Some Pleasant OLS Arithmetic

Paul Bousquet*

April 13, 2026

Abstract

This paper expands the representation of the least squares regression estimand to make clear its dependencies on the moments of the outcome variable and included regressors. Estimands on one regressor can be written as a linear combination of the covariance between the outcome and all other regressors or as a weighted average of the marginal effects of any given regressor without much loss of generality. A special case shows an implication for testing whether the marginal effects of one regressor depend on the level of another.

*Department of Economics, University of Virginia. Email: pbousquet@virginia.edu.

1 Introduction

Let $\mathbf{D} = (Y, X_1, \dots, X_K) = (Y, \mathbf{X})$ be a random vector with mean zero and assume $\{X_k\}_{k=1}^K$ are linearly independent. A linear projection of Y on $\mathbf{X}$ satisfies

$$Y = \sum_{k=1}^{K} \beta_k X_k + u,$$

where $\mathrm{Cov}(u, X_k) = 0 \ \forall k$. Multiplying by X_j and taking expectations yields

$$\mathrm{Cov}(Y, X_j) = \sum_{k=1}^{K} \beta_k \mathrm{Cov}(X_k, X_j).$$

This motivates a representation of the linear projection estimand of

$$\beta_j = \frac{\det(\Sigma^j)}{\det(\Sigma)},$$

where $\Sigma = \mathrm{E}[\mathbf{X}\mathbf{X}']$ and Σ^j replaces the j-th column of Σ with $\mathrm{E}[\mathbf{X}Y]$. It's apparent from this form that β_j is just a linear combination of the first and second moments of $\mathbf{D}$, but it obscures the exact nature of the internal dependencies. To fix ideas, let the conditional mean function be $m_k(x) = \mathrm{E}[Y \mid X_k = x]$. Under mild regularity conditions on the data generating process for $\mathbf{D}$ (**?**), we can equally write the linear projection estimand as

$$\beta_j = \int \omega_j(x) \cdot m_j'(x) \, \mathrm{d}x \tag{1.1}$$

$$\text{with} \ \ \omega_j(x) = \frac{\mathrm{Cov}(\mathbf{1}_{\{x \leq X_j\}}, X_j^{\perp})}{\mathrm{Var}(X_j^{\perp})}, \tag{1.2}$$

where $X_j^{\perp}$ is the residuals from projecting X_j on the remaining elements of $\mathbf{X}$. So whenever X_j is included in a regression, the estimand β_j will be a weighted average of the derivative of the conditional mean function of Y with respect to X_j. But absent further assumptions, the weights on β_j depend on the other elements of $\mathbf{X}$ that are included in the projection. And the exact nature of this dependency is unclear.

This paper provides three illustrations of the interdependence of included regressors. The first takes the Fresh-Waugh-Lovell (FWL) form of β_j and continues expanding until there are no remaining projection residuals. These steps are illuminating for the case where we want the estimand weights $\omega_j(\cdot)$ to take on a particular form, making explicit level of complexity the general case imposes. The second illustration shows how to write β_j as a weighted average of derivatives with respect to $\mathbf{X}$, not just X_j. This representation underscores the sensitivity of estimates to assumptions about the relationships within $\mathbf{X}$. The third illustration deals with an initial simplification where X_j is assumed to be independent of the remaining elements of $\mathbf{X}$. Including simple interaction terms as regressors can test whether the marginal effects of X_j depend on the other elements of $\mathbf{X}$.

Assumptions: Relevant integrability and regularity conditions (e.g., differentiability) are assumed throughout the main text and are provided formally in the Appendix.

2 Unpacking FWL

Define

$$C_j^Y = \mathrm{Cov}(Y, X_j), \qquad C_{jk} = \mathrm{Cov}(X_j, X_k).$$

For each $j = 1, \ldots, K$, there is a linear restriction on the system of projection coefficients

$$C_j^Y = \beta_1 C_{1j} + \beta_2 C_{2j} + \cdots + \beta_K C_{Kj}. \tag{2.1}$$

Define $X_j^{\perp}$ as the linear projection residuals from X_j on $\{X_k\}_{k=1,k\neq j}^K$. We know that $\beta_j = \frac{\mathrm{Cov}(Y, X_j^{\perp})}{\mathrm{Var}(X_j^{\perp})}$ by the Frisch–Waugh–Lovell (FWL) Theorem. Inside $X_j^{\perp}$ are other projection coefficients with their own FWL representation. Eventually, there will be no remaining regressors to partial out.

