The Minnesota Journal of Undergraduate Mathematics

On Least Squares Linear Regression Without Second Moment

Rajeshwari Majumdar

University of Connecticut

The Minnesota Journal of Undergraduate Mathematics

Volume 4 (2018-19 Academic Year)

Sponsored by
School of Mathematics
University of Minnesota
Minneapolis, MN 55455

The Minnesota Journal of Undergraduate Mathematics

Volume 4 (2018-19 Academic Year)

On Least Squares Linear Regression Without Second Moment

Rajeshwari Majumdar

University of Connecticut

Abstract. If X and Y are real valued random variables such that the first moments of X, Y, and XY exist and the conditional expectation of Y given X is an affine function of X, then the intercept and slope of the conditional expectation equal the intercept and slope of the least squares linear regression function, even though Y may not have a finite second moment. As a consequence, the affine in X form of the conditional expectation and zero covariance imply mean independence.

1. Introduction

If X and Y are real valued random variables such that the conditional expectation of Y given X is an affine function of X, then the intercept and slope of the conditional expectation equal, respectively, the intercept and slope of the least squares linear regression function. As explained in Remark 2.8, when both X and Y have finite second moments, this equality follows from the well-understood connection among conditional expectation, least squares linear regression, and the operation of projection. However, that this equality continues to hold when one only assumes that the first moments of X, Y, and XY exist is the most important finding of this note [Theorem 2.6].

Theorem 2.6 evolves from the investigation of the directional hierarchy of the interdependence among the notions of independence, mean independence, and zero covariance. It is well-known that, for random variables X and Y, independence implies mean independence and mean independence implies zero covariance, whenever the notions of mean independence and covariance make sense. Note that, the notion of covariance makes sense as long as the first moments of X, Y, and XY exist; it does not require X and Y to have finite second moments. We review well-known counterexamples to document that the direction of this hierarchy cannot be reversed in general. Theorem 2.6, above and beyond establishing that an affine in X form of the conditional expectation of Y given X implies its equality with the least squares linear regression function, also leads to the conclusion that mean independence is necessary and sufficient for zero covariance when the conditional expectation is affine in X.

* *Corresponding author*

Remark 2.9 examines the feasibility of obtaining the result of Theorem 2.6 by using the projection operator interpretation of conditional expectation and least squares linear regression elucidated in Remark 2.8 and the technique of extending operators to the closure of their domains, and concludes in the negative. Remark 2.10 explains how the relaxation of the assumption of Y having a finite second moment is non-trivial.

The following notational conventions are used throughout the note. Equality (or inequality) involving measurable functions defined on a probability space, unless otherwise indicated, indicates that the relation holds almost surely. The universal null set is denoted by $\mathfrak{K}$. The normal distribution with mean μ and variance σ^2 is denoted by $\mathcal{N}\left(\mu,\sigma^2\right)$.

2. Results

Let $(\Omega,\mathcal{F},P)$ be an arbitrary probability space. Let $\mathcal{L}_1$ (respectively, $\mathcal{L}_2$) denote the Banach (respectively, Hilbert) space of integrable (respectively, square integrable) real valued functions on $(\Omega,\mathcal{F},P)$. Let E denote the expectation induced by P and $\mathrm{E}^{\mathfrak{C}}$ the conditional expectation given a sub σ-algebra $\mathfrak{C}$ of $\mathcal{F}$. In what follows, we repeatedly use the *averaging property*

$$\mathrm{E}\left(\mathrm{E}^{\mathfrak{C}}(Z)I_{\mathcal{A}}\right)=\mathrm{E}(ZI_{\mathcal{A}}) \text{ for every } Z\in\mathcal{L}_1 \text{ and } \mathcal{A}\in\mathfrak{C},$$

the *pull-out property*

$$\mathrm{E}^{\mathfrak{C}}(ZT)=T\mathrm{E}^{\mathfrak{C}}(Z) \text{ if } Z\in\mathcal{L}_1,\ ZT\in\mathcal{L}_1, \text{ and } T \text{ is } \mathfrak{C}\text{-measurable},$$

