Returning to our original claim, write

$$P(\widehat{Y} = Y) = E\left[P(\widehat{Y} = Y \mid X)\right] \tag{1.53}$$

In that inner probability, "p" is

$$P(Y = 1 \mid X) = \mu(X) \tag{1.54}$$

which completes the proof.

1.19.5 Some Properties of Conditional Expectation

Since the regression function is defined as a conditional expected value, as in (1.3), for mathematical analysis we'll need some properties. First, a definition.

1.19.5.1 Conditional Expectation As a Random Variable

For any random variables U and V with defined expectation, either of which could be vector-valued, define a new random variable W, as follows. First note that the conditional expectation of V given $U = t$ is a function of t,

$$\mu(t) = E(V \mid U = t) \tag{1.55}$$

This is an ordinary function, just like, say, $\sqrt{t}$. But we can turn that ordinary function into a random variable by plugging in a random variable, say Q, for t: $R = \sqrt{Q}$ is a random variable. Thinking along these lines, we define the *random variable* version of conditional expectation accordingly. In the *function* $\mu(t)$ in (1.55), we plug in U for t:

$$W = E(V|U) = \mu(U) \tag{1.56}$$

This W is a random variable. As a simple example, say we choose a number U at random from the numbers 1 through 5. We then randomly choose a second number V, from the numbers 1 through U. Then

$$\mu(t) = E(V \mid U = t) = \frac{1+t}{2} \tag{1.57}$$

We now form a new random variable $W = (1 + U)/2$.

And, since W is a random variable, we can talk of *its* expected value, which turns out to be an elegant result:

1.19.5.2 The Law of Total Expectation

A property of conditional expected value, proven in many undergraduate probability texts, is

$$E(V) = EW = E[E(V \mid U)] \tag{1.58}$$

The foreboding appearance of this equation belies the fact that it is actually quite intuitive, as follows. Say you want to compute the mean height of all people in the U.S., and you already have available the mean heights in each of the 50 states. You cannot simply take the straight average of those state mean heights, because you need to give more weight to the more populous states. In other words, the national mean height is a *weighted* average of the state means, with the weight for each state being its proportion of the national population.

In (1.58), this corresponds to having V as height and U as state. State coding is an integer-valued random variable, ranging from 1 to 50, so we have

$$\begin{aligned} EV &= E[E(V \mid U)] && (1.59) \\ &= EW && (1.60) \\ &= \sum_{i=1}^{50} P(U = i)\ E(V \mid U = i) && (1.61) \end{aligned}$$

The left-hand side, EV, is the overall mean height in the nation; $E(V \mid U = i)$ is the mean height in state i; and the weights in the weighted average are the proportions of the national population in each state, $P(U = i)$.

Not only can we look at the mean of W, but also its variance. By using the various familiar properties of mean and variance, one can derive a similar relation for variance:

1.19.5.3 Law of Total Variance

For scalar V,

$$Var(V) = E[Var(V|U)] + Var[E(V|U)] \tag{1.62}$$

One might initially guess that we only need the first term. To obtain the national variance in height, we would take the weighted average of the state variances. But this would not take into account that the mean heights vary from state to state, thus also contributing to the national variance in height, hence the second term.

This is proven in Section 2.12.8.3.

1.19.5.4 Tower Property

Now consider conditioning on two variables, say U_1 and U_2. One can show that

$$E\left[E(V|U_1, U_2) \mid U_1\right] = E(V \mid U_1) \tag{1.63}$$

Here is an intuitive interpretation of that in the height example above. Take V, U_1 and U_2 to be height, state and gender, respectively, so that $E(V|U_1, U_2)$ is the mean height of all people in a certain state and of a certain gender. If we then take the mean of all these values for a certain state — i.e. take the average of the two gender-specific means in the state — we get the mean height in the state without regard to gender.

Again, note that we take the straight average of the two gender-specific means, because the two genders have equal proportions. If, say, U_2 were race instead of gender, we would need to compute a *weighted* average of the race-specific means, with the weights being the proportions of the various races in the given state.

This is proven in Section 7.8.1.

1.19.5.5 Geometric View

There is an elegant way to view all of this in terms of abstract vector spaces — (1.58) becomes the Pythagorean Theorem! — which we will address later in Mathematical Complements Sections 2.12.8 and 7.8.1.

So, we see that $\widehat{\beta}$ converges to

$$[E(XX')]^{-1}E(XY) = [E(XX')]^{-1}E(XX'\beta) = [E(XX')]^{-1}E(XX')\beta = \beta \tag{2.114}$$

2.12.7 Biased Nature of S

It was stated in Section 2.7.2 that S, even with the $n-1$ divisor, is a *biased* estimator of η, the population standard deviation. We'll derive that here.

$$\begin{aligned} 0 &< Var(S) && (2.115) \\ &= E(S^2) - (ES)^2 && (2.116) \\ &= \eta^2 - (ES)^2 && (2.117) \end{aligned}$$

since S^2 is an unbiased estimator of η^2. So,

$$ES < \eta \tag{2.118}$$

2.12.8 The Geometry of Conditional Expectation

Readers with a good grounding in vector spaces may find the material in this section helpful to their insight. It is recommended that the reader review Section 1.19.5 before continuing.[12]

2.12.8.1 Random Variables As Inner Product Spaces

Consider the set of all scalar random variables U defined in some probability space that have finite second moment, i.e. $E(U^2) < \infty$. This forms a linear space: The sum of two such random variables is another random variable with finite second moment, as is a scalar times such a random variable.

