This material comes from Hansen Appendix A and Chapter 2.

Linear Regression Notes 1: Conditional Expectation and Projection

Review of Matrix Algebra

H A.1

A **vector** a is a $k \times 1$ list of numbers. We will follow the convention of (primarily) using column vectors. That is,

$$a = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_k \end{pmatrix}$$

If $k = 1$, then a is a scalar. A **matrix A** is a $k \times r$ rectangular array of numbers which we will write as

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1r} \\ a_{21} & a_{22} & \cdots & a_{2r} \\ \vdots & \vdots & \ddots & \vdots \\ a_{k1} & a_{k2} & \cdots & a_{kr} \end{bmatrix}$$

I will typically capitalize and use bold-font to indicate a matrix in the course notes and will underline it on the board, e.g., $\underline{A}$ (since it is hard to write in bold on the board).

The **transpose** of a matrix, which will denote by $\mathbf{A}'$ is obtained by flipping the matrix on its diagonal. That is,

$$\mathbf{A}' = \begin{bmatrix} a_{11} & a_{21} & \cdots & a_{k1} \\ a_{12} & a_{22} & \cdots & a_{k2} \\ \vdots & \vdots & \ddots & \vdots \\ a_{1r} & a_{2r} & \cdots & a_{kr} \end{bmatrix}$$

Notice that $\mathbf{A}'$ is an $r \times k$ matrix. For a $k \times 1$ vector a, its tranpose, a', is a $1 \times k$ vector. For a scalar a, $a = a'$.

A matrix is **square** if $k = r$. A square matrix is **symmetric** if $\mathbf{A} = \mathbf{A}'$. A square matrix is **diagonal** if the off-diagonal elements are all zero. The **identity matrix** is the diagonal matrix where all the elements on the diagonal are equal to 1. It is common to denote the $k \times k$ identity

matrix by

$$\mathbf{I}_k = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}$$

Matrix Addition

H A.3

If two matrices $\mathbf{A} = (a_{ij})$ and $\mathbf{B} = (b_{ij})$ (here the notation just means that a_{ij} and b_{ij} are elements of each matrix) have the same dimension, then they can be added, and

$$\mathbf{A} + \mathbf{B} = (a_{ij} + b_{ij})$$

Matrix addition is commutative, that is, $\mathbf{A} + \mathbf{B} = \mathbf{B} + \mathbf{A}$. It is also associative: $\mathbf{A} + (\mathbf{B} + \mathbf{C}) = (\mathbf{A} + \mathbf{B}) + \mathbf{C}$.

Matrix Multiplication

H A.4

Let c denote a scalar, then (we define) $\mathbf{A}c = c\mathbf{A} = (a_{ij}c)$. If a and b are both $k \times 1$ vectors, then their **inner product** is

$$a'b = a_1b_1 + a_2b_2 + \cdots + a_kb_k = \sum_{j=1}^{k} a_jb_j$$

Further, notice that $a'b = b'a$. a and b are said to be **orthogonal** if $a'b = 0$.

If $\mathbf{A}$ is $k \times r$ and $\mathbf{B}$ is $r \times s$ (that is, the number of columns of $\mathbf{A}$ is the same as the number of rows of $\mathbf{B}$), then $\mathbf{A}$ and $\mathbf{B}$ are said to be **conformable** and the matrix produce $\mathbf{AB}$ is defined as

$$\mathbf{AB} = \begin{bmatrix} a_1' \\ a_2' \\ \vdots \\ a_k' \end{bmatrix} \begin{bmatrix} b_1 & b_2 & \cdots & b_s \end{bmatrix} = \begin{bmatrix} a_1'b_1 & a_1'b_2 & \cdots & a_1'b_s \\ a_2'b_1 & a_2'b_2 & \cdots & a_2'b_s \\ \vdots & \vdots & \ddots & \vdots \\ a_k'b_1 & a_k'b_2 & \cdots & a_k'b_s \end{bmatrix}$$

where, for example, $a_1' = (a_{11}, a_{12}, \ldots, a_{1r})$ (which is the first row of $\mathbf{A}$) and $b_1 = (b_{11}, b_{21}, \ldots, b_{r1})'$ is the first column of $\mathbf{B}$. Notice that the product is a $k \times s$ matrix.

Matrix multiplication is not commutative, i.e., in general, $\mathbf{AB} \neq \mathbf{BA}$. But it is associative: $\mathbf{A}(\mathbf{BC}) = (\mathbf{AB})\mathbf{C}$. And it is distributive: $\mathbf{A}(\mathbf{B} + \mathbf{C}) = \mathbf{AB} + \mathbf{AC}$.