and the *chain rule*

$$\mathrm{E}^{\mathfrak{B}}\left(\mathrm{E}^{\mathfrak{C}}(Z)\right)=\mathrm{E}^{\mathfrak{B}}(Z) \text{ if } \mathfrak{B} \text{ is a sub } \sigma\text{-algebra of } \mathfrak{C};$$

see [4, p. 105]. If $\mathfrak{C}=\sigma(U)$ for some random variable U defined on $(\Omega,\mathcal{F},P)$, we write E^U in place of $\mathrm{E}^{\mathfrak{C}}$. Recall that

$$\sigma(U)=\left\{U^{-1}(A): A \text{ is a Borel subset of } \mathbb{R}\right\}.$$

If $X,Y,XY\in\mathcal{L}_1$, and X and Y are independent, then

$$\mathrm{Cov}(X,Y)=0. \tag{1}$$

As is well-known, the reverse implication is not true in general; see Example 2.1.

Example 2.1. Let $X\sim\mathcal{N}(0,1)$ be independent of the Rademacher random variable W, defined by $P(W=-1)=P(W=1)=1/2$. Let $Y=WX$; then $Y\sim\mathcal{N}(0,1)$ and (1) holds. However, X and Y are not independent; because if they were, then $X+Y$ would have the $\mathcal{N}(0,2)$ distribution, implying $P(X+Y=0)=0$, whereas in actuality $2P(X+Y=0)=1$.

Definition 2.2. If $Y\in\mathcal{L}_1$ and

$$\mathrm{E}^X(Y)=\mathrm{E}(Y), \tag{2}$$

Y is said to be *mean independent* of X. Similarly, if $X\in\mathcal{L}_1$ and

$$\mathrm{E}^Y(X)=\mathrm{E}(X), \tag{3}$$

X is said to be *mean independent* of Y. Example 2.3 shows that Y can be mean independent of X without X being mean independent of Y.

Example 2.3. Consider the equi-probable discrete sample space $\Omega = \{-1,0,1\}$ and define random variables X and Y on Ω as $X(\omega) = I_{[\omega=0]}(\omega)$ and $Y(\omega) = \omega$. Since $\sigma(X) = \{\mathfrak{K}, \{0\}, \{-1,1\}, \Omega\}$ and $\mathrm{E}(YI_A)$ is trivially equal to 0 for all $A \in \sigma(X)$, we obtain $\mathrm{E}^X(Y) = 0 = \mathrm{E}(Y)$. However, since $X(\omega) = I_{[Y(\omega)=0]}(\omega)$, X is $\sigma(Y)$ measurable, and consequently $\mathrm{E}^Y(X) = X$, whereas $\mathrm{E}(X) = 1/3$.

It follows from the definition of independence and conditional expectation [2, p. 264] that for X and Y independent, (2) holds if $Y \in \mathcal{L}_1$, whereas (3) holds if $X \in \mathcal{L}_1$. The asymmetric nature of the notion of mean independence established in Example 2.3 shows that (2) (or, for that matter, (3)) does not imply independence of X and Y. Example 2.4 extends Example 2.1 to show that even (2) and (3) combined do not necessarily imply independence of X and Y.

Example 2.4. Let X, W, and Y be as in Example 2.1. Since X and W are independent and $-X \sim \mathcal{N}(0,1)$, by [1, Corollary 7.1.2],

$$\mathrm{E}\left(XI_{[Y\in B]}\right) = \frac{1}{2}\mathrm{E}\left(XI_{[X\in B]}\right) + \frac{1}{2}\mathrm{E}\left(XI_{[-X\in B]}\right) = 0,$$

implying that $\mathrm{E}^Y(X) = 0 = \mathrm{E}(X)$. Clearly, by the pull-out property,

$$\mathrm{E}^X(Y) = \mathrm{E}^X(WX) = X\mathrm{E}^X(W) = X\mathrm{E}(W) = 0 = \mathrm{E}(Y).$$

However, as observed in Example 2.1, X and Y are not independent.