The goal is to isolate β_k using scalar elimination to represent it purely in terms of covariances and variances, without any projection residuals. For indexing simplicity, results are presented for β_1. For each $m = 0, 1, \ldots, K-1$, let $S_m = \{1, m+2, \ldots, K\}$, meaning $S_0 = \{1, \ldots, K\}$, $S_1 = \{1, 3, \ldots, K\}$, and so on, until $S_{K-1} = \{1\}$. For $m = 1, \ldots, K-1$ and for all $j, k \in S_m$, define

$$R_j^{(m)} = R_j^{(m-1)} - \frac{D_{j,m+1}^{(m-1)}}{D_{m+1,m+1}^{(m-1)}} R_{m+1}^{(m-1)}, \tag{2.2}$$

$$D_{jk}^{(m)} = D_{jk}^{(m-1)} - \frac{D_{j,m+1}^{(m-1)} D_{m+1,k}^{(m-1)}}{D_{m+1,m+1}^{(m-1)}}, \tag{2.3}$$

and for $m = 0$ set $R_j^{(0)} = C_j^Y, D_{jk}^{(0)} = C_{jk}$. We want to show that for every $m = 0, 1, \ldots, K-1$ and every $j \in S_m$,

$$R_j^{(m)} = \sum_{k \in S_m} \beta_k D_{jk}^{(m)}. \tag{2.4}$$

We proceed by induction on m. For $m = 0$, equation (2.4) is exactly (2.1). Now suppose (2.4) holds for some $m-1$, where $1 \leq m \leq K-1$. Fix $j \in S_m$. Since $j \in S_{m-1}$ and $m+1 \in S_{m-1}$,

$$R_j^{(m-1)} = \sum_{k \in S_{m-1}} \beta_k D_{jk}^{(m-1)}, \tag{2.5}$$

$$R_{m+1}^{(m-1)} = \sum_{k \in S_{m-1}} \beta_k D_{m+1,k}^{(m-1)}. \tag{2.6}$$

Substituting (2.5) and (2.6) into the definition of $R_j^{(m)}$,

$$R_j^{(m)} = \sum_{k \in S_{m-1}} \beta_k D_{jk}^{(m-1)} - \frac{D_{j,m+1}^{(m-1)}}{D_{m+1,m+1}^{(m-1)}} \sum_{k \in S_{m-1}} \beta_k D_{m+1,k}^{(m-1)}.$$

Combining the sums,

$$R_j^{(m)} = \sum_{k \in S_{m-1}} \beta_k \left[D_{jk}^{(m-1)} - \frac{D_{j,m+1}^{(m-1)} D_{m+1,k}^{(m-1)}}{D_{m+1,m+1}^{(m-1)}} \right]. \tag{2.7}$$

Separate out the term $k = m + 1$. Its coefficient is

$$D_{j,m+1}^{(m-1)} - \frac{D_{j,m+1}^{(m-1)} D_{m+1,m+1}^{(m-1)}}{D_{m+1,m+1}^{(m-1)}} = 0.$$

So the β_{m+1} term is eliminated. For every remaining $k \in S_m$, the bracketed term in (2.7) is precisely $D_{jk}^{(m)}$ by definition (2.3), giving the result ■. Setting $m = K - 1$, we have $S_{K-1} = \{1\}$, so by (2.4)

$$\beta_1 = \frac{R_1^{(K-1)}}{D_{11}^{(K-1)}}. \tag{2.8}$$

Now we want to show $R_1^{(K-1)}$ is a linear combination of C_j^Y, the covariances between Y and X_j. For $m = 1, \ldots, K-1$, $j \in S_m$, $\ell = 1, \ldots, K$, define weights $W_{j\ell}^{(m)}$ by

