon_i^2\right) = \sigma^2, E(\varepsilon_i \varepsilon_j) = 0 \text{ when } i \neq j$$

In matrix notation

$$y = X\beta + \varepsilon$$
$$E(\varepsilon) = \mathbf{0}, E(\varepsilon\varepsilon') = \sigma^2 I$$

- The least squares estimator is

$$\widehat{\beta} = (X'X)^{-1} X'y$$

- Unbiasedness of $\widehat{\beta}$

$$\begin{aligned} E\left(\widehat{\beta}\right) &= E\left[(X'X)^{-1} X'y\right] = E\left[(X'X)^{-1} X'(X\beta + \varepsilon)\right] \\ &= E\left[\beta + (X'X)^{-1} X'\varepsilon\right] = \beta + (X'X)^{-1} X'E(\varepsilon) = \beta \end{aligned}$$

- Variance of $\widehat{\beta}$

$$\begin{aligned} Var\left(\widehat{\beta}\right) &= E\left[\left(\widehat{\beta} - E\left(\widehat{\beta}\right)\right)\left(\widehat{\beta} - E\left(\widehat{\beta}\right)\right)'\right] = E\left[\left(\widehat{\beta} - \beta\right)\left(\widehat{\beta} - \beta\right)\right] \\ &= E\left[(X'X)^{-1} X'\varepsilon\varepsilon'X(X'X)^{-1}\right] = (X'X)^{-1} X'E(\varepsilon\varepsilon')X(X'X)^{-1} \\ &= \sigma^2 (X'X)^{-1} X'IX(X'X)^{-1} = \sigma^2 (X'X)^{-1} \end{aligned}$$

- Residual vector and M matrix

$$
\begin{aligned}
e &= y - X\widehat{\beta} = y - X\left(X'X\right)^{-1}X'y = \left[I - X\left(X'X\right)^{-1}X'\right]y \\
&= My
\end{aligned}
$$

The matrices $P = X\left(X'X\right)^{-1}X'$ and $M = (I - P)$ are called projection matrix. 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 $\rho(M) = N - k$.

$$
\begin{aligned}
\rho(P) &= \rho\left(X\left(X'X\right)^{-1}X'\right) = \min\left(\rho(X), \rho\left(\left(X'X\right)^{-1}\right), \rho(X')\right) = \min(k,k,k) = k \\
\rho(M) &= tr(M) = tr(I - P) = tr(I) - tr(P) = tr(I) - \rho(P) = N - k
\end{aligned}
$$

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

(iii) $MX = \mathbf{0}$ and $P + M = I$

$$
\begin{aligned}
MX &= \left[I - X\left(X'X\right)^{-1}X'\right]X = X - X\left(X'X\right)^{-1}X'X = X - X = \mathbf{0} \\
P + M &= X\left(X'X\right)^{-1}X' + \left[I - X\left(X'X\right)^{-1}X'\right] = I
\end{aligned}
$$

- Estimation of σ^2

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

$$
\sum_{i=1}^{N}\left(y_i - \widehat{\beta}_1 - \widehat{\beta}_2 x_{i2} - \cdots - \widehat{\beta}_k x_{ik}\right)^2 = \sum_{i=1}^{N} e_i^2 = e'e
$$

Note that

$$
\begin{aligned}
e &= \left(y - X\widehat{\beta}\right) = \left(y - X\left(X'X\right)^{-1}X'\right)y = \left(I - X\left(X'X\right)^{-1}X'\right)y = My \\
&= M(X\beta + \varepsilon) = MX\beta + M\varepsilon = M\varepsilon
\end{aligned}
$$

Hence,

$$
e'e = (M\varepsilon)'(M\varepsilon) = \varepsilon'M'M\varepsilon = \varepsilon'MM\varepsilon = \varepsilon'M\varepsilon
$$

Now, taking expectation on both sides,

$$
\begin{aligned}
E(e'e) &= E(\varepsilon'M\varepsilon) \\
&= E\left[tr(\varepsilon'M\varepsilon)\right] \text{ since } \varepsilon'M\varepsilon \text{ is scalar} \\
&= E\left[tr(M\varepsilon\varepsilon')\right] \text{ since } tr(AB) = tr(BA) \\
&= tr\left[E(M\varepsilon\varepsilon')\right] \text{ since expectation is a linear operator} \\
&= tr\left[ME(\varepsilon\varepsilon')\right] \text{ since } M \text{ is non-stochastic} \\
&= tr\left[M\sigma^2 I\right] = \sigma^2 tr(M) \text{ since } tr(aA) = a\,tr(A) \text{ when } a \text{ is a scalar} \\
&= \sigma^2\rho(M) \text{ since } M \text{ is idempotent} \\
&= \sigma^2(N - k) \text{ from the argument above}
\end{aligned}
$$

