# Regression - Tutorial 3

Projection Matrices, The Linear Model, and Joint Distributions

## 1 Orthogonal Projection Matrices

We begin by revisiting some fundamental concepts from linear algebra, crucial for understanding the geometry of least squares estimation.

**Definition 1.1** (Idempotent Matrix)**.** A square matrix $\mathbf{A}$ is called **idempotent** if $\mathbf{A}^2 = \mathbf{A}$.

Think of an idempotent matrix as an operation that, once performed, has no further effect if applied again. Projecting a vector onto a subspace is a prime example: projecting the already projected vector doesn't change it.

**Definition 1.2** (Orthogonal Projection Matrix)**.** A square matrix $\mathbf{P}$ is called an **orthogonal projection matrix** if it is both:

1. **Symmetric**: $\mathbf{P}^\top = \mathbf{P}$.
2. **Idempotent**: $\mathbf{P}^2 = \mathbf{P}$.

Why these two properties? Symmetry ensures that the projection is orthogonal (we'll see this more clearly later), and idempotency captures the "projecting once is enough" idea.

**Claim 1.3.** *The eigenvalues of an orthogonal projection matrix $\mathbf{P}$ are either 0 or 1. The multiplicity of the eigenvalue 1 is equal to the rank of $\mathbf{P}$, $\operatorname{rank}(\mathbf{P})$, and the multiplicity of the eigenvalue 0 is equal to the dimension of the kernel (null space) of $\mathbf{P}$, $\dim(\operatorname{Ker}(\mathbf{P}))$.*

*Proof Sketch.* Let $\lambda$ be an eigenvalue of $\mathbf{P}$ with corresponding eigenvector $\boldsymbol{v} \neq \mathbf{0}$, so $\mathbf{P}\boldsymbol{v} = \lambda\boldsymbol{v}$. Applying $\mathbf{P}$ again, and using idempotency:

$$\mathbf{P}^2\boldsymbol{v} = \mathbf{P}(\mathbf{P}\boldsymbol{v}) = \mathbf{P}(\lambda\boldsymbol{v}) = \lambda(\mathbf{P}\boldsymbol{v}) = \lambda(\lambda\boldsymbol{v}) = \lambda^2\boldsymbol{v}$$

But $\mathbf{P}^2 = \mathbf{P}$, so $\mathbf{P}^2\boldsymbol{v} = \mathbf{P}\boldsymbol{v} = \lambda\boldsymbol{v}$. Therefore, we must have $\lambda\boldsymbol{v} = \lambda^2\boldsymbol{v}$. Since $\boldsymbol{v} \neq \mathbf{0}$, this implies $\lambda = \lambda^2$, which means $\lambda(\lambda - 1) = 0$. Thus, the only possible eigenvalues are $\lambda = 0$ or $\lambda = 1$.

Since $\mathbf{P}$ is symmetric, it is diagonalizable. Its trace equals the sum of its eigenvalues, and its rank equals the number of non-zero eigenvalues. Let $r = \operatorname{rank}(\mathbf{P})$. Since the only non-zero eigenvalue is 1, there must be $r$ eigenvalues equal to 1. The remaining $n - r$ eigenvalues must be 0, where $n$ is the dimension of the matrix. The dimension of the eigenspace corresponding to $\lambda = 0$ is $\dim(\operatorname{Ker}(\mathbf{P}))$, which is $n - r$. □

### 1.1 Projection onto the Column Space of X

A particularly important projection matrix in regression is the one that projects vectors onto the subspace spanned by the columns of the design matrix $\mathbf{X}$.

**Definition 1.4** (Projection Matrix onto $\operatorname{Im}(\mathbf{X})$)**.** Let $\mathbf{X}$ be an $n \times m$ matrix with **full column rank** (i.e., its columns are linearly independent, implying $m \leq n$). The **projection matrix onto the column space (image) of X** is defined as:

$$\mathbf{P}_{\mathbf{X}} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$$

*Remark* 1.5. The condition that $\mathbf{X}$ has full column rank ensures that the $m \times m$ matrix $\mathbf{X}^\top \mathbf{X}$ (often called the Gram matrix) is invertible. Why? If $\mathbf{X}^\top \mathbf{X}\boldsymbol{u} = \mathbf{0}$ for some $\boldsymbol{u} \in \mathbb{R}^m$, then $\boldsymbol{u}^\top \mathbf{X}^\top \mathbf{X}\boldsymbol{u} = 0$, which means $(\mathbf{X}\boldsymbol{u})^\top(\mathbf{X}\boldsymbol{u}) = \|\mathbf{X}\boldsymbol{u}\|^2 = 0$. This implies $\mathbf{X}\boldsymbol{u} = \mathbf{0}$. Since the columns of $\mathbf{X}$ are linearly independent, the only solution is $\boldsymbol{u} = \mathbf{0}$. Thus, $\mathbf{X}^\top \mathbf{X}$ is invertible.