As mentioned in the introduction, that mean independence implies zero covariance is well-known. Example 2.5 shows that zero covariance, that is (1), does not necessarily imply mean independence, as in (2).

Example 2.5. Let X be uniformly distributed over the interval $(-1,1)$ and $Y = X^2$. Since $\mathrm{E}(X) = 0 = \mathrm{E}\left(X^3\right) = \mathrm{E}(XY)$, (1) holds. Since $\mathrm{E}(Y) = \mathrm{E}\left(X^2\right) = \mathrm{Var}(X) = 1/3$ and $\mathrm{E}^X(Y) = \mathrm{E}^X\left(X^2\right) = X^2$, (2) does not hold.

Theorem 2.6 shows that if Y is mean independent of X (as in (2)), then (1) holds; it also characterizes the setup wherein the reverse implication holds.

Theorem 2.6. *Assume that $X, Y, XY \in \mathcal{L}_1$. Then, the following four conclusions hold:*

(i) $\mathrm{Var}(X)$ *is well defined, though it may be* ∞*; that is,* $\mathrm{Var}(X) \in [0,\infty]$.

(ii) If Y is mean independent of X (as in (2)*), then* (1) *holds; also,*

$$\mathrm{E}^X(Y) = \alpha + \beta X \tag{4}$$

for some $\alpha, \beta \in \mathbb{R}$, which are unique if $\mathrm{Var}(X) > 0$.

(iii) The affine form of $\mathrm{E}^X(Y)$ *in* (4) *implies*

$$\alpha = \mathrm{E}(Y) - \beta\mathrm{E}(X), \tag{5}$$

$$\mathrm{E}^X(Y) = \mathrm{E}(Y) + \beta(X - \mathrm{E}(X)), \tag{6}$$

and

$$\beta = \begin{cases} 0 & \textit{if } \mathrm{Var}(X) = 0 \\ \frac{Cov(X,Y)}{\mathrm{Var}(X)} & \textit{if } 0 < \mathrm{Var}(X) < \infty \\ 0 & \textit{if } \mathrm{Var}(X) = \infty. \end{cases} \tag{7}$$

(iv) If (4) *holds, then* (1) *implies* (2).

Remark 2.7. Clearly, if X is mean independent of Y as in (3), then (1) holds as well. Also, going back to Example 2.5, we now know why (1) does not imply (2) in that context; since $\mathrm{E}^X(Y) = X^2$, (4) does not hold, and by part (ii) of Theorem 2.6, (2) cannot hold.

We now present the proof of Theorem 2.6.

Proof. Since $X \in \mathcal{L}_1$, the assertion of part (i) is vacuously true.

The equality

$$\mathrm{E}(XY) = \mathrm{E}\left(\mathrm{E}^X(XY)\right) = \mathrm{E}\left(X\mathrm{E}^X(Y)\right) \tag{8}$$

is used in the proofs of both parts (ii) and (iii); it is substantiated by the averaging and pull-out properties.

If (2) holds, then (1) follows from (8). Clearly, (4) holds with $\alpha = \mathrm{E}(Y)$ and $\beta = 0$. If $\mathrm{E}^X(Y) = \alpha' + \beta' X$, then $\mathrm{E}(Y) = \alpha' + \beta' \mathrm{E}(X)$ by the averaging property, whence (2) implies

$$\beta'(X - \mathrm{E}(X)) = 0; \tag{9}$$

when $\mathrm{Var}(X) > 0$, (9) implies $\beta' = 0$ and consequently $\alpha' = \mathrm{E}(Y)$. That completes the proof of part (ii).

Now assume (4) holds. Taking expectations of both sides,

$$\mathrm{E}(Y) = \alpha + \beta \mathrm{E}(X), \tag{10}$$

whence (5) follows. The equality in (6) is a straightforward consequence of (4) and (5). We lay out the groundwork for proving (7) in the next paragraph.

