Estimation: Regression and Least Squares

This note explains how to use observations to estimate unobserved random variables.

Introduction

How would you guess the value of some random variable Y, say the weight of the next person to call you on the phone? One sensible guess, if you don't know anything else, is the expected value $E[Y]$ of Y. This is a good guess in a precise sense: the mean value $E[Y]$ is the constant a that minimizes the *mean squared error* $E[(Y-a)^2]$. To see this, note that, for any a, one has

$$\begin{aligned} E[(Y-a)^2] &= E[(Y-E[Y]+E[Y]-a)^2] \\ &= E((Y-E[Y])^2]+2E[(Y-E[Y])(E[Y]-a)]+(E[Y]-a)^2 \\ &= E((Y-E[Y])^2]+0+(E[Y]-a)^2 \\ &\geq E[(Y-E[Y])^2]. \end{aligned}$$

In the derivation above, the first identity is obvious, the second follows from expanding the square of the sum of two terms and using linearity of expectation, and the third follows from linearity and the fact that $E[(Y-E[Y])]=0$.

Thus, the derivation shows that $E[(Y-a)^2] \geq E[(Y-E[Y])^2]$ for any choice of a, which confirms that the value of a that minimizes the mean squared error is indeed $a=E[Y]$.

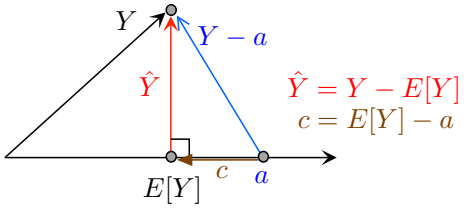
One graphical way of seeing the next-to-last identity is to observe that it looks like Pythagoras' theorem where the square of the 'hypothenuse' $Y-a$ is the sum of the squares of the two 'sides' $Y-E[Y]$ and $E[Y]-a$ of some right triangle. The triangle is right because the two sides are orthogonal in the sense that the mean value of their product is equal to zero. Figure 1 illustrates the connection between the algebra and Pythagoras' theorem.

Now, imagine that the person who calls you states his height X. Clearly that information should help you improve your guess about that person's weight. Indeed, a taller person is generally heavier than a short one. Figure 2 shows the height and weight of a number of people and confirms that the weight tends to increase with the height. For instance, say that the person's height is 1.42m, then a good guess about his weight is 44kg, which is substantially less that $E[Y]=51$, the average weight over all the people in that population. The practical question that we explore in this note is how one should use the information X to guess Y.

The basic estimation problem can be formulated as follows. There is a pair of random variables (X,Y) defined on some common probability space. The problem is to estimate Y from the observed value of X.

This problem admits a few different formulations:

- **Known Distribution**: We know the joint distribution of (X,Y);
- **Off-Line**: We observe a set of sample values of (X,Y);



$$E[\hat{Y}c] = 0 \Leftrightarrow \hat{Y} \perp c$$

$$\begin{aligned} E[(Y-a)^2] &= E[(\hat{Y}+c)^2] \\ &= E[\hat{Y}^2 + 2c\hat{Y} + c^2] \\ &= E[\hat{Y}^2] + c^2 \quad \text{(Pythagoras)} \end{aligned}$$

Figure 1: $E[Y]$ is the least squares estimates of Y.

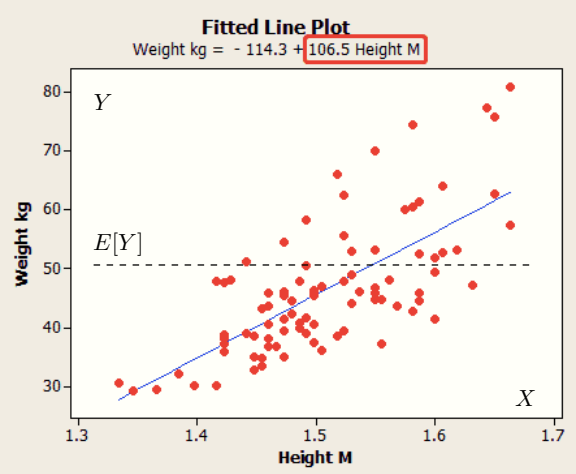


Figure 2: Height and weight of a number of people.

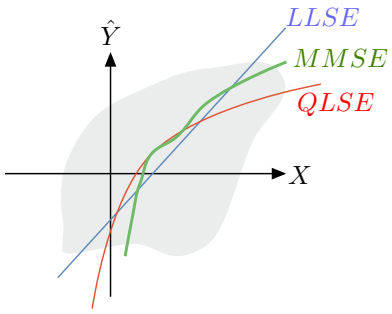


Figure 3: Least squares estimates $\hat{Y}$ of Y given X: *LLSE* is linear, *QLSE* is quadratic, and *MMSE* can be an arbitrary function.

- **On-Line**: We observe successive values of samples of (X,Y).

The objective of the estimation is to choose the inference function $g(\cdot)$ to minimize the expected error $C(g)$ where

$$C(g) = E(c(Y, g(X))).$$

In this expression, $g(X)$ is the guess $\hat{Y}$ that one makes about Y given the observed value X and $c(Y,\hat{Y})$ is the cost of guessing $\hat{Y}$ when the actual value is Y. A standard example is

$$c(Y,\hat{Y}) = |Y-\hat{Y}|^2.$$

We will also study the case when $Y \in \Re^d$ for $d > 1$. In such a situation, one uses $c(Y,\hat{Y}) = \|Y-\hat{Y}\|^2$. In this case, the corresponding best guess is said to be the *Least Squares Estimate (LSE)* of Y given X. If the function $g(\cdot)$ can be arbitrary, it is the *Minimum Mean Squares Estimate (MMSE)* of Y given X. If the function $g(\cdot)$ is restricted to be linear, i.e., of the form $a+BX$, it is the *Linear Least Squares Estimate (LLSE)* of Y given X. One may also restrict $g(\cdot)$ to be a polynomial of a given degree. For instance, one may define the Quadratic Least Squares Estimate *QLSE* of Y given X. See Figure 3.