Trace

H A.5

The **trace** of $k \times k$ square matrix $\mathbf{A}$ is the sum of its diagonal elements:

$$\text{tr}(\mathbf{A}) = \sum_{i=1}^{k} a_{ii}$$

Here are some useful properties of trace (where $\mathbf{A}$ and $\mathbf{B}$ are square matrices and c is a scalar):

1. $\text{tr}(c\mathbf{A}) = c\text{tr}(\mathbf{A})$
2. $\text{tr}(\mathbf{A}') = \text{tr}(\mathbf{A})$
3. $\text{tr}(\mathbf{A} + \mathbf{B}) = \text{tr}(\mathbf{A}) + \text{tr}(\mathbf{B})$
4. $\text{tr}(\mathbf{I}_k) = k$

Another useful property is that if $\mathbf{A}$ is $k \times r$ and $\mathbf{B}$ is $r \times k$, then $\text{tr}(\mathbf{AB}) = \text{tr}(\mathbf{BA})$. Unlike the previous results, this one is not obvious, so let's provide a quick proof:

$$\begin{aligned}
\text{tr}(\mathbf{AB}) &= \text{tr} \begin{bmatrix} a_1'b_1 & a_1'b_2 & \cdots & a_1'b_k \\ a_2'b_1 & a_2'b_2 & \cdots & a_2'b_s \\ \vdots & \vdots & \ddots & \vdots \\ a_k'b_1 & a_k'b_2 & \cdots & a_k'b_k \end{bmatrix} \\
&= \sum_{i=1}^{k} a_i'b_i \\
&= \sum_{i=1}^{k} b_i'a_i \\
&= \text{tr}(\mathbf{BA})
\end{aligned}$$

Rank and Inverse

H A.6

The rank of a $k \times r$ matrix (with $r \leq k$)

$$\mathbf{A} = \begin{bmatrix} a_1 & a_2 & \cdots & a_r \end{bmatrix}$$

written $\text{rank}(\mathbf{A})$, is the number of linearly independent columns of $\mathbf{A}$. $\mathbf{A}$ is said to have **full rank** if $\text{rank}(\mathbf{A}) = r$. Linear independence means that there is no non-zero $k \times 1$ vector c such that $\mathbf{A}'c = 0$. For example,

$$\text{rank} \begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix} = 1, \quad \text{rank} \begin{bmatrix} 1 & 2 \\ 2 & 3 \end{bmatrix} = 2$$

so that the second matrix has full rank but the first matrix does not (notice that the second column equals the first column times 2 so that they are not linearly independent; alternatively, you can notice that $\mathbf{A}'c = 0$ for $c = (2, -1)'$).

A square $k \times k$ matrix $\mathbf{A}$ is **nonsingular** if $\text{rank}(\mathbf{A}) = k$ (i.e., if it has full rank). If $\mathbf{A}$ is nonsingular, then it has an **inverse** $\mathbf{A}^{-1}$ that satisfies

$$\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}_k$$

For two non-singular matrices $\mathbf{A}$ and $\mathbf{C}$, another useful property is that

$$(\mathbf{A}^{-1})' = (\mathbf{A}')^{-1}$$

that is, for a nonsingular matrix, you can swap the order of transpose and inverse. Another useful property is that

$$(\mathbf{AC})^{-1} = \mathbf{C}^{-1}\mathbf{A}^{-1}$$

These properties are the ones that we'll use often though Appendix A.6 has several additional properties of nonsingular matrices that may be useful as a reference at some point.

Positive definite matrices

H A.10

A $k \times k$ symmetric matrix $\mathbf{A}$ is said to be **positive semi-definite** if $c'\mathbf{A}c \geq 0$ for any non-zero, $k \times 1$ vector c; this is often written $\mathbf{A} \geq 0$. $\mathbf{A}$ is said to be **positive definite** if $c'\mathbf{A}c > 0$ for any non-zero, $k \times 1$ vector c; this is often written $\mathbf{A} > 0$.

The textbook lists a number of properties of a positive definite matrix. One of these that we will use is that, if $\mathbf{A} > 0$, then $\mathbf{A}$ is nonsingular, $\mathbf{A}^{-1}$ exists, and $\mathbf{A}^{-1} > 0$.

Another is that, if $\mathbf{A}$ is positive definite, then we can find a square root matrix $\mathbf{A}^{1/2}$ such that $\mathbf{A} = \mathbf{A}^{1/2}\mathbf{A}^{1/2}$ where $\mathbf{A}^{1/2}$ is itself positive definite and symmetric.