[12] It should be noted that the treatment here will not be fully mathematically rigorous. For instance, we bring in projections below, without addressing the question of the conditions for their existence.

We can define an inner product on this space. For random variables S and T in this space, define

$$(S, T) = E(ST) \tag{2.119}$$

This defines the norm

$$||S|| = (S, S)^{1/2} = \sqrt{E(S^2)} \tag{2.120}$$

So, if $ES = 0$, then

$$||S|| = \sqrt{Var(S)} \tag{2.121}$$

Many properties for regression analysis can be derived quickly from this vector space formulation. Let's start with (2.82).

The famous *Cauchy-Schwartz Inequality* for inner product spaces states that for any vectors x and y, we have

$$|(x, y)| \ \leq ||x|| \ ||y|| \tag{2.122}$$

It is left as an exercise to the reader to show that this implies (2.82).

2.12.8.2 Projections

Inner product spaces also have the notion of a *projection.* Suppose we have an inner product space $\mathcal{V}$, and subspace $\mathcal{W}$. Then for any vector x, the projection z of x onto $\mathcal{W}$ is defined to be the closest vector to x in $\mathcal{W}$. An important property is that we have a "right triangle," i.e.

$$(z, x - z) = 0 \tag{2.123}$$

We say that z and $x - z$ are *orthogonal.* And the Pythagorean Theorem holds:

$$||x||^2 = ||z||^2 + ||x - z||^2 \tag{2.124}$$

2.12.8.3 Conditional Expectations As Projections

In regression terms, the discussion in Section 1.19.3 shows that the regression function, $E(Y \mid X) = \mu(X)$ has the property that

$$\mu(X) = \text{argmin}_g \; E[(Y - g(X))^2] = \text{argmin}_g \; ||Y - g(X)||^2 \tag{2.125}$$

as **g** ranges over all functions of X. Therefore, by definition, $\mu(X)$ is the projection of Y onto the subspace consisting of all random variables with finite variance that are functions of X. This view can be very useful.

We can also use (2.124) to derive the Law of Total Variance, (1.62). For convenience in present notation, rewrite that equation as

$$Var(Y) = E[Var(Y|X)] + Var[E(Y|X)] \tag{2.126}$$

The derivation will be less cluttered if we restrict attention to the case $EY = 0$. (For the general case, define a new random variable $W = Y - EY$, and apply the mean-0 result, left as an exercise for the reader.) Note that by the Law of Total Expectation (Section 1.19.5.2), this implies the $\mu(X)$ also has mean 0.

Then (2.124) and (2.121) say that

$$Var(Y) = E[\mu(X)^2] + E\left[(Y - \mu(X))^2\right] \tag{2.127}$$

Recalling that $E\mu(X) = 0$, the first term in (2.127) is

$$Var[\mu(X)] = Var[E(Y|X)] \tag{2.128}$$

which is exactly the second term in (1.62).

Now rewrite the second term in (2.127) using (1.58):

$$\begin{aligned} E\left[(Y - \mu(X))^2\right] &= E\{E\left[(Y - \mu(X))^2 \mid X\right]\} && (2.129) \\ &= E[Var(Y|X)] && (2.130) \end{aligned}$$

And, that last expression is exactly the first term in (1.62)! So, we are done with the derivation.

2.12.9 Predicted Values and Error Terms Are Uncorrelated

Assume a random-X context, and take x in (2.123) to be Y, so in that equation

$$z = \mu(X) \tag{2.131}$$

and thus

$$E[\mu(X)(Y - \mu(X))] = 0 \tag{2.132}$$

In other words, our prediction $\mu(X)$ is uncorrelated with our prediction error, $Y - \mu(X)$.

The above concerns the population level, but a similar argument can be made at the sample level for linear models. Here we will assume a fixed-X model (conditioning on X in the random-X case), and once again use the notation of Section 2.4.2.

Define

$$\widehat{\epsilon}_i = Y_i - \widetilde{X}_i \widehat{\beta} \tag{2.133}$$

Also define $\widehat{\epsilon}$ to be the vector of the $\widehat{\epsilon}_i$.

The claim is then that the correlation between $\widehat{\epsilon}_i$ and $\widehat{\beta}_j$ is 0 for any i and j. Again, a vector space argument can be made. In this case, take the full vector space to be $\mathcal{R}^n$, the space in which D roams, and the subspace will be that spanned by the columns of A.

The vector $A\widehat{\beta}$ is in that subspace, and because $b = \widehat{\beta}$ minimizes (2.25), $A\widehat{\beta}$ is then the projection of D onto that subspace. Again, that makes $D - A\widehat{\beta}$ and $A\widehat{\beta}$ orthogonal, i.e.

$$\widehat{\epsilon}' A\widehat{\beta} = (D - A\widehat{\beta})' A\widehat{\beta} = 0 \tag{2.134}$$

Since this must hold for all A, we see that each $\widehat{\epsilon}_i$ is uncorrelated with any component of $\widehat{\beta}$.