One may wonder at the choice of the square of the error as opposed to another measure. One key justification is simplicity. For instance, choosing $c(Y,\hat{Y}) = |Y-\hat{Y}|$ makes it much more complicated to derive the function $g(\cdot)$ that minimizes $C(g)$. To minimize the average value $E[|Y-a|]$ without any other information about Y than its distribution, one should choose a to be the median of that distribution, i.e., a value v such that $P(Y \geq v) = P(Y \leq v)$. If one makes an observation X, there is no easy way to calculate the linear function $a+bX$ that minimizes $E[|Y-a-bX|]$.

Thus, we will use the square of the error, mostly for convenience. Another justification is that the square of the error has the interpretation of noise power that is relevant in a number of applications, such as in signal processing.

Linear Least Squares Estimate

In this section, we study the linear least squares estimate. Recall the setup that we explained in the previous section. There is a pair (X,Y) of random variables with some joint distribution and the problem is to find the function $g(X) = a+bX$ that minimizes

$$C(g) = E(|Y-g(X)|^2).$$

↳ Expected cost of guessing Y value

We assume that the joint distribution of (X,Y) is known and we are looking for the linear function $g(X) = a+bX$ that minimizes

$$C(g) = E(|Y-g(X)|^2) = E(|Y-a-bX|^2).$$

We denote this function by $L[Y|X]$. Thus, we have the following definition.

Definition 26.1. *Linear Least Squares Estimate (LLSE) The LLSE of Y given X, denoted by $L[Y|X]$, is the linear function $a+bX$ that minimizes*

$$E(|Y-a-bX|^2).$$

Note that

$$\begin{aligned} C(g) &= E(Y^2+a^2+b^2X^2-2aY-2bYX+2abX) \\ &= E(Y^2)+a^2+b^2E(X^2)-2aE(Y)-2bE(YX)+2abE(X). \end{aligned}$$

To find the values of a and b that minimize that expression, we set to zero the partial derivatives with respect to a and b. This gives the following two equations:

$$0 = 2a-2E(Y)+2bE(X) \tag{1}$$

$$0 = 2bE(X^2)-2E(YX)+2aE(X). \tag{2}$$

Solving these equations for a and b, we find that

$$L[Y|X] = a+bX = E(Y)+\frac{cov(X,Y)}{var(X)}(X-E(X))$$

where we used the following definitions:

Definition 26.2. *Let X and Y be random variables defined on the same probability space. Then,*

$$cov(X,Y) := E(YX)-E(Y)E(X) \text{ and } var(X) := E(X^2)-E(X)^2.$$

The quantity $cov(X,Y)$ is called the covariance *of X and Y. These two random variables are said to be* uncorrelated *if $cov(X,Y)=0$,* positively correlated *if $cov(X,Y)>0$, and* negatively correlated *if $cov(X,Y)<0$.*

It is a simple exercise to show that

$$cov(a+bX,c+dY) = bd.cov(X,Y), var(X+Y) = var(X)+var(Y)+2cov(X,Y) \text{ and } var(a+bX) = b^2var(X).$$

Note also that X and Y are uncorrelated if they are independent, but that the converse is not true. For instance, if (X,Y) is uniformly distributed in $\{(-1,0),(0,1),(1,0)\}$, then X and Y are uncorrelated but not independent. Also, as was shown in a previous lecture, the variance of the sum of pairwise independent random variables is the sum of their variances.

We summarize the result above as a theorem.

Theorem 26.1. *Linear Least Squares Estimate*

One has

$$L[Y|X] = E(Y)+\frac{cov(X,Y)}{var(X)}(X-E(X)). \tag{3}$$

As a first example, assume that

$$X = \alpha Y + Z \tag{4}$$

where Y and Z are zero-mean and independent. The interpretation of this identity is that X is a noisy measurement of a scaled version of Y. For instance, Y could be the voltage at one end of a pair of wires and X the measured voltage at the other end. In this case, we find

$$\begin{aligned} cov(X,Y) &= E(YX) - E(Y)E(X) \\ &= E(Y(\alpha Y + Z)) = \alpha E(Y^2) \\ var(X) &= \alpha^2 var(Y) + var(Z) = \alpha^2 E(Y^2) + E(Z^2). \end{aligned}$$

Hence,

$$L[Y|X] = \frac{\alpha E(Y^2)}{\alpha^2 E(Y^2) + E(Z^2)} X = \frac{\alpha^{-1} X}{1 + SNR^{-1}}$$

where

$$SNR := \frac{\alpha^2 E(Y^2)}{\sigma^2}$$

is the *signal-to-noise ratio*, i.e., the ratio of the power $E(\alpha^2 Y^2)$ of the signal in X divided by the power $E(Z^2)$ of the noise. Note that if SNR is small, then $L[Y|X]$ is close to zero, which is the best guess about Y if one does not make any observation. Also, if SNR is very large, then $L[Y|X] \approx \alpha^{-1} X$, which is the correct guess if $Z = 0$.

As a second example, assume that

$$Y = \alpha X + \beta X^2 \tag{5}$$

where $E[X] = 1/2, E[X^2] = 1/3$ and $E[X^3] = 1/4$. The point of this example is that Y is a deterministic function of X, so that it is easy to calculate Y from X. However, what if one is required to approximate Y by a linear function of X?

