X is a random variable whose value we want to estimate. We generally have information available on which to base our estimate, typically other random variables whose values we observe. Thus our estimate is in general a function of other random variables and is therefore itself a random variable. In constructing our estimate we might be allowed to use only certain functions, e.g. linear or affine, of the observed random variables. In any event, if we denote our estimate $\widehat{X}$, the mean squared error of our estimate is

$$\mathbb{E}\left((X-\widehat{X})^2\right)\ , \tag{1}$$

and the goal of any minimum mean squared estimation error (MMSE) estimation scheme is to find $\widehat{X}$ that minimizes this quantity over all estimates that satisfy our information and functional-implementation constraints. In what follows, I'll consider a few different sets of constraints and see how they affect the solution to the problem and the resulting minimum value of (1).

We can always decompose the mean squared error (1) as

$$\mathbb{E}\left((X-\widehat{X})^2\right) = \mathrm{Var}(X-\widehat{X}) + (\mathbb{E}(X-\widehat{X}))^2\ . \tag{2}$$

The first term is the variance of the estimation error $\widetilde{X} = X - \widehat{X}$ and the second is the square of the expected value of $\widetilde{X}$. That expected value is known as the bias of the estimate $\widehat{X}$, and we call $\widehat{X}$ an unbiased estimate when $\mathbb{E}(\widetilde{X}) = 0$. All the MMSE estimators treated in this handout are unbiased and at the same time minimize the variance of the estimation error over all estimators satisfying the information and implementation constraints. It makes sense that a good MMSE estimator will have zero expected estimation error and small estimation-error variance, which together imply that the estimation error's probability mass is concentrated heavily around zero, but implementation constraints sometimes preclude simultaneous minimization of bias and estimation-error variance.

The most restrictive information constraint arises in the trivial situation where we have no extra information at all. In that case, our estimate $\widehat{X}$ must be a constant. Our MMSE estimation problem demands that we choose a number $\widehat{b}$ to minimize

$$\mathbb{E}\left((X-b)^2\right)$$

over all real numbers b. Perhaps it's obvious to you that the optimal choice of b is $\widehat{b} = \mathbb{E}(X)$, and indeed that's the case, but let's prove it. For any choice of b, equation (2) reads

$$\begin{aligned}\mathbb{E}\left((X-b)^2\right) &= \mathrm{Var}(X-b) + (\mathbb{E}(X-b))^2 \\ &= \mathrm{Var}(X) + (\mathbb{E}(X)-b)^2\ ,\end{aligned}$$

Thus our mean squared error is the sum of $\mathrm{Var}(X)$, a term we can do nothing about, and $(\mathbb{E}(X)-b)^2$, which we can zero out by setting $b = \mathbb{E}(X)$. Summarize as follows.

Constant MMSE Estimation: The constant $\widehat{X}_{\mathrm{opt}}$ that minimizes (1) over all constants $\widehat{X}$ is $\widehat{X}_{\mathrm{opt}} = \mathbb{E}(X)$.

Now suppose we have access to one other random variable Y, and we're allowed to choose as our estimate $\widehat{X} = h(Y)$, where $h(Y)$ is any function of Y. In class, we showed that the minimizing choice of $\widehat{X}$ in (1) is $\widehat{X}_{\mathrm{opt}} = \mathbb{E}(X \mid Y)$. Recall how the argument

went. For any function $h(Y)$, we have

$$\begin{aligned}\mathbb{E}\left((X-h(Y))^2\right) &= \mathbb{E}\left((X-\mathbb{E}(X\mid Y)+\mathbb{E}(X\mid Y)-h(Y))^2\right)\\ &= \mathbb{E}\left((X-\mathbb{E}(X\mid Y)^2\right)+2\,\mathbb{E}((X-\mathbb{E}(X\mid Y))\,(\mathbb{E}(X\mid Y)-h(Y)))\\ &\qquad +\mathbb{E}\left((\mathbb{E}(X\mid Y)-h(Y))^2\right)\ .\end{aligned}$$