$$W_{j\ell}^{(m)} = W_{j\ell}^{(m-1)} - \frac{D_{j,m+1}^{(m-1)}}{D_{m+1,m+1}^{(m-1)}} W_{m+1,\ell}^{(m-1)}, \tag{2.9}$$

and for $m = 0$ set $W_{j\ell}^{(0)} = \mathbf{1}\{j = \ell\}$. The formal claim to be proved is that for every $m = 0, 1, \ldots, K-1$ and every $j \in S_m$,

$$R_j^{(m)} = \sum_{\ell=1}^{K} W_{j\ell}^{(m)} C_\ell^Y. \tag{2.10}$$

For $m = 0$, $R_j^{(0)} = C_j^Y =$, so (2.10) holds. Now suppose (2.10) holds at stage $m - 1$. By nearly identical steps to before, there is satisfaction at m and the result is proved. Defining $w_{1\ell} = \frac{W_{1\ell}^{(K-1)}}{D_{11}^{(K-1)}}$ and substituting into (2.8),

$$\beta_1 = \sum_{\ell=1}^{K} w_{1\ell} C_\ell^Y. \tag{2.11}$$

As mentioned, linear algebra provides a much more elegant way to write this expression. In the general case for β_k, $w_{k\ell} = \frac{\mathrm{cof}_{\ell k}(\Sigma)}{\det(\Sigma)}$.

3 New Weighted Average Representations

Consider the case where all elements of $\mathbf{X}$ are functions of some fixed random variable X. From prior exposition, it would be natural to expect that β_j represents a weighted average of an object depending on the derivitive of the j-th function. But it follows immediately from earlier derivations this is not the case; conditioning on a function of X is equivalent to condition on X. So in fact, all projection estimands are a weighted average of the same object: the derivitive of the conditional mean function of Y with respect to X. This contrasts with the standard unit change interpretation of regresssion. In this case, the various functions in $\mathbf{X}$ simply change the weighting across estimands, as shown in the previous illustration. This builds intuition for the main result in this illustration, which is that β_j can be written as a weighted average of the derivatives of the conditional mean function of Y with respect to the entire random vector $\mathbf{X}$, not just X_j. The "equivalent conditioning" idea, combined with the previous illustration, shows why it would be a case that there exists a way to map the standard representation into an expanded form.

Fix j and let $\mathbf{W} = \{X_k : k \neq j\}$. Define the conditional mean function as $g(x, \mathbf{w}) = \mathrm{E}[Y \mid X_j = x, \mathbf{W} = \mathbf{w}]$. Recall $\omega_j(x) \propto \mathrm{E}[\mathbf{1}_{\{} X_j \geq x\} X_j^{\perp} \mid \mathbf{W}]$ and define

$$\widetilde{\omega_j}(x, \mathbf{w}) = \mathrm{E}[\mathbf{1}_{\mathbf{W} \geq \mathbf{w}} \omega_j(x)].$$

For a fixed $a \in \mathbb{R}$ define the recentered weighting kernel

$$\omega_{j,a}(x) = \omega_j(x) - \mathbf{1}\{a > x\} \mathrm{E}[X_j^{\perp} \mid \mathbf{W}].$$

Lemma 1: If $\mathrm{E}[X_j^{\perp} g(a, \mathbf{W})] = 0$, then $\beta_j = \int \omega_{j,a}(x) \cdot m_j'(x) \; \mathrm{d}x$.

We present the main result in the $K = 2$ case for simplicity. Recall $m_j'(x_j) = \partial_j \, g(x_j, \mathbf{W})$. For the $K = 2$ case, let $m_{1,2}'(x_1, x_2) = \partial_{1,2} \, g(x_1, x_2)$.