Idempotent Matrices

H A.11

A $k \times k$ square matrix $\mathbf{A}$ is **idempotent** if $\mathbf{AA} = \mathbf{A}$.

Matrix Calculus

H A.20

For this section, let $x = (x_1, x_2, \ldots, x_k)'$ denote a $k \times 1$ vector and $g(x) : \mathbb{R}^k \to \mathbb{R}$. Now, let's consider taking the partial derivatives of the function g with respect to each variable in x; in

particular,

$$\frac{\partial g(x)}{\partial x} = \begin{pmatrix} \frac{\partial g(x)}{\partial x_1} \\ \frac{\partial g(x)}{\partial x_2} \\ \vdots \\ \frac{\partial g(x)}{\partial x_k} \end{pmatrix}$$

I will typically follow the convention of taking vector derivatives like the previous one "down" (as above), but it is also useful to have a notation for taking vector derivatives "across" as in

$$\frac{\partial g(x)}{\partial x'} = \begin{pmatrix} \frac{\partial g(x)}{\partial x_1} & \frac{\partial g(x)}{\partial x_2} & \cdots & \frac{\partial g(x)}{\partial x_k} \end{pmatrix}$$

Sometimes, we will also take second derivatives, which are given by

$$\frac{\partial^2 g(x)}{\partial x \partial x'} = \begin{pmatrix} \frac{\partial^2 g(x)}{\partial x_1^2} & \frac{\partial^2 g(x)}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 g(x)}{\partial x_1 \partial x_k} \\ \frac{\partial^2 g(x)}{\partial x_1 \partial x_2} & \frac{\partial^2 g(x)}{\partial x_2^2} & \cdots & \frac{\partial^2 g(x)}{\partial x_2 \partial x_k} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 g(x)}{\partial x_1 \partial x_k} & \frac{\partial^2 g(x)}{\partial x_2 \partial x_k} & \cdots & \frac{\partial^2 g(x)}{\partial x_k^2} \end{pmatrix}$$

Notice that this is a $k \times k$ matrix which is symmetric and arises from taking the partial derivatives "down" and then "across".

Here are some examples (we will consider the case where a is a $k \times 1$ vector and $\mathbf{A}$ is a $k \times k$ symmetric matrix):

- $\frac{\partial}{\partial x}(a'x) = \frac{\partial}{\partial x}(x'a) = a$
- $\frac{\partial}{\partial x'}(\mathbf{A}x)_{k\times 1} = \mathbf{A}$ and $\frac{\partial}{\partial x}(x'\mathbf{A})_{1\times k} = \mathbf{A}$
- $\frac{\partial}{\partial x}(x'\mathbf{A}x) = 2\mathbf{A}x$ and $\frac{\partial}{\partial x'}(x'\mathbf{A}x) = 2x'\mathbf{A}$
- $\frac{\partial^2}{\partial x \partial x'}(x'\mathbf{A}x) = 2\mathbf{A}$

In my view, a main takeaway from the above examples is that matrix calculus behaves very much like scalar calculus as long as you pay close attention to keeping the dimensions of the matrices straight (and also pay some attention to where to put transposes).

Vec Operator and Kronecker Product

H A.21

Write the $k \times r$ matrix $\mathbf{A} = \begin{bmatrix} a_1 & a_2 & \cdots & a_r \end{bmatrix}$. Then, the **vec** of $\mathbf{A}$ is defined as

$$\text{vec}(\mathbf{A}) = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix}$$

which is a $kr \times 1$ vector that stacks all the columns of $\mathbf{A}$ into one long column.

Next, write $\mathbf{A} = (a_{ij})$, then the **Kronecker product** of $\mathbf{A}$ and $\mathbf{B}$ is defined as (note that there are not restrictions on the dimensions of the matrices):

$$\mathbf{A} \otimes \mathbf{B} = \begin{bmatrix} a_{11}\mathbf{B} & a_{12}\mathbf{B} & \cdots & a_{1r}\mathbf{B} \\ a_{21}\mathbf{B} & a_{22}\mathbf{B} & \cdots & a_{2r}\mathbf{B} \\ \vdots & \vdots & \ddots & \vdots \\ a_{k1}\mathbf{B} & a_{k2}\mathbf{B} & \cdots & a_{kr}\mathbf{B} \end{bmatrix}$$

If the dimension of $\mathbf{B}$ is $m \times n$, then the dimension of $\mathbf{A} \otimes \mathbf{B}$ is $km \times rn$. The book provides some additional properties of Kronecker products.