By the Conditional Jensen's inequality [2, Theorem 10.2.7], we obtain $\left|\mathrm{E}^X(Y)\right| \le \mathrm{E}^X(|Y|)$, implying $\left|X\mathrm{E}^X(Y)\right| \le |X|\mathrm{E}^X(|Y|)$. By the averaging and pull-out properties, $\mathrm{E}(|XY|) = \mathrm{E}\left(|X|\mathrm{E}^X(|Y|)\right)$. Since $XY \in \mathcal{L}_1$, it follows $X\mathrm{E}^X(Y) \in \mathcal{L}_1$ and, by (4), $X(\alpha + \beta X) \in \mathcal{L}_1$. Since $X \in \mathcal{L}_1$, $\alpha X \in \mathcal{L}_1$ for every $\alpha \in \mathbb{R}$; consequently,

$$\beta X^2 \in \mathcal{L}_1. \tag{11}$$

To prove (7), we consider the three cases separately.

If $\mathrm{Var}(X) = \infty$, that is, $X \notin \mathcal{L}_2$, then $\beta = 0$ by (11).

If $0 < \mathrm{Var}(X) < \infty$, then $X \in \mathcal{L}_2$. By (8) and (4),

$$\mathrm{E}(XY) = \mathrm{E}(X(\alpha + \beta X)) = \alpha \mathrm{E}(X) + \beta \mathrm{E}\left(X^2\right). \tag{12}$$

Multiplying both sides of (10) by $\mathrm{E}(X)$,

$$\mathrm{E}(X)\mathrm{E}(Y) = \alpha \mathrm{E}(X) + \beta (\mathrm{E}(X))^2. \tag{13}$$

Subtracting (13) from (12),

$$\operatorname{Cov}(X,Y) = \beta \operatorname{Var}(X),$$

whence $\beta = \frac{\operatorname{Cov}(X,Y)}{\operatorname{Var}(X)}$.

Note that $\operatorname{Var}(X) = 0$ implies $X = \mathrm{E}(X)$ and consequently, $\operatorname{Cov}(X,Y) = 0$, that is, (1) holds. Since $X = \mathrm{E}(X)$, for any Borel subset A of $\mathbb{R}$,

$$P([X \in A]) = \begin{cases} 0 & \text{if } \mathrm{E}(X) \notin A \\ 1 & \text{if } \mathrm{E}(X) \in A, \end{cases}$$

whence

$$\mathrm{E}\left(\mathrm{E}^X(Y) I_{[X\in A]}\right) = \begin{cases} 0 & \text{if } \mathrm{E}(X) \notin A \\ \mathrm{E}(Y) & \text{if } \mathrm{E}(X) \in A \end{cases} = \mathrm{E}\left(\mathrm{E}(Y) I_{[X\in A]}\right),$$

and (2) follows from the definition of conditional expectation, without using the assumption that (4) holds. As in the proof of part (ii), (2) implies (4) with $\alpha = \mathrm{E}(Y)$ and $\beta = 0$. However, since $\operatorname{Var}(X) = 0$, $(\mathrm{E}(Y), 0)$ is not the unique choice for (α, β) (see (9)). In other words, when $\operatorname{Var}(X) = 0$, we cannot algebraically conclude from (4) that $\beta = 0$ (α and β can be anything subject to (10)). That said, since $\operatorname{Var}(X) = 0$ implies that (2) holds, the canonical choice of parameters in the affine model for $\mathrm{E}^X(Y)$ specified in (4) is $\alpha = \mathrm{E}(Y)$ and $\beta = 0$. That completes the proof of part (iii).