Proposition 1: Suppose $\exists$ $\mathbf{a} = (a_1, a_2) \in \mathbb{R}^2$ s.t $g(\mathbf{a}) = 0$. Consider a linear projection of Y on X_1, X_2. The estimand on β_1 satisfies

$$\begin{aligned} \beta_1 = & \int \widetilde{\omega}_1(x_1, -\infty) \, m_1'(x_1, a_2) \; \mathrm{d}x_1 + \int \widetilde{\omega}_1(-\infty, x_2) \, m_2'(a_1, x_2) \; \mathrm{d}x_2 \\ & + \iint \Omega_{12}(x_1, x_2; \mathbf{a}) \, m_{1,2}'(x_1, x_2) \; \mathrm{d}x_2 \; \; \mathrm{d}x_1, \end{aligned} \tag{3.1}$$

where $\Omega_{12}(x_1, x_2; \mathbf{a}) = \widetilde{\omega}_1(x_1, x_2) - \mathbf{1}_{a_2 > x_2} \widetilde{\omega}_1(x_1, -\infty) - \mathbf{1}_{a_1 > x_1} \widetilde{\omega}_1(-\infty, x_2)$.

The Appendix contains the proofs and also offers discussion on how restrictive the sufficient condition is. For higher dimensional examples, this representation requires further assumptions on the smoothness of the conditional mean function. But these representations are informative to show another sense in which linear projections "partial out" the effects of other regressors, in addition to putting forth different weighted average representations.

4 Testing for Marginal Effect Dependence

Assume X_j is independent of the other elements of $\mathbf{X}$. The complex interdependencies discussed in the previous two illustrations immediately vanish. However, it is still possible the marginal effects of X_j on Y depend on the remaining elements of $\mathbf{X}$. The following result provides a simple test of this possibility. For simplicity, we look at the $K = 2$ case, but the result generalizes.

Proposition 2: Suppose (Y, X_1, X_2) is a mean zero random vector and $X_1 \perp X_2$. If $m_1'(\cdot)$ is independent of X_2, the estimand on the interaction term in a projection of Y on $X_1, X_1 X_2$ is zero.

Proof: The projection coefficient is 0 if $\mathrm{E}[X_1 X_2 Y] = 0$. By the L.I.E, $\mathrm{E}[X_2 X_1 Y] = \mathrm{E}[X_2 \mathrm{E}[X_1 Y \mid X_2]]$. Under the assumed sufficient condition, we can invoke Fubini's Theorem and the Fundamental Theorem of Calculus and write $\mathrm{E}[X_1 Y \mid X_2] = \mathrm{E}[\int \omega(a) m_1'(x) \, \mathrm{d}x]$ for some weighting function $\omega_1(\cdot)$. Therefore, $\mathrm{E}[X_1 Y \mid X_2]$ is equal to some constant $\kappa_h \in \mathbb{R}$ and $\mathrm{E}[X_2 X_1 Y] = 0$.

Though this result may appear trivial ex post, it shows the mean zero normalization is a necessary step to test for joint dependence of marginal effects.

5 Conclusion

These three illustrations are motivated by an attempt to understand what linear projections multivariate projections uncover, in particular how the choice of included regressors affects the estimand. The primary motivation is the weighted average representation of linear projections and seeking to better clarify how the weighting schemes integrate over the entire joint distribution. Though the first two illustrations are primarily expository in nature, they are informative for thinking about different applications of linear projections, particularly the case of targeting a particular weighting function. The lasst illustration provides an analog to test for dependence in the continuous case, which is useful because the saturated specification will often be infeasible.

Appendix

Assumptions

First, $Y, X_1, \ldots, X_K$ are square integrable, mean zero random variables and the covariance matrix Σ_{XX} is nonsingular. For each fixed j, let $\mathbf{W}$ denote the remaining elements of $\{X_k\}_{k=1}^K$, and let

$$g(x, \mathbf{w}) = \mathrm{E}[Y \mid X_j = x, \mathbf{W} = \mathbf{w}]$$

be a jointly measurable version of the conditional mean function. Assume that for almost every $\mathbf{w}$, the map $x \mapsto g(x, \mathbf{w})$ is absolutely continuous. For every $a \in \mathbb{R}$ used below, assume

$$\mathrm{E}\left[\left|X_j^{\perp} g(a, \mathbf{W})\right| + \int_{\mathbb{R}} \left|(\mathbf{1}\{X_j \geq x\} - \mathbf{1}\{a > x\})\, X_j^{\perp} \frac{\partial g(x, \mathbf{W})}{\partial x}\right| \mathrm{d}x\right] < \infty.$$