Vector norms

H A.22

A **norm** is a function $\rho : \mathbb{R}^k \to \mathbb{R}$ that satisfies the following properties:

1. $\rho(ca) = c\rho(a)$ for any scalar c and $a \in \mathbb{R}^k$
2. $\rho(a + b) \leq \rho(a) + \rho(b)$. This is called the triangle inequality.
3. If $\rho(a) = 0$, then $a = 0$.

The three most common norm functions are

- The Euclidean norm: $||a|| = (a'a)^{1/2}$
- The 1-norm: $||a||_1 = \sum_{i=1}^{k} |a_i|$
- The sup-norm: $||a||_\infty = \max\{|a_1|, \ldots, |a_k|\}$

Conditional expectation and projection

H: 2.5

We will spend much time this semester thinking about **conditional expectations**, that is, $\mathbb{E}[Y|X = x]$. This is the mean of Y conditional on X taking the particular value x. The book sometimes uses the short-hand notation $m(x) := \mathbb{E}[Y|X = x]$. You can think of $\mathbb{E}[Y|X = x]$ as a

function — that is, when you plug in a new value of x, the value of the conditional expectation function can change. For example, if Y is a person's earnings and X is their years of education, the $\mathbb{E}[Y|X = 12]$ could differ (perhaps substantially) from, say, $\mathbb{E}[Y|X = 16]$. Sometimes it will be useful to view $\mathbb{E}[Y|X]$ as a function of the random variable X; in this case, $\mathbb{E}[Y|X]$ is itself random (because X is random). This is different from when you plug in a particular value of x, $\mathbb{E}[Y|X = x]$ is no longer random; it is equal to some number (though, in most cases, it is unlikely that we know the value of this sort of population quantity).

Law of iterated expectations

H: 2.7

One of the most useful tools (that you likely will be familiar with already) is the law of iterated expectations. We'll provide a simple version and a general version (both require the regularity condition that $\mathbb{E}|Y| < \infty$). The simple version is

$$\mathbb{E}[Y] = \mathbb{E}[\mathbb{E}[Y|X]]$$

In words, this says that the expectation of the conditional expectation is equal to the unconditional expectation. Here is an example. Continue to suppose that Y is a person's earnings and X is their years of education. $\mathbb{E}[Y|X = x]$ can vary arbitrarily for different values of x. But if you know $\mathbb{E}[Y|X = x]$ for all possible values of x, this will pin down the value of $\mathbb{E}[Y]$ (i.e., if you know mean earnings for all years of education, then you also should be able to recover the overall mean value of earnings). And, in particular, the law of iterated expectations says that the overall mean is equal to the mean of the conditional expectations (i.e., it puts more "weight'' on conditional expectations of relatively common values of X).

A more general version of the law of iterated expectations is the following: for any two random vectors X_1 and X_2,

$$\mathbb{E}[Y|X_1] = \mathbb{E}[\mathbb{E}[Y|X_1, X_2]|X_1]$$

The inside expectation is an expectation of Y conditional on X_1 and X_2, the outside expectation is over the distribution of X_2 conditional on X_1.

CEF Error

H: 2.8

We define the CEF error as the difference between Y and $\mathbb{E}[Y|X]$. That is,

$$e := Y - m(X)$$

Rearranging terms also implies the following expression that we will use often

$$Y = m(X) + e$$

That is, the actual outcome Y is equal to the CEF plus the CEF error. Notice that (besides regularity conditions that expectations exist), we are not making any assumptions here — we can essentially always write that Y is equal to its conditional expection plus CEF error.

A fundamental property of the CEF error is that $\mathbb{E}[e|X] = 0$. Let us show why this holds

$$\begin{aligned}
\mathbb{E}[e|X] &= \mathbb{E}[Y - m(X)|X] \\
&= \mathbb{E}[Y|X] - \mathbb{E}[m(X)|X] \\
&= m(X) - m(X) \\
&= 0
\end{aligned}$$

From the law of iterated expectations, this also implies that

$$\mathbb{E}[e] = \mathbb{E}[\underbrace{\mathbb{E}[e|X]}_{=0}] = 0$$

The condition that $\mathbb{E}[e|X] = 0$ is called **mean independence**. It is weaker than "full'' independence. To give an example, even if $\mathbb{E}[e|X] = 0$, it could be the case that $\mathbb{E}[e^2|X]$ varies with X which would imply that e and X are not independent.

We can also define the variance of the CEF error as

$$\sigma^2 = \text{var}(e) = \mathbb{E}[(e - \mathbb{E}[e])^2] = \mathbb{E}[e^2]$$

σ^2 measures the amount of variation in Y that is not accounted for by $\mathbb{E}[Y|X]$.