The conclusion of part (iv) follows from (6) and (7). As noted in the preceeding paragraph, the validity of (2) when $\operatorname{Var}(X) = 0$ does not depend on the canonical choice of parameters. $\square$

Remark 2.8. Is there an element of surprise in the conclusion of part (iii) of Theorem 2.6 which asserts that (4) implies (5) and (7)? The answer is no when $X, Y \in \mathcal{L}_2$, since in that case it follows from the well-understood connection (outlined below) among conditional expectation, least squares linear regression, and the operation of projection.

For a fixed real valued measurable function X on $(\Omega, \mathcal{F}, P)$, let $\mathcal{L}_2(X)$ denote the Hilbert space of real valued measurable functions f on $(\Omega, \sigma(X), P)$ such that $\mathrm{E}\left(f^2\right) < \infty$. Clearly, $\mathcal{L}_2(X) \subset \mathcal{L}_2$. Let $\|\cdot\|_2$ denote the $\mathcal{L}_2$ norm on $\mathcal{L}_2$, and by inheritance, on $\mathcal{L}_2(X)$. Define

$$\mathcal{M}_2 = \{f \in \mathcal{L}_2 : \text{for some } g \in \mathcal{L}_2(X),\ f = g \text{ outside of a } P\text{-null set in } \mathcal{F}\}.$$

Clearly, $\mathcal{M}_2$ is a subspace of $\mathcal{L}_2$ that contains $\mathcal{L}_2(X)$. If $\{f_n : n \geq 1\}$ is a sequence in $\mathcal{M}_2$ that converges to $f \in \mathcal{L}_2$, then, since every convergent sequence is Cauchy, $\|f_n - f_m\|_2 \to 0$ as $m, n \to \infty$. By definition of $\mathcal{M}_2$, there exists a sequence $\{g_n : n \geq 1\}$ in $\mathcal{L}_2(X)$ such that $\|f_n - f_m\|_2 = \|g_n - g_m\|_2$, implying that $\{g_n : n \geq 1\}$ is a Cauchy sequence in $\mathcal{L}_2(X)$. Since $\mathcal{L}_2(X)$ is complete, there exists $g \in \mathcal{L}_2(X)$ such that $\|g_n - g\|_2 \to 0$ as $n \to \infty$. Since $\|f_n - g_n\|_2 = 0$ for every $n \geq 1$, by the triangle inequality in $\mathcal{L}_2$, $\|f - g\|_2 = 0$, showing that $f \in \mathcal{M}_2$, that is, $\mathcal{M}_2$ is closed in $\mathcal{L}_2$.

Since $\left(\mathrm{E}^X(h)\right)^2 \leq \mathrm{E}^X\left(h^2\right)$ by the Conditional Jensen's Inequality, by the averaging property $\mathrm{E}^X(h) \in \mathcal{L}_2(X)$ if $h \in \mathcal{L}_2$. Let T denote the map from $\mathcal{L}_2$ to $\mathcal{L}_2(X)$ that takes $h \in \mathcal{L}_2$ to $\mathrm{E}^X(h) \in \mathcal{L}_2(X)$. Note that, T depends on X, but we suppress that dependence for notational convenience. Clearly, T is a linear map. Since $\left(\mathrm{E}^X(h)\right)^2 \leq \mathrm{E}^X\left(h^2\right)$, $\|Th\|_2 \leq \|h\|_2$ by the averaging property. Thus, T is a contraction operator.

For any $g \in \mathcal{L}_2(X)$ and $h \in \mathcal{L}_2$, using the pull-out property,

$$\left\langle h - Th, Th - g \right\rangle = 0,$$

where $\left\langle \cdot,\cdot \right\rangle$ denotes the inner product in $\mathcal{L}_2$, and consequently,

$$\|h-g\|_2^2 = \|h-Th\|_2^2 + \|Th-g\|_2^2,$$

showing that Th is the unique minimizer of $\|h-g\|_2$ as g varies over $\mathcal{L}_2(X)$.