We find

$$\begin{aligned} E(Y) &= \alpha E(X) + \beta E(X^2) = \alpha/2 + \beta/3; \\ cov(X,Y) &= E(YX) - E(Y)E(X) \\ &= E(\alpha X^2 + \beta X^3) - (\alpha/2 + \beta/3)(1/2) \\ &= \alpha/3 + \beta/4 - \alpha/4 - \beta/6 \\ &= (\alpha + \beta)/12 \\ var(X) &= E(X^2) - E(X)^2 = 1/3 - (1/2)^2 = 1/12. \end{aligned}$$

Hence,

$$L[Y|X] = \alpha/2 + \beta/3 + (\alpha + \beta)(X - 1/2) = -\beta/6 + (\alpha + \beta)X.$$

This estimate is sketched in Figure 4.

Projection

There is an insightful interpretation of $L[Y|X]$ as a projection that also helps understand more complex estimates. This interpretation is that $L[Y|X]$ is the projection of Y onto the set $\mathscr{L}(X)$ of linear functions of X.

This interpretation is sketched in Figure 5. In that figure, random variables are represented by points and $\mathscr{L}(X)$ is shown as a plane since the linear combination of points in that set is again in the set. (A linear

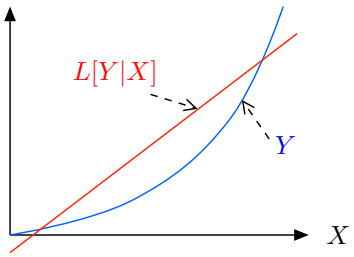


Figure 4: The figure shows $L[Y|X]$ where $Y = \alpha X + \beta X^2$.

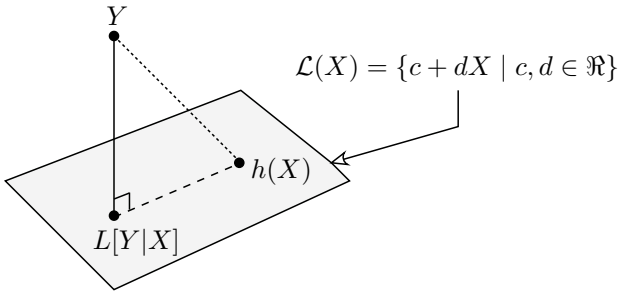


Figure 5: $L[Y|X]$ is the projection of Y onto $\mathscr{L}(X)$.

combination of linear functions of X is also a linear function of X.) In the figure, the square of the length of a vector from a random variable V to another random variable W is $E(|V-W|^2)$. Also, we say that two vectors V and W are orthogonal if $E(VW) = 0$. Thus, $L[Y|X] = a + bX$ is the projection of Y onto $\mathscr{L}(X)$ if $Y - L[Y|X]$ is orthogonal to every linear function of X, i.e., if

$$E((Y-a-bX)(c+dX)) = 0, \forall c, d \in \Re.$$

Equivalently,

$$E(Y) = a + bE(X) \text{ and } E((Y-a-bX)X) = 0. \tag{6}$$

These two equations are the same as (1)-(2). We call the identities (6) the projection property.

Figure 6 illustrates the projection when

$$E(Y) = 0, var(Y) = 1 \text{ and } X = Y + Z \text{ where } E(Z) = 0 \text{ and } E(Z^2) = \sigma^2.$$

In this figure, the length of Z is equal to $\sqrt{E(Z^2)} = \sigma$, the length of Y is $\sqrt{E(Y^2)} = 1$ and the vectors Y and Z are orthogonal because $E(YZ) = 0$.

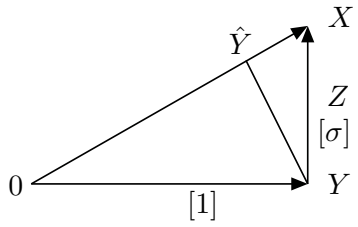


Figure 6: Example of projection.

We see that the triangles $0\hat{Y}Y$ and $0YX$ are similar. Hence,

$$\frac{\|\hat{Y}\|}{\|Y\|} = \frac{\|Y\|}{\|X\|},$$

so that

$$\frac{\|\hat{Y}\|}{1} = \frac{1}{\sqrt{1+\sigma^2}} = \frac{\|X\|}{1+\sigma^2},$$

since $\|X\| = \sqrt{1+\sigma^2}$. This shows that

$$\hat{Y} = \frac{1}{1+\sigma^2}X.$$

To see why the projection property implies that $L[Y|X]$ is the closest point to Y in $\mathscr{L}(X)$, as suggested by Figure 5, we verify that

$$E(|Y - L[Y|X]|^2) \leq E(|Y - h(X)|^2),$$

for any given $h(X) = c + dX$. The idea of the proof is to verify Pythagora's identity on the right triangle with vertices $Y, L[Y|X]$ and $h(X)$. We have

$$\begin{aligned} E(|Y-h(X)|^2) &= E(|Y - L[Y|X] + L[Y|X] - h(X)|^2) \\ &= E(|Y - L[Y|X]|^2) + E(|L[Y|X] - h(X)|^2) \\ &\quad + 2E((Y - L[Y|X])(L[Y|X] - h(X))). \end{aligned}$$

Now, the projection property (6) implies that the last term in the above expression is equal to zero. Indeed, $L[Y|X] - h(X)$ is a linear function of X. It follows that

$$\begin{aligned} E(|Y-h(X)|^2) &= E(|Y - L[Y|X]|^2) + E(|L[Y|X] - h(X)|^2) \\ &\geq E(|Y - L[Y|X]|^2), \end{aligned}$$

as was to be proved.