Let's verify that $\mathbf{P_X}$ is indeed an orthogonal projection matrix and projects onto the intended space.

**Claim 1.6.** $\mathbf{P_X} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$ *is an orthogonal projection matrix onto* $\mathrm{Im}(\mathbf{X})$.

*Proof.* We need to show three things:

1. **Symmetry:** We check if $\mathbf{P}_\mathbf{X}^\top = \mathbf{P_X}$.

$$\begin{aligned}
\mathbf{P}_\mathbf{X}^\top &= \left(\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\right)^\top \\
&= (\mathbf{X}^\top)^\top \left((\mathbf{X}^\top\mathbf{X})^{-1}\right)^\top \mathbf{X}^\top \\
&= \mathbf{X}\left((\mathbf{X}^\top\mathbf{X})^\top\right)^{-1}\mathbf{X}^\top \quad \text{(since } (A^{-1})^\top = (A^\top)^{-1}\text{)} \\
&= \mathbf{X}\left(\mathbf{X}^\top(\mathbf{X}^\top)^\top\right)^{-1}\mathbf{X}^\top \\
&= \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top \\
&= \mathbf{P_X}
\end{aligned}$$

   Thus, $\mathbf{P_X}$ is symmetric.

2. **Idempotency:** We check if $\mathbf{P}_\mathbf{X}^2 = \mathbf{P_X}$.

$$\begin{aligned}
\mathbf{P}_\mathbf{X}^2 &= \left(\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\right)\left(\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\right) \\
&= \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}(\mathbf{X}^\top\mathbf{X})(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top \\
&= \mathbf{X}\mathbf{I}_m(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top \\
&= \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top \\
&= \mathbf{P_X}
\end{aligned}$$

   Thus, $\mathbf{P_X}$ is idempotent.

3. **Projection onto** $\mathrm{Im}(\mathbf{X})$**:** We need to show that for any vector $\boldsymbol{v} \in \mathbb{R}^n$, the result $\mathbf{P_X}\boldsymbol{v}$ lies in the column space of $\mathbf{X}$, i.e., $\mathbf{P_X}\boldsymbol{v} \in \mathrm{Im}(\mathbf{X})$. Let $\boldsymbol{v} \in \mathbb{R}^n$. Consider $\mathbf{P_X}\boldsymbol{v} = \mathbf{X}[(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\boldsymbol{v}]$. Let $\boldsymbol{u} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\boldsymbol{v}$. Note that $\boldsymbol{u}$ is an $m \times 1$ vector (a vector of coefficients). Then $\mathbf{P_X}\boldsymbol{v} = \mathbf{X}\boldsymbol{u}$. This is, by definition, a linear combination of the columns of $\mathbf{X}$ with coefficients given by the entries of $\boldsymbol{u}$. Therefore, $\mathbf{P_X}\boldsymbol{v}$ must be in the column space (image) of $\mathbf{X}$.

Since $\mathbf{P_X}$ is symmetric and idempotent, it is an orthogonal projection matrix. Since its output always lies in $\mathrm{Im}(\mathbf{X})$, it projects onto a subspace of $\mathrm{Im}(\mathbf{X})$. We will see in Property 8 below that it actually projects onto the entirety of $\mathrm{Im}(\mathbf{X})$. □

## 1.2 Key Properties of Projection Matrices

The following properties are fundamental for understanding least squares and related concepts. They are likely among the most important results in this course.

**Proposition 1.7** (Properties of $\mathbf{P_X}$)**.** *Let* $\mathbf{X}$ *be an* $n \times m$ *matrix with full column rank* $(m \leq n)$*. Then the projection matrix* $\mathbf{P_X} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$ *has the following properties:*

1. *$\mathbf{P}_{\mathbf{X}}$ is symmetric. (Proven above)*

2. *$\mathbf{P}_{\mathbf{X}}$ is idempotent, $\mathbf{P}_{\mathbf{X}}^2 = \mathbf{P}_{\mathbf{X}}$. (Proven above)*

3. *$\mathbf{P}_{\mathbf{X}}\mathbf{X} = \mathbf{X}$.*

   *Proof.* $\mathbf{P}_{\mathbf{X}}\mathbf{X} = \left(\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\right)\mathbf{X} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}(\mathbf{X}^\top\mathbf{X}) = \mathbf{X}\mathbf{I}_m = \mathbf{X}$. (Interpretation: Projecting the columns of $\mathbf{X}$ onto their own span leaves them unchanged.) □