Now recall that $\mathcal{M}_2$ is a closed subspace of $\mathcal{L}_2$; let the orthogonal projection from $\mathcal{L}_2$ to $\mathcal{M}_2$ be denoted by $\Pi_{\mathcal{M}_2}$. Then, for any $f \in \mathcal{M}_2$ and $h \in \mathcal{L}_2$,

$$\|h-f\|_2^2 = \left\|h - \Pi_{\mathcal{M}_2} h\right\|_2^2 + \left\|\Pi_{\mathcal{M}_2} h - f\right\|_2^2,$$

showing that $\Pi_{\mathcal{M}_2} h$ is the unique minimizer of $\|h-f\|_2$ as f varies over $\mathcal{M}_2$.

Since

$$\{\|h-g\|_2 : g \in \mathcal{L}_2(X)\} = \{\|h-f\|_2 : f \in \mathcal{M}_2\},$$

$\Pi_{\mathcal{M}_2} h$ equals Th outside of a P-null set in $\mathcal{F}$; that is, as elements of $\mathcal{L}_2$, $\Pi_{\mathcal{M}_2} h = Th$.

Note that if the probability space $(\Omega, \sigma(X), P)$ is complete, so that the almost sure limit of a sequence of measurable functions is measurable, $\mathcal{L}_2(X)$ becomes a closed subspace of $\mathcal{L}_2$, and in our identification of the conditional expectation as a projection, we can avoid the construction involving $\mathcal{M}_2$.

Let $\mathcal{H}$ denote the two-dimensional linear space spanned by J and X, where J is the real valued function defined on $(\Omega, \mathcal{F}, P)$ that is almost surely equal to 1; since $X \in \mathcal{L}_2$, we obtain $\mathcal{H} \subset \mathcal{L}_2(X) \subset \mathcal{M}_2$.

Let $\mathcal{A}$ denote the subspace of $\mathcal{L}_2$ that consists of all $f \in \mathcal{L}_2$ such that $Tf \in \mathcal{H}$, that is, (4) holds for f.

Using $\Pi_{\mathcal{M}_2} h = Th$ for every $h \in \mathcal{L}_2$ and $\mathcal{H} \subset \mathcal{L}_2(X) \subset \mathcal{M}_2$, for $f \in \mathcal{A}$ we obtain

$$Tf = \Pi_{\mathcal{M}_2} f = \Pi_{\mathcal{H}}\left(\Pi_{\mathcal{M}_2} f\right) = \Pi_{\mathcal{H}} f, \tag{14}$$

where $\Pi_{\mathcal{H}}$ denotes the orthogonal projection from $\mathcal{L}_2$ to $\mathcal{H}$.

Since $X \in \mathcal{L}_2$, we obtain $\mathrm{Var}(X) < \infty$. If $\mathrm{Var}(X) > 0$, equivalently, X and J are linearly independent, applying the Gram-Schmidt orthonormalization process to the basis $\{J, X\}$ of $\mathcal{H}$ we obtain that $\{J, X^*\}$, where

$$X^* = \frac{X - \left\langle J, X \right\rangle J}{\left\| X - \left\langle J, X \right\rangle J \right\|_2},$$

is an orthonormal basis of $\mathcal{H}$. Consequently, by (14),

$$TY = \left\langle J, Y \right\rangle J + \left\langle X^*, Y \right\rangle X^*,$$

which, in more familiar notation, asserts

$$\mathrm{E}^X(Y) = \mathrm{E}(Y) + \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)}(X - \mathrm{E}(X)).$$

If $\mathrm{Var}(X) = 0$, equivalently, X and J are linearly dependent, $\mathcal{H}$ is simply the span of J, implying $TY = \langle J, Y \rangle J$, that is, $\mathrm{E}^X(Y) = \mathrm{E}(Y)$, leading to the conclusion of part (iii) of Theorem 2.6 for $X, Y \in \mathcal{L}_2$.

Remark 2.9. Can the conclusion of part (iii) of Theorem 2.6, when we only have $Y \in \mathcal{L}_1 \backslash \mathcal{L}_2$, $X \in \mathcal{L}_1$, and $XY \in \mathcal{L}_1$, be obtained using the projection operator tools employed in Remark 2.8? The answer, as far as we can tell, is no.