Note the similarity of this argument with the proof given at the beginning of this note that $E[Y]$ is the value of a that minimizes $E[(Y-a)^2]$. There we needed the error $Y - E[Y]$ to be orthogonal to all the constants, which only requires that $E[Y - E[Y]] = 0$. Thus, $E[Y]$ is the projection of Y onto the set of constants because the error is orthogonal to the 'plane' of constants. Here, $L[Y|X]$ is characterized by the property that the error $Y - L[Y|X]$ is orthogonal to the plane of linear functions $c + dX$, which requires $E[Y - L[Y|X]] = 0$ and $E[(Y - L[Y|X])X] = 0$.

Linear Regression

Assume now that, instead of knowing the joint distribution of (X,Y), we observe K samples $(X_1,Y_1),\ldots,(X_K,Y_K)$ of these random variables. Our goal is to choose the values a and b that minimize

$$\frac{1}{K}\sum_{k=1}^{K} |Y_k - a - bX_k|^2. \tag{7}$$

The resulting expression for $a+bX$ is then called the *linear regression* of Y over X.

Note that in this formulation we do not have any prior distribution of the random variables X and Y. Instead, we only observe sample values. We say that this is a *non-Bayesian* formulation. Many practitioners tend to

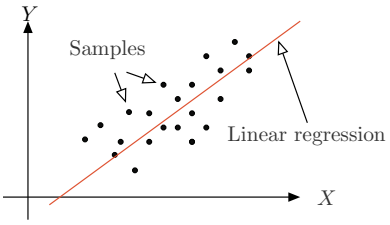


Figure 7: The linear regression of Y over X.

trust a non-Bayesian approach more than the Bayesian one covered in the previous section that relies on a possibly incorrect assumption about the prior distribution.

To find the best values of a and b, we note that the cost (7) is equal to

$$E[|Y-a-bX|^2]$$

where the random pair (X,Y) is defined to be equally likely to take any of the values $\{(X_1,Y_1),\ldots,(X_K,Y_K)\}$. Indeed, with that definition of (X,Y), one finds

$$E[|Y-a-bX|^2] = \sum_{(x,y)} |x-a-by|^2 P[Y=x,X=y] = \sum_k |Y_k-a-bX_k|^2 \frac{1}{K},$$

which agrees with (7).

Consequently, the linear regression of Y over X given the K samples (X_k,Y_k) is given by (3) where the expected values, the covariance, and the variance are calculated for the random variables (X,Y) uniformly distributed in the set of sample values. For instance, $E[X] = \frac{1}{K}\sum_{k=1}^K X_k$ and similarly for $E[Y]$ and $cov(X,Y) = \frac{1}{K}\sum_{k=1}^K X_k Y_k - E[X]E[Y]$.

One has the following result.

Theorem 26.2. *Linear Regression Converges to LLSE*

Assume that the samples $(X_k,Y_k), k \geq 1$ are i.i.d. with distribution $P(X=x,Y=y)$. As the number of samples increases, the linear regression approaches the LLSE calculated from the distribution of (X,Y).

Proof:

As $K \to \infty$, the Law of Large Numbers implies that the sample means converge to the expected values. For instance, $\frac{1}{K}\sum_{k=1}^K X_k$ approaches $E[X]$ and similarly for the other quantities.

□

Formula (3) and the linear regression provide an intuitive meaning of the covariance $cov(X,Y)$. If this covariance is zero, then $L[Y|X]$ does not depend on X. If it is positive (negative), it increases (decreases, respectively) with X. Thus, $cov(X,Y)$ measures a form of dependency in terms of linear regression. For instance, the random variables in Figure 8 are uncorrelated since $L[Y|X]$ does not depend on Y.

MMSE

In the previous section, we examined the problem of finding the linear function $a+bX$ that best approximates Y, in the mean squared error sense. We could develop the corresponding theory for quadratic approximations

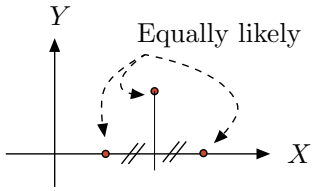


Figure 8: The random variables Y and X are uncorrelated. Note that they are not independent.

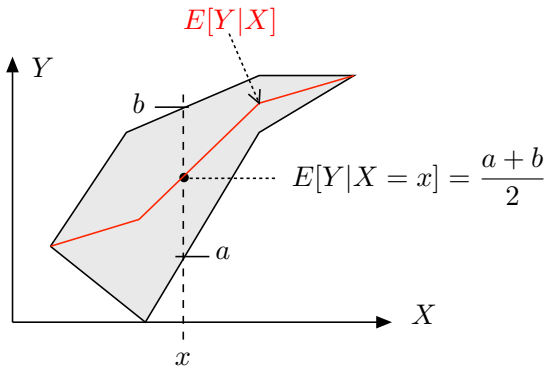


Figure 9: The conditional expectation $E[Y|X]$ when the pair (X,Y) is picked uniformly in the shaded area.

$a+bX+cX^2$, or for polynomial approximations of a given degree. The ideas would be the same and one would have a similar projection interpretation.

In principle, a higher degree polynomial approximates Y better than a lower degree one since there are more such polynomials. The question of fitting the parameters with a given number of observations is more complex. For instance, if one has K samples, one can generically find a polynomial of degree K that fits the observations exactly. However, this does not imply that this polynomial results in a smaller mean squared error than a lower-degree polynomial. This issue is called "over fitting."

For now, assume that we know the joint distribution of (X,Y) and consider the problem of finding the function $g(X)$ that minimizes

$$E(|Y-g(X)|^2),$$

per all the possible functions $g(\cdot)$. The best function is called the MMSE of Y given X.