The middle term is zero because, by the law of iterated expectations,

$$\begin{aligned}\mathbb{E}((X-\mathbb{E}(X\mid Y))\,(\mathbb{E}(X\mid Y)-h(Y))) &= \mathbb{E}(\,\mathbb{E}(X-\mathbb{E}(X\mid Y))\,(\mathbb{E}(X\mid Y)-h(Y))\mid Y)\\ &= \mathbb{E}(\,(\mathbb{E}(X\mid Y)-h(Y)\,)\,\mathbb{E}(X-\mathbb{E}(X\mid Y)\mid Y)\,)\\ &= \mathbb{E}(\,\mathbb{E}(X\mid Y)-h(Y)\,)\times 0)\\ &= 0\ ,\end{aligned}$$

where the second line holds because $\mathbb{E}(X\mid Y)-h(Y)$ is a function of Y and the third because $\mathbb{E}(\mathbb{E}(X\mid Y)\mid Y)=\mathbb{E}(X\mid Y)$. Since the middle term is zero, we have

$$\mathbb{E}\left((X-h(Y))^2\right)=\mathbb{E}\left((X-\mathbb{E}(X\mid Y)^2\right)++\mathbb{E}\left((\mathbb{E}(X\mid Y)-h(Y))^2\right)\ . \tag{3}$$

This makes it obvious that we must choose $h(Y)=\mathbb{E}(X\mid Y)$ to minimize the mean squared error. That choice zeroes the second term in (3) and makes the mean squared error equal to $\mathbb{E}\left((X-\mathbb{E}(X\mid Y))^2\right)$, a term we can't affect by choice of h. Here's the summary of this important result.

General MMSE Estimation: The random variable $\widehat{X}_{\rm opt}$ that minimizes (1) over all functions of Y is $\widehat{X}_{\rm opt}=\mathbb{E}(X\mid Y)$.

One noteworthy feature of the general MMSE is that it is unbiased, i.e.

$$\mathbb{E}\left(\widehat{X}_{\rm opt}\right)=\mathbb{E}(\mathbb{E}(X\mid Y))=\mathbb{E}(X)\ ,$$

which is equivalent to

$$\mathbb{E}(\widetilde{X})=0\ ,$$

where $\widetilde{X}=X-\widehat{X}_{\rm opt}$ is the optimal estimation error. Furthermore, the optimal estimation error is uncorrelated with Y and in fact with any function $h(Y)$. That's because

$$\begin{aligned}\mathrm{Cov}\left(\widetilde{X}_{\rm opt},h(Y)\right) &= \mathbb{E}\left(\widetilde{X}_{\rm opt}\left(h(Y)-\widetilde{X}_{\rm opt}\right)\right)\\ &= \mathbb{E}\left(\mathbb{E}\left(\widetilde{X}_{\rm opt}\left(h(Y)-\widetilde{X}_{\rm opt}\right)\mid Y\right)\right)\\ &= \mathbb{E}\left(\left(h(Y)-\widetilde{X}_{\rm opt}\right)\,\mathbb{E}\left(\widetilde{X}_{\rm opt}\mid Y\right)\right)\\ &= 0\ ,\end{aligned}$$

where the first line holds because $\widetilde{X}$ has zero mean, the second by the law of iterated expectations, the third because $h(Y)-\mathbb{E}(h(Y))$ is a function of Y, and the fourth because the inner conditional expectation on the third line is zero.

An identical argument leads to an extended General MMSE Estimation result that applies when multiple random variables are available to estimate X in terms of.

Extended General MMSE Estimation: The random variable $\widehat{X}_{\rm opt}$ that minimizes (1) over all functions of Y_1, Y_2, ... , Y_n is $\widehat{X}_{\rm opt}=\mathbb{E}(X\mid Y_1,Y_2,\ldots,Y_n)$.

We find in this case that the estimation error $\widetilde{X} = X - \widehat{X}_{\rm opt}$ has zero mean as before and is uncorrelated with any function $h(Y_1, Y_2, \ldots, Y_n)$.

Suppose now that we want to estimate X in terms of Y using a linear — strictly speaking, affine — estimator of the form

$$\widehat{X} = aY + b \; .$$

The mean square error for such an estimator, using (2), is

$$\mathbb{E}\left((X - aY - b)^2\right) = \mathrm{Var}(X - aY - b) + \left(\mathbb{E}(X - aY - b)\right)^2 \; .$$

Regarding the variance,

$$\begin{aligned} \mathrm{Var}(X - aY - b) &= \mathrm{Var}(X - aY) \\ &= \mathrm{Var}(X) - 2a\mathrm{Cov}(X, Y) + a^2\mathrm{Var}(Y) \; , \end{aligned}$$

The value of a minimizing this quantity, which you can calculate by taking the derivative, is

$$\widehat{a} = \frac{\mathrm{Cov}(X, Y)}{\mathrm{Var}(Y)} \; ,$$

and by choosing

$$b = \widehat{b} = \mathbb{E}(X) - \widehat{a}\,\mathbb{E}(Y)$$

we can zero out the bias term in the mean square error expression. Summarize as follows.