Best Predictor

H: 2.11

Next, we will show that the conditional expectation function is the "best'' predictor of Y given X. We can write any predictor as a function $g(X)$; i.e., a function that takes values of X and makes predictions about what the outcome will be. In order to evaluate how well a predictor makes

predictions, we need some criteria. The most common criteria is **mean squared prediction error**. This is given by

$$\mathbb{E}[(Y - g(X))^2]$$

The squared difference between Y and its prediction $g(X)$ is a measure of the distance between Y and $g(X)$ — in particular, it is always non-negative and gets larger as when Y and $g(X)$ are further away from each other. The outside expectation averages this distance over the distribution of Y and X.

Next, we show that $m(X)$ minimizes mean squared prediction error. To this end, notice that

$$\begin{aligned}
\mathbb{E}[(Y - g(X))^2] &= \mathbb{E}\Big[\big((Y - m(X)) + (m(X) - g(X))\big)^2\Big] \\
&= \mathbb{E}\Big[\big(e + (m(X) - g(X))\big)^2\Big] \\
&= \mathbb{E}[e^2] + 2\mathbb{E}[e(m(X) - g(X))] + \mathbb{E}[(m(X) - g(X))^2]
\end{aligned}$$

where the first equality holds by adding and subtracting $m(X)$, the second equality holds by the definition of e, and the last equality holds by squaring the main term and pushing the expectation through the sum. Now, let's consider each term. First, $\mathbb{E}[e^2] = \sigma^2$ which does not depend on our prediction function $g(X)$. Next, consider the middle term

$$\begin{aligned}
\mathbb{E}[e(m(X) - g(X))] &= \mathbb{E}[(m(X) - g(X))\mathbb{E}[e|X]] \\
&= 0
\end{aligned}$$

where the first equality holds by the law of iterated expectations and the second equality holds because $\mathbb{E}[e|X] = 0$. Finally, the third term is minimized at 0 by setting $g(X) = m(X)$. This implies that setting $g(X) = m(X)$ minimizes mean squared prediction error.

Regression Derivatives

H 2.14

Following the textbook, we'll use the shorthand notation $m(x) := \mathbb{E}[Y|X = x]$. We will often be interested in the **regression derivative**. An example of a regression derivative is

$$\frac{\partial\, \mathbb{E}[Y|X = x]}{\partial\, x_1}$$

which holds when x_1 is continuously distributed. This derivative should be interpreted as how much Y changes, on average, when x_1 increases by one unit holding the other regressors constant.

You can also define a regression derivative when X_1 is discrete. For example, suppose that X_1 is binary (so it only takes the value 0 or 1), then the regression derivative is given by

$$\mathbb{E}[Y|X_1 = 1, X_2 = x_2, \ldots, X_k = x_k] - \mathbb{E}[Y|X_1 = 0, X_2 = x_2, \ldots X_k = x_k]$$

You could similarly define a regression derivative for the case where X_1 was discrete but took more possible values.

In order to unify notation, we write

$$\nabla_1 m(x) := \begin{cases} \frac{\partial\, \mathbb{E}[Y|X=x]}{\partial x_1} & \text{if } x_1 \text{ is continuous} \\ \mathbb{E}[Y|X_1 = 1, X_2 = x_2, \ldots, X_k = x_k] - \mathbb{E}[Y|X_1 = 0, X_2 = x_2, \ldots X_k = x_k] & \text{if } x_1 \text{ is binary} \end{cases}$$

There is nothing unique about defining partial effects for just X_1, and we can likewise define partial effects for $X_2, \ldots, X_k$, for example, $\nabla_2 m(x)$ is the partial effect of X_2.

The regression derivatives above are also sometimes called the "partial effect" of x_1 or the "marginal effect" of x_1.

Some Comments

- First, partial effects hold other regressors constant. But they do not hold other variables that are not in the model constant.
- Second, you should notice that $\nabla_1 m(x)$ is a function of x. If you plug in different values of x, then the value of this function could change. For example, if you take X_1 to be a binary variable indicating whether or not an individual attended college, Y to be their earnings, and X_2 to be a person's age, you could imagine that the partial effect of college differs depending on a person's age.
- Third, partial effects are really about averages rather than individual-level effects. Continuing the example of the return to going to college – you can easily imagine that, holding age constant, the effect of going to college on a person's earnings may vary (perhaps tremendously across different people). The regression derivative averages over all of these individual-level effects while holding age constant.

Linear CEF

H: 2.15

An important special case is when $\mathbb{E}[Y|X] = X'\beta$; that is, when the CEF is linear. Importantly, unlike the previous generic discussion of CEFs, in many cases, it may be a strong assumption to impose a linear CEF.