Before proceeding with the formal discussion of the solution, let us see what it should be like. We saw that the best guess for Y without any information other than the distribution is $E[Y]$. Now say that we observe that $X=x$. That observation changes the likelihood that $Y=y$ from $P(Y=y)$ to $P[Y=y|X=x]=P(X=x,Y=y)/P(X=x)$. Thus, observing $X=x$ replaces the distribution of Y by its conditional distribution given $X=x$. That conditional distribution is $\{(y,P[Y=y|X=x]),y\in\Re\}$. Since Y has this new distribution, the best guess for the value of Y given that $X=x$ should be the expected value of Y under this new distribution. This expected value is called the conditional expectation of Y given that $X=x$. Let us make all this precise.

We have the following theorem:

Theorem 26.3. *The MMSE is the Conditional Expectation*

The MMSE of Y given X is given by

$$g(X)=E[Y|X]$$

where $E[Y|X]$ is the conditional expectation *of Y given X.*

The conditional expectation is defined as follows.

Definition 26.3. *Conditional Expectation*

The conditional expectation of Y given X is defined by

$$E[Y|X=x] = \sum_y yP[Y=y|X=x] = \sum_y y\frac{P(X=x,Y=y)}{P(X=x)}.$$

Figure 9 illustrates the conditional expectation. That figure assumes that the pair (X,Y) is picked uniformly in the shaded area. Thus, if one observes that $X=x$, the point Y is uniformly distributed along the segment $[a,b]$. Accordingly, the average value of Y is the mid-point $(a+b)/2$ of that segment, as indicated in the figure. The red graph shows how that mean value depends on X and it defines $E[Y|X]$. Of course, if the point (X,Y) is not chosen uniformly in the shaded area, then $E[Y|X]$ is different.

The following result is a direct consequence of the definition.

Lemma 26.1. *Orthogonality Property of MMSE*

(a) For any function $\phi(\cdot)$, one has

$$E((Y-E[Y|X])\phi(X)) = 0. \tag{8}$$

(b) Moreover, if the function $g(X)$ is such that

$$E((Y-g(X))\phi(X)) = 0, \forall \phi(\cdot), \tag{9}$$

then $g(X) = E[Y|X]$.

Proof:

(a) To verify (8) note that

$$\begin{aligned} E(E[Y|X]\phi(X)) &= \sum_y E[Y|X=y]\phi(y)P(X=y) \\ &= \sum_y \left[\sum_x x\frac{P(Y=x,X=y)}{P(X=y)}\right]\phi(y)P(X=y) \\ &= \sum_y\sum_x x\phi(y)P(Y=x,X=y) = E(Y\phi(X)), \end{aligned}$$

which proves (8).

(b) To prove the second part of the lemma, note that

$$\begin{aligned} &E(|g(X)-E[Y|X]|^2) \\ &\quad = E((g(X)-E[Y|X])\{(g(X)-Y)-(E[Y|X]-Y)\}) = 0, \end{aligned}$$

because of (8) and (9) with $\phi(X) = g(X)-E[Y|X]$.

Note that the second part of the lemma simply says that the projection property characterizes uniquely the conditional expectation. In other words, there is only one projection of Y onto $\mathcal{G}(X)$. □

We can now prove the theorem.

Proof of Theorem 26.3.

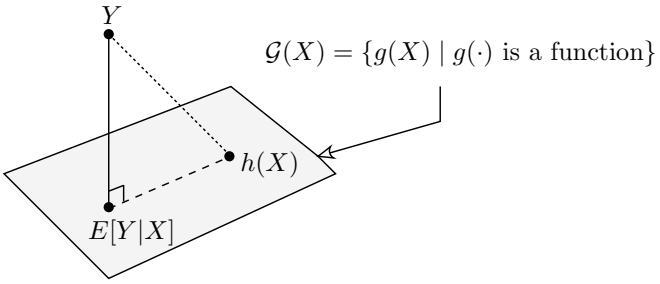


Figure 10: The conditional expectation $E[Y|X]$ as the projection of Y on the set $\mathcal{G}(X)$ of functions of X.

The identity (8) is the projection property. It states that $Y - E[Y|X]$ is orthogonal to the set $\mathcal{G}(X)$ of functions of X, as shown in Figure 10. In particular, it is orthogonal to $h(X) - E[Y|X]$. As in the case of the LLSE, this projection property implies that

$$E(|Y - h(X)|^2) \geq E(|Y - E[Y|X]|^2),$$

for any function $h(\cdot)$. This implies that $E[Y|X]$ is indeed the MMSE of Y given X. □

From the definition, we see how to calculate $E[Y|X]$ from the conditional density of Y given X. However, in many cases one can calculate $E[Y|X]$ more simply. One approach is to use the following properties of conditional expectation.

Theorem 26.4. *Properties of Conditional Expectation (a) Linearity:*

$$E[a_1Y_1 + a_2Y_2|X] = a_1E[Y_1|X] + a_2E[Y_2|X];$$

(b) Factoring Known Values:

$$E[h(X)Y|X] = h(X)E[Y|X];$$

(c) Smoothing:

$$E(E[Y|X]) = E(Y);$$

(d) Independence: If Y and X are independent, then

$$E[Y|X] = E(Y).$$

Proof:

(a) By Lemma 26.1(b), it suffices to show that

$$a_1Y_1 + a_2Y_2 - (a_1E[Y_1|X] + a_2E[Y_2|X])$$

is orthogonal to $\mathcal{G}(X)$. But this is immediate since it is the sum of two terms

$$a_i(Y_i - E[Y_i|X])$$

for $i = 1,2$ that are orthogonal to $\mathcal{G}(X)$.