Linear MMSE Estimation I: The random variable $\widehat{X}_{\rm opt}$ that minimizes (1) over all linear estimators $\widehat{X} = aY + b$ is

$$\begin{aligned} \widehat{X}_{\rm opt} &= \widehat{a}Y + \widehat{b} \\ &= \frac{\mathrm{Cov}(X, Y)}{\mathrm{Var}(Y)} Y + \left(\mathbb{E}(X) - \frac{\mathrm{Cov}(X, Y)}{\mathrm{Var}(Y)}\mathbb{E}(Y)\right) \\ &= \frac{\mathrm{Cov}(X, Y)}{\mathrm{Var}(Y)} (Y - \mathbb{E}(Y)) + \mathbb{E}(X) \; . \end{aligned}$$

Again in this case the optimal estimator is unbiased in the sense that the expected value of $\widehat{X}_{\rm opt}$ is the same as the expected value of X. Furthermore, the optimal estimation error $\widetilde{X} = X - \widehat{X}_{\rm opt}$ is again uncorrelated with Y, and in fact with any linear function $aY + b$. That's because for any a and b we have

$$\begin{aligned} \mathrm{Cov}(\widetilde{X}, aY + b) &= \mathrm{Cov}(X - \widehat{a}Y - \widehat{b}, aY + b) \\ &= a\,\mathrm{Cov}(X - \widehat{a}Y, Y) \\ &= a\left(\mathrm{Cov}(X, Y) - \widehat{a}\mathrm{Var}(Y)\right) \\ &= 0 \; , \end{aligned}$$

where the last line holds because $\widehat{a} = \mathrm{Cov}(X, Y)/\mathrm{Var}(Y)$.

Suppose now that we want to estimate X linearly in terms of several random variables. First let's consider the case where $W_1, \ldots, W_n$ are uncorrelated random variables and we want to minimize (1) using a linear estimator of the form

$$\widehat{X} = \sum_{k=1}^{n} a_k W_k + b \; .$$

Conveniently, it turns out that the optimal coefficients are

$$\widehat{a}_k = \frac{\mathrm{Cov}(X, W_k)}{\mathrm{Var}(W_k)} \quad 1 \le k \le n$$

and

$$\widehat{b} = \mathbb{E}(X) = \sum_{k=1}^{n} \widehat{a}_k \mathbb{E}(W_k) \ .$$

Thus the optimal estimator is

$$\begin{aligned} \widehat{X}_{\text{opt}} &= \sum_{k=1}^{n} \widehat{a}_k W_k + \widehat{b} \\ &= \left(\sum_{k=1}^{n} \frac{\text{Cov}(X, W_k)}{\text{Var}(W_k)} (W_k - \mathbb{E}(W_k)) \right) + \mathbb{E}(X) \ . \end{aligned}$$

You can prove this in any number of ways, for example by induction, but here's a simple way to verify that these coefficients do the job. For any a_k, $1 \le k \le n$, and b, we have by equation (2)

$$\mathbb{E}\left((X - \sum_{k=1}^{n} a_k W_k - b)^2 \right) = \text{Var}\left(X - \sum_{k=1}^{n} a_k W_k - b \right) + \left(\mathbb{E}\left(X - \sum_{k=1}^{n} a_k W_k - b \right) \right)^2 \ .$$

Take this decomposition of the mean squared error and proceed as follows.