Therefore, to get an unbiased estimator of σ^2, we propose;

$$
s^2 = \frac{e'e}{(N - k)}
$$

Then,

$$E\left(s^2\right) = \frac{1}{(N-k)} E\left(e'e\right) = \frac{\sigma^2\left(N-k\right)}{(N-k)} = \sigma^2$$

- Distribution of s^2

  Fact-you can actually prove this, try-.

$$\frac{(N-k)\,s^2}{\sigma^2} = \frac{e'e}{\sigma^2} \sim \chi^2\left(N-k\right)$$

Then,

$$\begin{aligned}
E\left(\frac{e'e}{\sigma^2}\right) &= (N-k) \Rightarrow E\left(e'e\right) = \sigma^2\left(N-k\right) \\
Var\left(\frac{e'e}{\sigma^2}\right) &= 2\left(N-k\right) \Rightarrow Var\left(e'e\right) = 2\sigma^4\left(N-k\right)
\end{aligned}$$

- A matrix

$$A \equiv I - \mathbf{1}\left(\mathbf{1}'\mathbf{1}\right)^{-1}\mathbf{1}'$$

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;

$$\begin{aligned}
Ay &= \left[I - \mathbf{1}\left(\mathbf{1}'\mathbf{1}\right)^{-1}\mathbf{1}'\right] y = y - \mathbf{1}\left(\mathbf{1}'\mathbf{1}\right)^{-1}\mathbf{1}'y \\
&= \begin{bmatrix} y_1 \\ y_2 \\ \cdots \\ y_N \end{bmatrix} - \begin{bmatrix} 1 \\ 1 \\ \cdots \\ 1 \end{bmatrix} \left[\begin{bmatrix} 1 & 1 & \cdots & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ \cdots \\ 1 \end{bmatrix}\right]^{-1} \begin{bmatrix} 1 & 1 & \cdots & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \cdots \\ y_N \end{bmatrix} \\
&= \begin{bmatrix} y_1 \\ y_2 \\ \cdots \\ y_N \end{bmatrix} - \begin{bmatrix} 1 \\ 1 \\ \cdots \\ 1 \end{bmatrix} \frac{1}{N} \begin{bmatrix} 1 & 1 & \cdots & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \cdots \\ y_N \end{bmatrix} \\
&= \begin{bmatrix} y_1 \\ y_2 \\ \cdots \\ y_N \end{bmatrix} - \frac{1}{N} \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ \cdots & \cdots & \cdots & \cdots \\ 1 & 1 & \cdots & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \cdots \\ y_N \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ \cdots \\ y_N \end{bmatrix} - \frac{1}{N} \begin{bmatrix} \sum_{i=1}^N y_i \\ \sum_{i=1}^N y_i \\ \cdots \\ \sum_{i=1}^N y_i \end{bmatrix} \\
&= \begin{bmatrix} y_1 \\ y_2 \\ \cdots \\ y_N \end{bmatrix} - \begin{bmatrix} \overline{y} \\ \overline{y} \\ \cdots \\ \overline{y} \end{bmatrix} = \begin{bmatrix} y_1 - \overline{y} \\ y_2 - \overline{y} \\ \cdots \\ y_N - \overline{y} \end{bmatrix}
\end{aligned}$$

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

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

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,

$$\begin{aligned}
\widehat{\beta} = \begin{bmatrix} \widehat{\beta}_1 \\ \widehat{\beta}_2 \end{bmatrix} &= \left(X'X\right)^{-1} X'y = \left(\begin{bmatrix} \mathbf{1}' \\ X_2' \end{bmatrix} \begin{bmatrix} \mathbf{1} & X_2 \end{bmatrix}\right)^{-1} \begin{bmatrix} \mathbf{1}' \\ X_2' \end{bmatrix} y \\
&= \begin{bmatrix} \mathbf{1}'\mathbf{1} & \mathbf{1}'X_2 \\ X_2'\mathbf{1} & X_2'X_2 \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{1}'y \\ X_2'y \end{bmatrix}
\end{aligned}$$

What is the lower right block of the inverse matrix? From the formula for the inverse of the partitioned matrix,