(b) By Lemma 26.1(b), it suffices to show that

$$h(X)Y - h(X)E[Y|X]$$

iid = independent, identically distributed.

is orthogonal to $\mathcal{G}(X)$, i.e., that

$$E((h(X)Y - h(X)E[Y|X])\phi(X)) = 0, \forall \phi(\cdot).$$

Now,

$$E((h(X)Y - h(X)E[Y|X])\phi(X)) = E((Y - E[Y|X])h(X)\phi(X)) = 0,$$

because $Y - E[Y|X]$ is orthogonal to $\mathcal{G}(X)$ and therefore to $h(X)\phi(X)$.

(c) Letting $\phi(X) = 1$ in (8), we find

$$E(Y - E[Y|X]) = 0,$$

which is the identity we wanted to prove.

(d) By Lemma 26.1(b), it suffices to show that

$$Y - E(Y)$$

is orthogonal to $\mathcal{G}(X)$. Now,

$$E((Y - E(Y))\phi(X)) = E(Y - E(Y))E(\phi(X)) = 0.$$

The first equality follows from the fact that $Y - E(Y)$ and $\phi(X)$ are independent since they are functions of independent random variables.

□

As an example, assume that Y, X, Z are i.i.d. with mean $1/2$ and second moment $1/3$. We want to calculate

$$E[(Y + 2X)^2|X].$$

We find

$$\begin{aligned}
E[(Y+2X)^2|X] &= E[Y^2 + 4X^2 + 4YX|X] \\
&= E[Y^2|X] + 4E[X^2|X] + 4E[YX|X], \text{ by linearity} \\
&= E(Y^2) + 4E[X^2|X] + 4E[YX|X], \text{ by independence} \\
&= E(Y^2) + 4X^2 + 4XE[Y|X], \text{ by factoring known values} \\
&= E(Y^2) + 4X^2 + 4XE(Y), \text{ by independence} \\
&= \frac{1}{3} + 4X^2 + 2X.
\end{aligned}$$

Note that we don't have enough information to calculate the conditional distribution of $Z := (Y+2X)^2$ given X, so that a direct calculation $E[Z|X = y] = \sum_z zP[Z = z|X = y]$ would not have been possible.

In some situations, one may be able to exploit symmetry to evaluate the conditional expectation. Here is one representative example. Assume that Y, X, Z are i.i.d. Then, we claim that

$$E[Y|Y + X + Z] = \frac{1}{3}(Y + X + Z). \tag{10}$$

To see this, note that, by symmetry,

$$E[Y|Y + X + Z] = E[X|Y + X + Z] = E[Z|Y + X + Z].$$

Denote by V the common value of these random variables. Note that their sum is

$$3V = E[Y + X + Z|Y + X + Z],$$

by linearity. Thus, $3V = Y + X + Z$, which proves our claim.

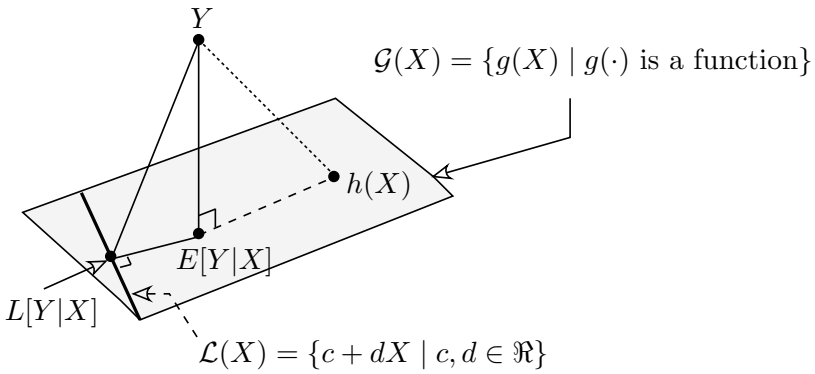


Figure 11: The MMSE and LLSE are generally different.

MMSE for Jointly Gaussian[1]

In general $L[Y|X] \neq E[Y|X]$. As a trivial example, Let X be uniformly distributed in $\{-1,0,1\}$ and $Y = X^2$. Then $E[Y|X] = X^2$ and $L[Y|X] = E(Y) = 2/3$ since $cov(X,Y) = E(YX) - E(Y)E(X) = 0$.

Figure 11 recalls that $E[Y|X]$ is the projection of Y onto $\mathcal{G}(X)$ whereas $L[Y|X]$ is the projection of Y onto $\mathcal{L}(X)$. Since $\mathcal{L}(X)$ is a subspace of $\mathcal{G}(X)$, one expects the two projections to be different, in general. The figure also shows that $E[Y|X]$ is closer to Y than $L[Y|X]$, in the mean squared error sense. This should be obvious since there are more functions of X than linear functions. Note also that the figure shows that $L[Y|X]$ is the projection of $E[Y|X]$ onto $\mathcal{L}(X)$, which says that $L[Y|X] = L[E[Y|X]|X]$.

However, there are examples where $E[Y|X]$ happens to be linear. We saw one such example in (10) and it is not difficult to construct many other examples.

There is an important class of situations where this occurs. It is when Y and X are jointly Gaussian. We state that result as a theorem. Unfortunately, we have not discussed the necessary concepts to explain what this result means. (We have not even defined what Gaussian means!) We state the result to pick your curiosity.

Theorem 26.5. *MMSE for Jointly Gaussian RVs*

Let Y,X be jointly Gaussian random variables. Then

$$E[Y|X] = L[Y|X].$$

Vector Case[2]

So far, to keep notation at a minimum, we have considered $L[Y|X]$ and $E[Y|X]$ when Y and X are single random variables. In this section, we discuss the vector case, i.e., $L[\mathbf{Y}|\mathbf{X}]$ and $E[\mathbf{Y}|\mathbf{X}]$ when $\mathbf{Y}$ and $\mathbf{X}$ are random vectors. The only difficulty is one of notation. Conceptually, there is nothing new.