- Rewrite the variance term as

  $$\text{Var}\left(X - Q + Q - \sum_{k=1}^{n} a_k W_k \right) \ ,$$

  where Q is shorthand for $\widehat{X}_{\text{opt}}$ above. The b disappears because it doesn't affect the variance.
- Note that $\mathbb{E}(Q) = \mathbb{E}(X)$, so $\mathbb{E}(X - Q) = 0$.
- Prove that $X - Q$ is uncorrelated with W_m for $1 \le m \le n$. This is because

  $$\begin{aligned} \text{Cov}(X - Q, W_m) &= \text{Cov}((X - \widehat{a}_m W_m, W_m) \\ &= \text{Cov}(X, W_m) - \widehat{a}_m \text{Var}(W_m) \\ &= 0 \ , \end{aligned}$$

  where the first line holds because the W_k are uncorrelated, so the only term in Q not uncorrelated with W_m is the $k = m$ term.
- Since $X - Q$ is uncorrelated with all the W_m, the variance term decomposes as

  $$\text{Var}(X - Q) + \text{Var}\left(\sum_{k=1}^{n} (\widehat{a}_k - a_k) W_k \right) \ .$$

  The first term we can't affect by choice of a_k, whereas we can zero the second term by setting $a_k = \widehat{a}_k$ for all k, so we do that.
- Having chosen $a_k = \widehat{a}_k$ for all k, we can zero the bias term by setting

  $$b = \widehat{b} = \mathbb{E}(X) - \sum_{k=1}^{n} \widehat{a}_k \mathbb{E}(W_k) \ .$$

Summarize as follows.

Linear MMSE Estimation II: If W_k, $1 \le k \le n$, are uncorrelated random variables, the random variable $\widehat{X}_{\text{opt}}$ that minimizes (1) over all linear estimators $\widehat{X} = \sum_{k=1}^{n} a_k W_k + b$

is

$$\begin{aligned}\widehat{X}_{\text{opt}} &= \sum_{k=1}^{n} \widehat{a}_k W_k + \widehat{b} \\ &= \sum_{k=1}^{n} \frac{\text{Cov}(X, W_k)}{\text{Var}(W_k)} W_k + \left(\mathbb{E}(X) - \sum_{k=1}^{n} \frac{\text{Cov}(X, W_k)}{\text{Var}(W_k)} \mathbb{E}(W_k) \right) \\ &= = \sum_{k=1}^{n} \frac{\text{Cov}(X, W_k)}{\text{Var}(W_k)} (W_k - \mathbb{E}(W_k)) + \mathbb{E}(X) \ .\end{aligned}$$

Again in this case the optimal estimator is unbiased and the optimal estimation error $\widetilde{X} = X - \widehat{X}_{\text{opt}}$ is uncorrelated with any linear combination of $W_1, \ldots, w_n$.

Suppose finally that we want to estimate X linearly in terms of several random variables $Y_1, \ldots, Y_n$ that aren't necessarily uncorrelated. As usual, we want to minimize (1) using a linear estimator of the form

$$\widehat{X} = \sum_{k=1}^{n} a_k Y_k + b \ .$$

One effective approach to the problem begins by generating uncorrelated random variables $W_1, \ldots, W_n$ that "span the same space" as $Y_1, \ldots Y_n$. You can do this using a procedure much like Gram-Schmidt orthogonalization of vectors or by diagonalizing the covariance matrix Σ of the Y_k, whose (i, j)-entry is $\text{Cov}(Y_i, Y_j)$ for $1 \leq i, j \leq n$. Having generated the W_k, you estimate X as above in terms of the uncorrelated random variables $W_1, \ldots, W_n$ and then rewrite that estimate in terms of $Y_1, \ldots, Y_n$. The math works out so that the following result holds.

Linear MMSE Estimation III: If $Y_1, \ldots, Y_n$ are not necessarily uncorrelated random variables, the random variable $\widehat{X}_{\text{opt}}$ that minimizes (1) over all linear estimators $\widehat{X} = \sum_{k=1}^{n} a_k Y_k + b$ is

$$\widehat{X}_{\text{opt}} = (\text{Cov}(X, \mathbf{Y}))^T \ \Sigma^{-1} \ (\mathbf{Y} - \mathbb{E}(\mathbf{Y})) \ + \ \mathbb{E}(X) \ ,$$

where $\mathbf{Y}$ is the column n-vector whose kth entry is Y_k, and $\text{Cov}(X, \mathbf{Y})$ is the column n-vector whose kth entry is $\text{Cov}(X, Y_k)$, and Σ is the $(n \times n)$ matrix whose (i, j) entry is $\text{Cov}\,(Y_i, Y_j)$, i.e. $\Sigma = \mathbb{E}\left(\mathbf{Y}\mathbf{Y}^T\right)$.