In the bivariate case, write

$$g(x_1, x_2) = \mathrm{E}[Y \mid X_1 = x_1, X_2 = x_2].$$

Assume g is twice continuously differentiable and that for every $\mathbf{a} = (a_1, a_2) \in \mathbb{R}^2$,

$$\mathrm{E}\left[|X_1^{\perp}|\left(\int_{\mathbb{R}} \left|P_1(x_1)\frac{\partial g(x_1, a_2)}{\partial x_1}\right| \mathrm{d}x_1 + \int_{\mathbb{R}} \left|P_2(x_2)\frac{\partial g(a_1, x_2)}{\partial x_2}\right| \mathrm{d}x_2 \right.\right.$$
$$\left.\left. + \int_{\mathbb{R}}\int_{\mathbb{R}} \left|P_1(x_1)P_2(x_2)\frac{\partial^2 g(x_1, x_2)}{\partial x_1 \partial x_2}\right| \mathrm{d}x_2 \mathrm{d}x_1\right)\right] < \infty,$$

where $P_i(x_i) = \mathbf{1}\{X_i \geq x_i\} - \mathbf{1}\{a_i > x_i\}$ for $i = 1, 2$. Finally, whenever Section 4 invokes the one-dimensional representation inside a conditional expectation, assume the same absolute continuity and integrability conditions hold conditionally on the relevant conditioning variables. These assumptions are stronger than necessary, and they cover every use of iterated expectations, Fubini or Tonelli, and the Fundamental Theorem of Calculus.

Proofs

Lemma 2. Fix $a, t \in \mathbb{R}$. If $f : \mathbb{R} \to \mathbb{R}$ is absolutely continuous, then

$$f(t) - f(a) = \int \big(\mathbf{1}\{t \geq x\} - \mathbf{1}\{a > x\}\big) f'(x)\, dx.$$

Proof. If $t \geq a$, then the integrand equals $f'(x)$ on $[a, t]$ and 0 elsewhere. If $t < a$, then it equals $-f'(x)$ on $[t, a)$ and 0 elsewhere. In either case the integral is $f(t) - f(a)$.

Proof of Lemma 1. Assume that, for almost every $\mathbf{w}$, the map $x \mapsto g(x, \mathbf{w})$ is absolutely continuous and

$$\mathrm{E}\left[\left|X_j^{\perp} g(a, \mathbf{W})\right| + \int \left|\big(\mathbf{1}\{X_j \geq x\} - \mathbf{1}\{a > x\}\big) X_j^{\perp} \frac{\partial g(x, \mathbf{W})}{\partial x}\right| dx\right] < \infty.$$

By Lemma 2,

$$g(X_j, \mathbf{W}) - g(a, \mathbf{W}) = \int \big(\mathbf{1}\{X_j \geq x\} - \mathbf{1}\{a > x\}\big) \frac{\partial g(x, \mathbf{W})}{\partial x}\, dx.$$

Multiply by $X_j^{\perp}$ and take expectations. Fubini's theorem yields

$$\begin{aligned}
\mathrm{E}[X_j^{\perp} g(X_j, \mathbf{W})] &= \mathrm{E}[X_j^{\perp} g(a, \mathbf{W})] \\
&\quad + \mathrm{E}\left[\int \frac{\mathrm{E}\Big[\big(\mathbf{1}\{X_j \geq x\} - \mathbf{1}\{a > x\}\big) X_j^{\perp} \mid \mathbf{W}\Big]}{\mathrm{Var}(X_j^{\perp})} \frac{\partial g(x, \mathbf{W})}{\partial x}\, dx\right] \mathrm{Var}(X_j^{\perp}) \\
&= \mathrm{E}[X_j^{\perp} g(a, \mathbf{W})] + \mathrm{Var}(X_j^{\perp}) \mathrm{E}\left[\int \omega_{j,a}(x, \mathbf{W}) \frac{\partial g(x, \mathbf{W})}{\partial x}\, dx\right].
\end{aligned}$$

Now divide by $\mathrm{Var}(X_j^{\perp})$ and invoke the definition of $\widetilde{\omega}$. If $\mathrm{E}[X_j^{\perp} g(a, \mathbf{W})] = 0$, the result follows immediately.