Definition 26.4. *LLSE of Random Vectors*

Let $\mathbf{Y}$ and $\mathbf{X}$ be random vectors of dimensions m and n, respectively. Then

$$L[\mathbf{Y}|\mathbf{X}] = \mathbf{a} + B\mathbf{X}$$

[1] This topic will not be covered this semester

[2] This topic will not be covered this semester

where B is the $m \times n$ matrix and $\mathbf{a}$ the vector in $\Re^m$ that minimize

$$E(\|\mathbf{Y} - \mathbf{a} - B\mathbf{X}\|^2).$$

Thus, as in the scalar case, the LLSE is the linear function of the observations that best approximates $\mathbf{Y}$, in the mean squared error sense.

Before proceeding, we define $\Sigma_{\mathbf{X}}$ and $\Sigma_{\mathbf{Y},\mathbf{X}}$. By $\mathbf{v}'$ we indicate the transposed of the vector $\mathbf{v}$ and by $\mathbf{B}'$ the transposed of the matrix $\mathbf{B}$.

Definition 26.5. *Let $\mathbf{Y} = [Y_1, \ldots, Y_m]'$ and $\mathbf{X} = [X_1, \ldots, X_n]'$ be two random (column) vectors. Then*

$$E[\mathbf{Y}] := [E(Y_1), \ldots, E(Y_m)]', E[\mathbf{X}] := [E(X_1), \ldots, E(X_n)]',$$

i.e., the mean values of the random vectors are the vectors of the mean values of their components. The product $\mathbf{Y}\mathbf{X}'$ is an $m \times n$ matrix whose component (i, j) is $Y_i X_j$, and similarly for $\mathbf{Y}\mathbf{Y}'$ and $\mathbf{X}\mathbf{X}'$. We define $E[\mathbf{Y}\mathbf{X}']$ to be the $m \times n$ matrix whose component (i, j) is $E[Y_i X_j]$. That is, the mean value of a random matrix is defined to be the matrix of the mean values of its components. The definitions of $E[\mathbf{Y}\mathbf{Y}']$ and $E[\mathbf{X}\mathbf{X}']$ are similar.

Finally, we define

$$\Sigma_{\mathbf{X}} := E[\mathbf{X}\mathbf{X}'] - E[\mathbf{X}]E[\mathbf{X}]' \text{ and } \Sigma_{\mathbf{Y},\mathbf{X}} := E[\mathbf{Y}\mathbf{X}'] - E[\mathbf{Y}]E[\mathbf{X}]'.$$

We call $\Sigma_{\mathbf{X}}$ the covariance matrix *of $\mathbf{X}$ and $\Sigma_{\mathbf{Y},\mathbf{X}}$ the* covariance matrix *of $\mathbf{Y}$ and $\mathbf{X}$.*

Theorem 26.6. *LLSE of Vectors*

Let $\mathbf{Y}$ and $\mathbf{X}$ be random vectors such that $\Sigma_{\mathbf{X}}$ is nonsingular, i.e., such that the matrix admits an inverse.

(a) Then

$$L[\mathbf{Y}|\mathbf{X}] = E(\mathbf{Y}) + \Sigma_{\mathbf{Y},\mathbf{X}}\Sigma_{\mathbf{X}}^{-1}(\mathbf{X} - E(\mathbf{X})). \tag{11}$$

(b) Moreover,

$$E(\|\mathbf{Y} - L[\mathbf{Y}|\mathbf{X}]\|^2) = tr(\Sigma_{\mathbf{Y}} - cov(\mathbf{Y},\mathbf{X})\Sigma_{\mathbf{X}}^{-1}cov(\mathbf{X},\mathbf{Y})). \tag{12}$$

In this expression, for a square matrix M, $tr(M) := \sum_i M_{i,i}$ is the trace of the matrix.

Proof:

(a) The proof is similar to the scalar case. Let $\mathbf{Z}$ be the right-hand side of (11). One shows that the error $\mathbf{Y} - \mathbf{Z}$ is orthogonal to all the linear functions of $\mathbf{X}$. One then uses that fact to show that $\mathbf{Y}$ is closer to $\mathbf{Z}$ than to any other linear function $h(\mathbf{X})$ of $\mathbf{X}$.

First we show the orthogonality. Since $E(\mathbf{Y} - \mathbf{Z}) = 0$, we have

$$E((\mathbf{Y} - \mathbf{Z})(\mathbf{a} + B\mathbf{X})') = E((\mathbf{Y} - \mathbf{Z})(B\mathbf{X})') = E((\mathbf{Y} - \mathbf{Z})\mathbf{X}')B'.$$

Next, we show that $E((\mathbf{Y} - \mathbf{Z})\mathbf{X}') = 0$. To see this, note that

$$\begin{aligned} E((\mathbf{Y} - \mathbf{Z})\mathbf{X}') &= E((\mathbf{Y} - \mathbf{Z})(\mathbf{X} - E(\mathbf{X}))') \\ &= E((\mathbf{Y} - E(\mathbf{Y}))(\mathbf{X} - E(\mathbf{X}))') \\ &\quad - \Sigma_{\mathbf{Y},\mathbf{X}}\Sigma_{\mathbf{X}}^{-1}E((\mathbf{X} - E(\mathbf{X}))(\mathbf{X} - E(\mathbf{X}))') \\ &= \Sigma_{\mathbf{Y},\mathbf{X}} - \Sigma_{\mathbf{Y},\mathbf{X}}\Sigma_{\mathbf{X}}^{-1}\Sigma_{\mathbf{X}} = 0. \end{aligned}$$