$$
\begin{aligned}
\widehat{\beta}_2 &= -\left(X_2'X_2 - X_2'\mathbf{1}\left(\mathbf{1}'\mathbf{1}\right)^{-1}\mathbf{1}'X_2\right)^{-1} X_2'\mathbf{1}\left(\mathbf{1}'\mathbf{1}\right)^{-1}\mathbf{1}'y \\
&+ \left(X_2'X_2 - X_2'\mathbf{1}\left(\mathbf{1}'\mathbf{1}\right)^{-1}\mathbf{1}'X_2\right)^{-1} X_2'y \\
&= -\left[X_2'\left(I - \mathbf{1}\left(\mathbf{1}'\mathbf{1}\right)^{-1}\mathbf{1}'\right)X_2\right]^{-1} X_2'\mathbf{1}\left(\mathbf{1}'\mathbf{1}\right)^{-1}\mathbf{1}'y \\
&+ \left[X_2'\left(I - \mathbf{1}\left(\mathbf{1}'\mathbf{1}\right)^{-1}\mathbf{1}'\right)X_2\right]^{-1} X_2'y \\
&= -\left[X_2'AX_2\right]^{-1} X_2'\mathbf{1}\left(\mathbf{1}'\mathbf{1}\right)^{-1}\mathbf{1}'y + \left[X_2'AX_2\right]^{-1} X_2'y \\
&= \left[X_2'AX_2\right]^{-1} X_2'\left[I - \mathbf{1}\left(\mathbf{1}'\mathbf{1}\right)^{-1}\mathbf{1}'\right]y = \left[X_2'AX_2\right]^{-1}\left[X_2'Ay\right] \\
&= \left[X_2'A'AX_2\right]^{-1}\left[X_2'A'Ay\right] = \left[\left(AX_2\right)'\left(AX_2\right)\right]^{-1}\left[\left(AX_2\right)'\left(Ay\right)\right]
\end{aligned}
$$

Now consider another approach to the estimation;

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

Premultiplying both sides with A gives;

$$
\begin{aligned}
Ay &= \beta_1 A\mathbf{1} + AX_2\beta_2 + A\varepsilon \\
&= AX_2\beta_2 + A\varepsilon
\end{aligned}
$$

since

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

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;

$$
\begin{aligned}
\widehat{\beta}_2 &= \left(X_2^{*\prime}X_2^*\right)^{-1} X_2^{*\prime}y^* = \left[\left(AX_2\right)'\left(AX_2\right)\right]^{-1}\left[\left(AX_2\right)'Ay\right] \\
&= \left[X_2'A'AX_2\right]^{-1}\left[X_2'A'Ay\right] = \left[X_2'AX_2\right]^{-1}\left[X_2'Ay\right]
\end{aligned}
$$

which is identical to the least squares estimator for β_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;

$$
\widehat{\beta}_1 = \overline{y} - \widehat{\beta}_2\overline{x}_2 - \widehat{\beta}_3\overline{x}_3 - \cdots - \widehat{\beta}_k\overline{x}_k
$$

which can be derived easily from the first order condition.

- Variance matrix from the two regressions

  In model without transformation, we know that

$$
\begin{aligned}
Var\left(\widehat{\beta}\right) &= \begin{bmatrix} Var\left(\widehat{\beta}_1\right) & Cov\left(\widehat{\beta}_1, \widehat{\beta}_2\right) \\ Cov\left(\widehat{\beta}_1, \widehat{\beta}_2\right) & Var\left(\widehat{\beta}_2\right) \end{bmatrix} \\
&= \sigma^2\left(X'X\right)^{-1} = \sigma^2 \begin{bmatrix} \mathbf{1}'\mathbf{1} & \mathbf{1}'X_2 \\ X_2'\mathbf{1} & X_2'X_2 \end{bmatrix}^{-1}
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
Var\left(\widehat{\beta}_2\right) &= \sigma^2 \left(X_2'X_2 - X_2'\mathbf{1}\left(\mathbf{1}'\mathbf{1}\right)^{-1}\mathbf{1}'X_2\right)^{-1} \\
&= \sigma^2 \left[X_2'\left(I - \mathbf{1}\left(\mathbf{1}'\mathbf{1}\right)^{-1}\mathbf{1}'\right)X_2\right]^{-1} = \sigma^2\left[X_2'AX_2\right]^{-1}
\end{aligned}
$$

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

$$
Var\left(\widehat{\beta}_2\right) = \sigma^2\left(X_2^{*\prime}X_2^*\right)^{-1} = \sigma^2\left(X_2'AX_2\right)^{-1}
$$

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;

$$
TSS = \sum_{i=1}^{N}\left(y_i - \overline{y}\right)^2 = \sum_{i=1}^{N}\left(y_i - \overline{y}\right)\left(y_i - \overline{y}\right) = \left(Ay\right)'\left(Ay\right) = y'Ay
$$

On the other hand,

$$
\begin{aligned}
y'Ay &= \left(Ay\right)'\left(Ay\right) = \left(A\widehat{y} + Ae\right)'\left(A\widehat{y} + Ae\right) = \left(A\widehat{y} + e\right)'\left(A\widehat{y} + e\right) \\
&= \widehat{y}'A\widehat{y} + e'e
\end{aligned}
$$

Hence,

$$
\begin{aligned}
R^2 &= \frac{\widehat{y}'A\widehat{y}}{y'Ay} = \frac{\left(X\widehat{\beta}\right)'A\left(X\widehat{\beta}\right)}{y'Ay} = \frac{\widehat{\beta}'\left(X'AX\right)\widehat{\beta}}{y'Ay} \\
&= 1 - \frac{e'e}{y'Ay}
\end{aligned}
$$