4. *$\mathbf{X}^\top(\mathbf{I} - \mathbf{P}_{\mathbf{X}}) = \mathbf{0} \in \mathbb{R}^{m\times n}$.*

   *Proof.* $\mathbf{X}^\top(\mathbf{I}-\mathbf{P}_{\mathbf{X}}) = \mathbf{X}^\top\mathbf{I}-\mathbf{X}^\top\mathbf{P}_{\mathbf{X}} = \mathbf{X}^\top-\mathbf{X}^\top(\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top) = \mathbf{X}^\top-(\mathbf{X}^\top\mathbf{X})(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top = \mathbf{X}^\top - \mathbf{I}_m\mathbf{X}^\top = \mathbf{X}^\top - \mathbf{X}^\top = \mathbf{0}$. (Interpretation: The columns of $\mathbf{X}$ are orthogonal to the "residual projection" $\mathbf{I} - \mathbf{P}_{\mathbf{X}}$. We'll see this matrix projects onto the orthogonal complement space.) □

5. *$\mathbf{P}_{\mathbf{X}}\boldsymbol{v} \in \mathrm{Im}(\mathbf{X})$ for all $\boldsymbol{v} \in \mathbb{R}^n$. (Proven above)*

6. *If $m = n$ and $\mathbf{X}$ is invertible, then $\mathbf{P}_{\mathbf{X}} = \mathbf{I}$.*

   *Proof.* If $\mathbf{X}$ is $n \times n$ and invertible, then $\mathbf{X}^\top$ is also invertible. $\mathbf{P}_{\mathbf{X}} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top = \mathbf{X}\mathbf{X}^{-1}(\mathbf{X}^\top)^{-1}\mathbf{X}^\top = \mathbf{I}\mathbf{I} = \mathbf{I}$. (Interpretation: If the columns of $\mathbf{X}$ span the entire space $\mathbb{R}^n$, projecting onto that space leaves every vector unchanged.) □

7. *$(\mathbf{I} - \mathbf{P}_{\mathbf{X}})\boldsymbol{v} \in \mathrm{Im}(\mathbf{X})^\perp$ for all $\boldsymbol{v} \in \mathbb{R}^n$. (See Proposition 1.10 below.)*

8. *If $\boldsymbol{w} \in \mathrm{Im}(\mathbf{X})$, then $\mathbf{P}_{\mathbf{X}}\boldsymbol{w} = \boldsymbol{w}$.*

   *Proof.* If $\boldsymbol{w} \in \mathrm{Im}(\mathbf{X})$, then $\boldsymbol{w} = \mathbf{X}\boldsymbol{a}$ for some $\boldsymbol{a} \in \mathbb{R}^m$. Then $\mathbf{P}_{\mathbf{X}}\boldsymbol{w} = \mathbf{P}_{\mathbf{X}}(\mathbf{X}\boldsymbol{a}) = (\mathbf{P}_{\mathbf{X}}\mathbf{X})\boldsymbol{a}$. By Property 3, $\mathbf{P}_{\mathbf{X}}\mathbf{X} = \mathbf{X}$. So, $\mathbf{P}_{\mathbf{X}}\boldsymbol{w} = \mathbf{X}\boldsymbol{a} = \boldsymbol{w}$. (Interpretation: Vectors already in the subspace are unaffected by the projection.) □

9. *If $\boldsymbol{w} \in \mathrm{Im}(\mathbf{X})^\perp$, then $\mathbf{P}_{\mathbf{X}}\boldsymbol{w} = \mathbf{0}$.*

   *Proof.* If $\boldsymbol{w} \in \mathrm{Im}(\mathbf{X})^\perp$, then $\boldsymbol{w}$ is orthogonal to every column of $\mathbf{X}$. This means $\mathbf{X}^\top\boldsymbol{w} = \mathbf{0}$. Then $\mathbf{P}_{\mathbf{X}}\boldsymbol{w} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}(\mathbf{X}^\top\boldsymbol{w}) = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{0} = \mathbf{0}$. (Interpretation: Vectors orthogonal to the subspace are projected to the zero vector.) □

10. *If $\mathbf{Z}$ is another $n \times k$ matrix such that $\mathrm{Im}(\mathbf{Z}) = \mathrm{Im}(\mathbf{X})$, then $\mathbf{P}_{\mathbf{Z}} = \mathbf{P}_{\mathbf{X}}$. This means that $\mathbf{P}_{\mathbf{X}}$ depends on $\mathbf{X}$ only through the subspace spanned by its columns. Hence, for an arbitrary linear subspace $M \subseteq \mathbb{R}^n$, we can define the projection matrix $\mathbf{P}_M$ onto $M$. An explicit form for $\mathbf{P}_M$ can be obtained by taking any basis of $M$, stacking its elements as columns in a matrix $\mathbf{X}$, then forming $\mathbf{P}_M := \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$.*