Second, we show that $\mathbf{Z}$ is closer to $\mathbf{Y}$ than any linear $h(\mathbf{X})$. We have

$$\begin{aligned}
E(\|\mathbf{Y}-h(\mathbf{X})\|^2) &= E((\mathbf{Y}-h(\mathbf{X}))'(\mathbf{Y}-h(\mathbf{X}))) \\
&= E((\mathbf{Y}-\mathbf{Z}+\mathbf{Z}-h(\mathbf{X}))'(\mathbf{Y}-\mathbf{Z}+\mathbf{Z}-h(\mathbf{X}))) \\
&= E(\|\mathbf{Y}-\mathbf{Z}\|^2)+E(\|\mathbf{Z}-h(\mathbf{X})\|^2)+2E((\mathbf{Y}-\mathbf{Z})'(\mathbf{Z}-h(\mathbf{X}))).
\end{aligned}$$

We claim that the last term is equal to zero. To see this, note that

$$E((\mathbf{Y}-\mathbf{Z})'(\mathbf{Z}-h(\mathbf{X}))=\sum_{i=1}^{n}E((Y_i-Z_i)(Z_i-h_i(\mathbf{X}))).$$

Also,

$$E((Y_i-Z_i)(Z_i-h_i(\mathbf{X})))=E((\mathbf{Y}-\mathbf{Z})(\mathbf{Z}-h(\mathbf{X}))')_{i,i}$$

and the matrix $E((\mathbf{Y}-\mathbf{Z})(\mathbf{Z}-h(\mathbf{X}))')$ is equal to zero since $\mathbf{Y}-\mathbf{X}$ is orthogonal to any linear function of $\mathbf{X}$ and, in particular, to $\mathbf{Z}-h(\mathbf{X})$.

(Note: an alternative way of showing that the last term is equal to zero is to write

$$E((\mathbf{Y}-\mathbf{Z})'(\mathbf{Z}-h(\mathbf{X}))=\mathrm{tr}E((\mathbf{Y}-\mathbf{Z})(\mathbf{Z}-h(\mathbf{X}))')=0,$$

where the first equality comes from the fact that $\mathrm{tr}(AB)=\mathrm{tr}(BA)$ for matrices of compatible dimensions.)

(b) Let $\tilde{\mathbf{Y}}:=\mathbf{Y}-E[\mathbf{Y}|\mathbf{X}]$ be the estimation error. Thus,

$$\tilde{\mathbf{Y}}=\mathbf{Y}-E(\mathbf{Y})-\Sigma_{\mathbf{Y},\mathbf{X}}\Sigma_{\mathbf{X}}^{-1}(\mathbf{X}-E(\mathbf{X})).$$

Now, if $\mathbf{V}$ and $\mathbf{W}$ are two zero-mean random vectors and M a matrix,

$$\begin{aligned}
cov(\mathbf{V}-M\mathbf{W}) &= E((\mathbf{V}-M\mathbf{W})(\mathbf{V}-M\mathbf{W})') \\
&= E(\mathbf{V}\mathbf{V}'-2M\mathbf{W}\mathbf{V}'+M\mathbf{W}\mathbf{W}'M') \\
&= cov(\mathbf{V})-2Mcov(\mathbf{W},\mathbf{V})+Mcov(\mathbf{W})M'.
\end{aligned}$$

Hence,

$$\begin{aligned}
cov(\tilde{\mathbf{Y}}) &= \Sigma_{\mathbf{Y}}-2\Sigma_{\mathbf{Y},\mathbf{X}}\Sigma_{\mathbf{X}}^{-1}cov(\mathbf{X},\mathbf{Y}) \\
&\quad +\Sigma_{\mathbf{Y},\mathbf{X}}\Sigma_{\mathbf{X}}^{-1}\Sigma_{\mathbf{X}}\Sigma_{\mathbf{X}}^{-1}cov(\mathbf{X},\mathbf{Y}) \\
&= \Sigma_{\mathbf{Y}}-\Sigma_{\mathbf{Y},\mathbf{X}}\Sigma_{\mathbf{X}}^{-1}cov(\mathbf{X},\mathbf{Y}).
\end{aligned}$$

To conclude the proof, note that, for a zero-mean random vector $\mathbf{V}$,

$$E(\|\mathbf{V}\|^2)=E(\mathrm{tr}(\mathbf{V}\mathbf{V}'))=\mathrm{tr}(E(\mathbf{V}\mathbf{V}'))=\mathrm{tr}(\Sigma_{\mathbf{V}}).$$

□

Summary

- LLSE, linear regression and MMSE;
- Projection characterization;
- MMSE of jointly Gaussian random variables is linear;

Key equations & formulas:

LLSE	$L[Y\|X] = E(Y) + cov(X,Y)var(X)^{-1}(X - E(X))$	T. 26.1
Projection	$Y - L[Y\|X] \perp a + bX$	(6)
Linear Regression	(X,Y) uniform in set of samples; converges to $L[Y\|X]$	T.26.2
Conditional Expectation	$E[Y\|X] := ...$	D.26.3
Projection	$Y - E[Y\|X] \perp g(X)$	L.26.1
MMSE = CE	$MMSE[Y\|X] = E[Y\|X]$	T.26.3
Properties of CE	Linearity, smoothing, etc...	T.26.4
CE for J.G.	If Y,X J.G., then $E[Y\|X] = L[Y\|X]$	T.26.5
LLSE vectors	$L[\mathbf{Y}\|\mathbf{X}] = E(\mathbf{Y}) + \Sigma_{\mathbf{Y},\mathbf{X}}\Sigma_{\mathbf{X}}^{-1}(\mathbf{X} - E(\mathbf{X}))$	T.26.6