The domain of the map T can be expanded to $\mathcal{L}_1$, causing the range to be expanded to $\mathcal{L}_1(X)$, the Banach space of real valued measurable functions f on $(\Omega, \sigma(X), P)$ such that $\mathrm{E}(|f|) < \infty$. The linearity of T is not impacted by this expansion of domain. Since $\left|\mathrm{E}^X(h)\right| \leq \mathrm{E}^X(|h|)$ (again by the Conditional Jensen's Inequality), $\|Th\|_1 \leq \|h\|_1$ by the averaging property, where $\|\cdot\|_1$ denotes the $\mathcal{L}_1$ norm on $\mathcal{L}_1$, and by inheritance, on $\mathcal{L}_1(X)$. Thus, T remains a contraction operator on $\mathcal{L}_1$. In fact, as observed by [4], T on $\mathcal{L}_2$ is uniformly $\mathcal{L}_1$-continuous and its extension to a linear and continuous map on $\mathcal{L}_1$ is unique up to almost sure equivalence.

The definition of $\mathcal{A}$ can be extended to denote the subspace of $\mathcal{L}_1$ that consists of all $f \in \mathcal{L}_1$ such that $Tf \in \mathcal{H}$. Since $\mathcal{H}$ is a finite-dimensional, hence closed, subspace of $\mathcal{L}_1$, and T is continuous, $\mathcal{A}$ is a closed subspace of $\mathcal{L}_1$. Part (iii) of Theorem 2.6 asserts that the representation of T as the orthogonal projection to $\mathcal{H}$ (for $X \in \mathcal{L}_2$) that is valid on $\mathcal{A} \cap \mathcal{L}_2$ can be extended to hold on $\mathcal{A} \cap \mathcal{B}$, where $\mathcal{B}$ denotes the subspace of $\mathcal{L}_1$ that consists of all $f \in \mathcal{L}_1$ such that $Xf \in \mathcal{L}_1$. For that conclusion to be drawn using the technique of operator extension, we need to have the closure of $\mathcal{A} \cap \mathcal{L}_2$ in $\mathcal{L}_1$ equal $\mathcal{A} \cap \mathcal{B}$. However, even if we assume X is bounded, we can only conclude that $\mathcal{B} = \mathcal{L}_1$ and $\mathcal{A} \cap \mathcal{B} = \mathcal{A}$, implying that the closure of $\mathcal{A} \cap \mathcal{L}_2$ in $\mathcal{L}_1$ is contained in $\mathcal{A} \cap \mathcal{B}$ (since $\mathcal{A}$ is a closed subspace of $\mathcal{L}_1$), but the reverse inclusion is not necessarily true.

Remark 2.10. What does the relaxation of the structural assumption from $X, Y \in \mathcal{L}_2$ to $Y \in \mathcal{L}_1 \backslash \mathcal{L}_2$, $X \in \mathcal{L}_1$, and $XY \in \mathcal{L}_1$ entail? One can conceivably argue that the Cauchy-Schwartz inequality remains the primary tool for verifying that $XY \in \mathcal{L}_1$ when X and Y are dependent random variables, and as such, this relaxation of assumption is neither insightful nor useful. While that argument may have some merit, we would like to point out that if X is bounded, then obviously $X \in \mathcal{L}_1$, and $Y \in \mathcal{L}_1 \backslash \mathcal{L}_2$ implies $XY \in \mathcal{L}_1$.

A classic example of (4) holding for a bounded random variable X is the Bernoulli random variable, since, for any measurable function $h(X)$ of the Bernoulli random variable X, we have $h(X) = h(0) + (h(1) - h(0))X$, and $\mathrm{E}^X(Y)$ is a measurable function of X.