Lemma 3. Let $g : \mathbb{R}^2 \to \mathbb{R}$ be twice continuously differentiable. For any (x_1, x_2) and (a_1, a_2),

$$\begin{aligned}
g(x_1, x_2) - g(a_1, a_2) &= \int \big(\mathbf{1}\{x_1 \geq z_1\} - \mathbf{1}\{a_1 > z_1\}\big) m_1(z_1, a_2)\, dz_1 \\
&\quad + \int \big(\mathbf{1}\{x_2 \geq z_2\} - \mathbf{1}\{a_2 > z_2\}\big) m_2(a_1, z_2)\, dz_2 \\
&\quad + \iint \big(\mathbf{1}\{x_1 \geq z_1\} - \mathbf{1}\{a_1 > z_1\}\big) \\
&\qquad\qquad \times \big(\mathbf{1}\{x_2 \geq z_2\} - \mathbf{1}\{a_2 > z_2\}\big) m_{1,2}(z_1, z_2)\, dz_2\, dz_1.
\end{aligned}$$

Proof. Apply Lemma 2 to the map $z_1 \mapsto g(z_1, a_2)$ to obtain

$$g(x_1, a_2) - g(a_1, a_2) = \int \big(\mathbf{1}\{x_1 \geq z_1\} - \mathbf{1}\{a_1 > z_1\}\big) m_1(z_1, a_2)\, dz_1.$$

Apply Lemma 2 again to the map $z_2 \mapsto g(a_1, z_2)$ to obtain

$$g(a_1, x_2) - g(a_1, a_2) = \int \big(\mathbf{1}\{x_2 \geq z_2\} - \mathbf{1}\{a_2 > z_2\}\big) m_2(a_1, z_2)\, dz_2.$$

For the mixed term, first apply Lemma 2 to the map $z_2 \mapsto m_1(z_1, z_2)$ for fixed z_1, giving

$$\int \big(\mathbf{1}\{x_2 \geq z_2\} - \mathbf{1}\{a_2 > z_2\}\big) m_{1,2}(z_1, z_2)\, dz_2 = m_1(z_1, x_2) - m_1(z_1, a_2).$$

Apply Lemma 2 once more in z_1 to obtain

$$\begin{aligned}
&\iint (\mathbf{1}\{x_1 \geq z_1\} - \mathbf{1}\{a_1 > z_1\})(\mathbf{1}\{x_2 \geq z_2\} - \mathbf{1}\{a_2 > z_2\}) m_{1,2}(z_1, z_2)\, dz_2\, dz_1 \\
&\quad = \int (\mathbf{1}\{x_1 \geq z_1\} - \mathbf{1}\{a_1 > z_1\})(m_1(z_1, x_2) - m_1(z_1, a_2))\, dz_1 \\
&\quad = g(x_1, x_2) - g(a_1, x_2) - g(x_1, a_2) + g(a_1, a_2).
\end{aligned}$$

Combining the three displayed identities gives the claim.

Proof of Proposition 1. Let

$$P_1(x_1) = \mathbf{1}\{X_1 \geq x_1\} - \mathbf{1}\{a_1 > x_1\}, \qquad P_2(x_2) = \mathbf{1}\{X_2 \geq x_2\} - \mathbf{1}\{a_2 > x_2\}.$$

By Lemma 3,

$$\begin{aligned}
g(X_1, X_2) - g(a_1, a_2) = \int P_1(x_1) m_1(x_1, a_2)\, dx_1 + \int P_2(x_2) m_2(a_1, x_2)\, dx_2 \\
+ \iint P_1(x_1) P_2(x_2) m_{1,2}(x_1, x_2)\, dx_2\, dx_1.
\end{aligned}$$