Notation:
I'll follow the convention in the book by writing

$$\mathbb{E}[Y|X = x] = x_1\beta_1 + x_2\beta_2 + \cdots + x_{k-1}\beta_{k-1} + \beta_k$$

so that the "intercept" is in the last position. More specifically,

$$X = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_{k-1} \\ 1 \end{pmatrix} \qquad \beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}$$

so that X and β are both $k \times 1$ vectors. And Y, the outcome, is a scalar.
When referring to particular observations, I'll use the notation

$$X_i = \begin{pmatrix} X_{i1} \\ X_{i2} \\ \vdots \\ X_{ik} \end{pmatrix} \qquad \text{and} \qquad Y_i$$

where X_i is $k \times 1$ and Y_i is a scalar.

In this case, the regression derivative is given by $\nabla_1 m(x) = \beta_1$. This is a major simplification from the general version of $\nabla_1 m(x)$ that we discussed above. It has pros and cons. If the CEF really is linear, then (i) it will be much easier to estimate $m(x) = x'\beta$ (and, hence, $\nabla_1 m(x) = \beta_1$) than in the more general case. Additionally, given an estimate of β_1, it is easy to report/fully summary the partial effect of X_1 on $\mathbb{E}[Y|X]$. The main disadvantage is that, often, we may not have a good reason to think/impose that the CEF is actually linear.

Linear CEF with Nonlinear Effects

H: 2.16-2.17

One way to add some additional "flexibility" to the linear CEF model that we have been discussing is to add higher order terms and interactions into the model. For example,

$$m(x_1, x_2) = x_1\beta_1 + x_2\beta_2 + x_1^2\beta_3 + x_2^2\beta_4 + x_1x_2\beta_5 + \beta_6$$

This would still be considered a linear CEF as it is still *linear in the parameters* though it allows for

nonlinear effects of the regressors. In particular, notice that

$$\nabla_1 m(x_1, x_2) = \beta_1 + 2x_1\beta_3 + x_2\beta_5$$

which can depend on the values of x_1 and x_2. To be clear, this is still less flexible than the general version of $\nabla_1 m(x)$ that we discussed earlier, but it does allow for partial effects to depend on the values of different regressors.

One important special case of the nonlinear effects CEF mentioned above is when, say, x_2 is a binary variable. For simplicity, let's consider a slight modification of the previous model such that

$$m(x_1, x_2, x_3) = x_1\beta_1 + x_2\beta_2 + x_1x_2\beta_3 + x_3\beta_4$$

Suppose also that Y is a person's income, X_1 is a person's years of education, X_2 is whether or not a person is married, and X_3 is how old a person is. In this case,

$$\nabla_1(x_1, x_2, x_3) = x_1 + x_2\beta_3$$

which depends on whether $x_2 = 1$ or 0; that is $\nabla_1(x_1, x_2 = 1, x_3) = x_1 + \beta_3$ while $\nabla_1(x_1, x_2 = 0, x_3) = x_1$, so that (in the example), the partial effect of education can depend on whether or not a person is married. Further, rearranging, we have that

$$\beta_3 = \nabla_1(x_1, x_2 = 1, x_3) - \nabla_1(x_1, x_2 = 0, x_3)$$

so that β_3 is equal to the difference between partial effects of education for married people relative to unmarried people.

Please read H: 2.16-2.17 for additional discussions on nonlinear effects in linear CEF models. I think much of this will be review from your undergrad econometrics, but these are things that you should make sure that you fully understand. Relatedly, please read the discussion in H: 2.4 about taking the logarithm of positive random variables (which is relatively common in economics).

Best Linear Predictor

H: 2.18

In many cases, we may be hesitant (or not have a good reason to believe) that the CEF is actually linear. Even in this case, we can still "run a regression" of Y on X.

In this section, we'll consider the best *linear* predictor of Y given X. A linear predictor for Y is a function $X'b$ for some $b \in \mathbb{R}^k$. We will choose the best possible value of b. For this section, we will make the following regularity assumptions (Assumption 2.1 in the textbook)

1. $\mathbb{E}[Y^2] < \infty$
2. $\mathbb{E}||X||^2 < \infty$

3. $\mathbb{E}[XX']$ is positive definite

where $||x|| = (x'x)^{1/2}$ is the Euclidean length of the vector x. The first two assumptions imply that Y and X have finite second moments (and, therefore, finite means, variances, and covariances). As earlier, we will try to minimize mean squared prediction error; that is,

$$\beta = \underset{b}{\text{argmin}}\ S(b)$$

where we define

$$S(b) = \mathbb{E}[(Y - X'b)^2]$$

We can use tools from calculus to solve this. As a first step, it is helpful to notice that

$$\mathbb{E}[(Y - X'b)^2] = \mathbb{E}[Y^2] - 2\beta'\mathbb{E}[XY] + \beta'\mathbb{E}[XX']\beta$$