Working with the non-central t-distribution with 2 degrees of freedom, the following example presents X, Y such that $X \in \mathcal{L}_1$ but is not bounded, $Y \in \mathcal{L}_1 \backslash \mathcal{L}_2$, $XY \in \mathcal{L}_1$, and (4) holds. Let $X \sim \mathcal{N}(0,1)$ and given X, $W \sim \mathcal{N}(X,1)$. Let $V \sim \chi_2^2$ be independent of (W, X). Let $Y = W/\sqrt{V/2}$.

Clearly, $X \in \mathcal{L}_1$ but is not bounded.

To verify that $Y \in \mathcal{L}_1 \backslash \mathcal{L}_2$, we first obtain the marginal distribution of W. Using the form of the conditional density of W given $X = x$ and the form of the marginal density of X, we obtain that the joint density of (W,X) is given by

$$f(w,x)=\frac{\exp\left(-\frac{w^2}{2}+wx-x^2\right)}{2\pi};$$

since $x\mapsto \exp\left(-(x-w/2)^2\right)\Big/\sqrt{\pi}$ represents the density of the $\mathcal{N}\left(\frac{w}{2},\frac{1}{2}\right)$ distribution, completing the square in x we obtain the marginal density of W to be

$$f_W(w)=\int_{\mathbb{R}} f(w,x)\,dx=\frac{\exp\left(-\frac{w^2}{4}\right)}{\sqrt{2\pi}\sqrt{2}}\int_{\mathbb{R}}\frac{\exp\left(-\left(x-\frac{w}{2}\right)^2\right)}{\sqrt{\pi}}dx=\frac{\exp\left(-\frac{w^2}{4}\right)}{\sqrt{2\pi}\sqrt{2}},$$

showing that $W\sim\mathcal{N}(0,2)$. Since $V=2U$, where U is a $\Gamma(1)$ random variable [3, Definition 3.3.6], by [3, Theorem 3.3.9],

$$\mathrm{E}\left(\frac{1}{\sqrt{V/2}}\right)=\mathrm{E}\left(U^{-\frac{1}{2}}\right)=\frac{\Gamma\left(\frac{1}{2}\right)}{\Gamma(1)}=\sqrt{\pi}; \tag{15}$$

since W and V are independent, $Y\in\mathcal{L}_1$. However, $\mathrm{E}(2/V)=\mathrm{E}\left(U^{-1}\right)=\infty$, showing that $Y\notin\mathcal{L}_2$.

To show that $XY\in\mathcal{L}_1$, by independence of (W,X) and V, along with (15), it suffices to show that $XW\in\mathcal{L}_1$, which follows readily from the Cauchy-Schwartz inequality.

Finally, by [1, Corollary 7.1.2] and (15),

$$\mathrm{E}^{W,X}(Y)=\sqrt{\pi}W,$$

implying, by the chain rule, $\mathrm{E}^X(Y)=\sqrt{\pi}X$, that is, (4) holds.

3. Acknowledgments

I would like to thank Professor Jungbin Hwang, whose graduate econometrics seminar got me started on this research, for carefully reading multiple drafts of this paper and providing helpful suggestions.

References

[1] Chow, Y. S., and Teicher, H. *Probability Theory, Independence, Interchangeability, Martingales.* Springer-Verlag, New York, 2nd Ed. (1988).
[2] Dudley, R. M. *Real Analysis and Probability.* Wadsworth & Brooks/Cole, Pacific Grove, 1st Ed. (1989).
[3] Fabian, V., and Hannan, J. *Introduction to Probability and Mathematical Statistics.* Wiley, New York (1985).
[4] Kallenberg, O. *Foundations of Modern Probability.* Springer-Verlag, New York, 2nd Ed. (2002).

Student biographies

Rajeshwari Majumdar: (*Corresponding author:* rajeshwari.majumdar@uconn.edu) Rajeshwari Majumdar recently graduated from the University of Connecticut with majors in Economics, Mathematics-Statistics, and Political Science. She will begin graduate school at New York University in the fall.