Under the assumption $g(\mathbf{a}) = 0$, this becomes an expansion for $g(X_1, X_2)$. Multiply by $X_1^{\perp}$ and take expectations. Since $\mathrm{E}[X_1^{\perp}] = 0$, Fubini's theorem gives

$$\begin{aligned}
\mathrm{E}[X_1^{\perp} g(X_1, X_2)] = \int \mathrm{E}[X_1^{\perp} P_1(x_1)] m_1(x_1, a_2)\, dx_1 + \int \mathrm{E}[X_1^{\perp} P_2(x_2)] m_2(a_1, x_2)\, dx_2 \\
+ \iint \mathrm{E}[X_1^{\perp} P_1(x_1) P_2(x_2)] m_{1,2}(x_1, x_2)\, dx_2\, dx_1.
\end{aligned}$$

Divide by $\mathrm{Var}(X_1^{\perp})$ and invoke the definition of $\widetilde{\omega}$. The first kernel is

$$\frac{\mathrm{E}[X_1^{\perp} P_1(x_1)]}{\mathrm{Var}(X_1^{\perp})} = \frac{\mathrm{E}[X_1^{\perp} \mathbf{1}\{X_1 \geq x_1\}]}{\mathrm{Var}(X_1^{\perp})} = \widetilde{\omega}_1(x_1, -\infty),$$

because $\mathrm{E}[X_1^{\perp}] = 0$. Similarly,

$$\frac{\mathrm{E}[X_1^{\perp} P_2(x_2)]}{\mathrm{Var}(X_1^{\perp})} = \frac{\mathrm{E}[X_1^{\perp} \mathbf{1}\{X_2 \geq x_2\}]}{\mathrm{Var}(X_1^{\perp})} = \widetilde{\omega}_1(-\infty, x_2).$$

For the mixed term,

$$\frac{\mathrm{E}[X_1^{\perp} P_1(x_1) P_2(x_2)]}{\mathrm{Var}(X_1^{\perp})} = \widetilde{\omega}_1(x_1, x_2) - \mathbf{1}_{\{a_2 > x_2\}} \widetilde{\omega}_1(x_1, -\infty) - \mathbf{1}_{\{a_1 > x_1\}} \widetilde{\omega}_1(-\infty, x_2),$$

where the term involving $\mathbf{1}_{\{a_1 > x_1\}} \mathbf{1}_{\{a_2 > x_2\}} \mathrm{E}[X_1^{\perp}]$ vanishes. Substituting these kernels into the preceding display yields (3.1).

Discussion on Sufficient Condition for Proposition 1

The second illustration posits that there exists some constant term that allows a particular term to be zeroed out. A natural question is what classes of data generating processes would this not exist. After transforming $(X, Y, \mathbf{W})$ to be mean 0 and stationary, an informal but still exhaustive answer is

- If $g(x, \cdot)$ has a large jump every time it changes sign.

- If **both** (i) and (ii) are true: (i) the conditional means of both X and Y are highly nonlinear in $\mathbf{W}$ in the same way (i.e., the linear projection residuals are highly correlated) and (ii) $|g(\cdot, \mathbf{W})| >> 0$ (either $g|(x, \cdot)|$ is almost everywhere increasing with $|\mathbf{E}[Y \mid \mathbf{W}]| >> 0$ or $g(x, \cdot)$ does not vary much with x (i.e., both $g(x, \cdot)$ and $g'(x, \cdot)$ are not divergent in the positive and negative infinite limits)).

If the second tenet is still a problem after the variable transformations, this setting is unlikely to be of interest to begin with, so this should not be restrictive in practice. The first tenet notes a non-negligible restriction: a particularly "adversarial" threshold/state-dependent form for $g(\cdot)$. But this is a still a knife-edge case we are ruling out; a worst-case bias is still 0 if the function were highly non-smooth but still continuous (e.g., sigmoidal behavior around the points of sign change).

It's therefore reasonable to argue that representating β with the recentered ω is generic.