As a side-comment, notice that the function that we are minimizing is a scalar (we will likely see other parameters/estimators this semester that minimize or maximize some objective function — these objective functions always return a scalar). Second, the expansion in the previous equation may not be obvious; in general, I think you can do quite well at linear algebra by keeping track of the dimensions of the terms that you are working with. In particular, you would know that you were making a mistake if the dimension of any term above were not scalar. Finally, we combined the two terms $\mathbb{E}[YX]'\beta$ and $\beta'\mathbb{E}[XY]$ for the middle term above — these are equal to the transpose of each other and since they are both scalars, they are exactly equal to each other.

Next, let's take the derivative of the previous equation, set it equal to 0 and solve for β. Before we do this, let's be clear about exactly what we are doing. We are taking the derivative of a function $S(b) : \mathbb{R}^k \to \mathbb{R}$ (that is a function that takes in a k dimensional vector and returns a scalar). And, in particular, this vector derivative is given by

$$\frac{\partial S(b)}{\partial b} = \begin{pmatrix} \frac{\partial S(b)}{\partial b_1} \\ \vdots \\ \frac{\partial S(b)}{\partial b_k} \end{pmatrix}$$

which is a $k \times 1$ vector. As a side-comment, I follow the convention (which is also used in the textbook) of taking first derivatives of scalar-valued functions with respect to a vector "down" (so that the first derivative is $k \times 1$ vector rather than a $1 \times k$ vector) and (if needed) second derivatives "across' ' (so that the second derivative would result in a $k \times k$ matrix). For more details about taking derivatives with respect to a vector, see the discussion on p.38 of the textbook and in Appendix A.20 (I consider this to be review material, but it is definitely material you should know).

Returning to our present problem, notice that

$$0 = \frac{\partial S(b)}{\partial b}\Big|_{b=\beta} = -2\mathbb{E}[XY] + 2\mathbb{E}[XX']\beta$$

In terms of mechanics, this derivative is just like the scalar case (the first term is a linear term and the second term is a quadratic term) except you just need to make sure that you get a $k \times 1$ vector (rather than, especially, a $1 \times k$ vector).

We can immediately solve the previous equation for β. The regularity conditions above imply that all of the moments here exist and that the matrix $\mathbb{E}[XX']$ is invertible.

$$\beta = \mathbb{E}[XX']^{-1}\mathbb{E}[XY]$$

In this context, we'll refer to β as the **linear projection coefficient** and $X'\beta$ as the **best linear predictor**

We can also define the **projection error**

$$e = Y - X'\beta$$

which is the difference between the actual value of Y and the best linear predictor of Y given X.

An important property of the projection error is that $\mathbb{E}[Xe] = 0$. To see this, notice that

$$\begin{aligned}
\mathbb{E}[Xe] &= \mathbb{E}[X(Y - X'\beta)] \\
&= \mathbb{E}[XY] - \mathbb{E}[XX']\beta \\
&= \mathbb{E}[XY] - \mathbb{E}[XX']\mathbb{E}[XX']^{-1}\mathbb{E}[XY] \\
&= 0
\end{aligned}$$

Notice that $\mathbb{E}[Xe]$ is a $k \times 1$ vector. When X includes an intercept (so that $X_k = 1$), this implies that $\mathbb{E}[e] = 0$.

Best linear approximation

H 2.25

Next, we show another interesting property/interpretation for the linear projection coefficient β. Suppose that we are interested in learning about the conditional expectation function $m(x) = \mathbb{E}[Y|X = x]$, but we have no reason to suppose that it is linear. Therefore, we might be interested in trying to construct the best linear approximation to $m(x)$. That is, let's consider choosing β in the following way

$$\beta = \underset{b}{\operatorname{argmin}}\ \mathbb{E}[(m(X) - X'b)^2]$$

Following roughly the same strategy as earlier, notice that

$$\mathbb{E}[(m(X) - X'b)^2] = \mathbb{E}[m(X)^2] - 2\beta'\mathbb{E}[XY] + \beta'\mathbb{E}[XX']\beta$$

Taking the derivative and setting equal to 0 implies that

$$0 = -2\mathbb{E}[Xm(X)] + 2\mathbb{E}[XX']\beta$$

which further implies that

$$\begin{aligned}\beta &= \mathbb{E}[XX']^{-1}\mathbb{E}[Xm(X)] \\ &= \mathbb{E}[XX']^{-1}\mathbb{E}[X\mathbb{E}[Y|X]] \\ &= \mathbb{E}[XX']^{-1}\mathbb{E}[XY]\end{aligned}$$

where the second equality holds by the definition of $m(X)$ and the last equality holds by the law of iterated expectations.

This is exactly the same expression for β as we derived earlier under the motivation of best linear predictor. This implies that $X'\beta$ can additionally be interpreted as the best linear approximation to the underlying CEF — even if the CEF is nonlinear. This is a nice property for the linear projection model to have. That being said, even the best linear approximation to a nonlinear CEF can sometimes be quite poor. See Section 2.28 for an example and some discussion.

Omitted variable bias

H: 2.24

To conclude this section, let's briefly talk about omitted variable bias. Let's partition X as follows $X = (X_1', X_2')'$ and likewise partition β into $\beta = (\beta_1', \beta_2')$. Suppose that we are interested in β_1 from the linear projection of Y onto X_1 and X_2:

$$Y = X_1'\beta_1 + X_2'\beta_2 + e \tag{1}$$

Since this is a linear projection, it implies that $\mathbb{E}[Xe] = 0$.

However, let's suppose that X_2 is not observed, so that it is infeasible to run a regression of Y on X_1 and X_2. In this section, we consider properties of the following **short regression**

$$Y = X_1'\gamma_1 + u$$

which is the linear projection of Y on X_1 only. Since this is a linear projection, we also have that

$\mathbb{E}[X_1 u] = 0$ and that

$$\begin{aligned}
\gamma_1 &= \mathbb{E}[X_1 X_1']^{-1}\mathbb{E}[X_1 Y] \\
&= \mathbb{E}[X_1 X_1']^{-1}\mathbb{E}[X_1(X_1\beta_1 + X_2\beta_2 + e)] \\
&= \beta_1 + \mathbb{E}[X_1 X_1']^{-1}\mathbb{E}[X_1 X_2]\beta_2 \\
&= \beta_1 + \Gamma_{12}\beta_2
\end{aligned}$$

where the first equality holds by the definition of linear projection of Y on X_1, the second equality holds by substituting for Y, the third equality combines and cancels terms and also holds because $\mathbb{E}[X_1 e] = 0$ (since $\mathbb{E}[Xe] = 0$), and the last equality holds because we define $\Gamma_{12} = \mathbb{E}[X_1 X_1']^{-1}\mathbb{E}[X_1 X_2]$. Notice that Γ_{12} is the coefficient from the linear projection of X_2 on X_1.

Importantly, the previous expression implies that γ_1 is not generally equal to β_1; that is, in general, we are not able to recover the parameter of interest β_1 from the feasible regression of Y on X_1. This is probably not surprising — otherwise, our lives would be much easier! The difference between γ_1 and β_1 is called **omitted variable bias** and is a very important concern in many applications.

The only case where $\gamma_1 = \beta_1$ is when $\Gamma_{12}\beta_2 = 0$ which can happen when either $\Gamma_{12} = 0$ or $\beta_2 = 0$. $\Gamma_{12} = 0$ if $\mathbb{E}[X_1 X_2] = 0$ which would be the case if X_1 and X_2 are uncorrelated. $\beta_2 = 0$ occurs when the coefficient on X_2 in Equation 1 is equal to 0. In words, the cases where you can recover β_1 while only using the short regression are (i) if the omitted variables are uncorrelated with the included variables or (ii) if the omitted variables have no effect on the outcome.

Side-Comment: There are a number of cases where you might be able to figure out the sign of the omitted variable bias. The textbook gives the following simple example. Consider the case where Y is a person's earnings, X_1 is a person's years of education, and X_2 is a person's "ability", and where you are interested in β_1 (the coefficient on years of education). However, suppose that ability is not observed. In this case, it might be reasonable to suppose that $\beta_2 > 0$ (i.e., that, conditional on years of education, individuals with higher ability tend to have higher earnings) and that $\Gamma_{12} > 0$ (i.e., that higher ability is positively correlated with more education). Under these conditions, it would be the case that $\gamma_1 > \beta_1$.
This discussion suggests that a regression that only includes years of education would overestimate the effect of years of education relative to a model that included both education and ability. This sort of argument is quite common in applied work — something like: "even though we are not able to control for some important variable, it's correlation with the observed variable of interest and likely sign in the long regression indicate that the estimate of our coefficient of interest is likely a lower (or upper) bound."