11. *If $L$ and $M$ are two subspaces with $L \subseteq M$, then $\mathbf{P}_M\mathbf{P}_L = \mathbf{P}_L\mathbf{P}_M = \mathbf{P}_L$.*

    *Proof.* If $\boldsymbol{v} \in \mathbb{R}^n$, then $\mathbf{P}_L\boldsymbol{v} \in L$. Since $L \subseteq M$, we have $\mathbf{P}_L\boldsymbol{v} \in M$. By Property 8 applied to $\mathbf{P}_M$, if $\boldsymbol{w} = \mathbf{P}_L\boldsymbol{v} \in M$, then $\mathbf{P}_M\boldsymbol{w} = \boldsymbol{w}$. Thus, $\mathbf{P}_M(\mathbf{P}_L\boldsymbol{v}) = \mathbf{P}_L\boldsymbol{v}$. Since this holds for all $\boldsymbol{v}$, $\mathbf{P}_M\mathbf{P}_L = \mathbf{P}_L$. For the other equality, $\mathbf{P}_L\mathbf{P}_M$: take transpose $(\mathbf{P}_M\mathbf{P}_L)^\top = \mathbf{P}_L^\top\mathbf{P}_M^\top = \mathbf{P}_L\mathbf{P}_M$. Since $\mathbf{P}_L$ is symmetric, $(\mathbf{P}_L)^\top = \mathbf{P}_L$. Thus $\mathbf{P}_L\mathbf{P}_M = \mathbf{P}_L$. (Interpretation: Projecting onto a smaller subspace $L$ and then onto a larger subspace $M$ containing $L$ is the same as just projecting onto $L$. Projecting onto $M$ first and then $L$ is also just projecting onto $L$.) □

## 1.3 The Orthogonal Complement Space

**Definition 1.8** (Orthogonal Complement)**.** Let $V$ be an inner product space (e.g., $\mathbb{R}^n$ with the standard dot product) and let $U \subseteq V$ be a subspace. The **orthogonal complement** of $U$, denoted $U^\perp$, is defined as:

$$U^\perp = \{\boldsymbol{v} \in V \mid \boldsymbol{u}^\top \boldsymbol{v} = 0 \text{ for all } \boldsymbol{u} \in U\}$$

$U^\perp$ is the set of all vectors in $V$ that are orthogonal to every vector in $U$. It is also a subspace of $V$.

**Theorem 1.9** (Direct Sum Decomposition)**.** *Let $U$ be a subspace of a finite-dimensional inner product space $V$. Then $V$ can be written as the direct sum of $U$ and its orthogonal complement $U^\perp$:*

$$V = U \oplus U^\perp$$

*This means that every vector $\boldsymbol{v} \in V$ can be uniquely written as $\boldsymbol{v} = \boldsymbol{u} + \boldsymbol{w}$, where $\boldsymbol{u} \in U$ and $\boldsymbol{w} \in U^\perp$. Furthermore, $\boldsymbol{u}$ is the orthogonal projection of $\boldsymbol{v}$ onto $U$, and $\boldsymbol{w}$ is the orthogonal projection of $\boldsymbol{v}$ onto $U^\perp$.*

*Proof.* Omitted (standard result from linear algebra, often proven using Gram-Schmidt orthogonalization). □

This theorem is geometrically intuitive: any vector can be uniquely decomposed into a component lying within a subspace and a component orthogonal to that subspace. The matrix $\mathbf{P_X}$ finds the component in $U = \mathrm{Im}(\mathbf{X})$. What finds the component in $U^\perp$?

**Proposition 1.10** (Projection onto the Orthogonal Complement)**.** *Let $\mathbf{P_X}$ be the orthogonal projection matrix onto $M = \mathrm{Im}(\mathbf{X})$. Then:*

1. *$\mathbf{I} - \mathbf{P_X}$ is the orthogonal projection matrix onto the orthogonal complement $M^\perp = \mathrm{Im}(\mathbf{X})^\perp$. We denote this $\mathbf{P}_{M^\perp} = \mathbf{I} - \mathbf{P_X}$.*
2. *If $L$ and $M$ are two subspaces of $\mathbb{R}^n$ with $L \subseteq M$, then $\mathbf{P}_M - \mathbf{P}_L$ is the orthogonal projection matrix onto the subspace $M \cap L^\perp$ (the part of $M$ that is orthogonal to $L$). We denote this $\mathbf{P}_{M\cap L^\perp} = \mathbf{P}_M - \mathbf{P}_L$.*

*Proof Sketch for Part 1.* Let $\mathbf{Q} = \mathbf{I} - \mathbf{P_X}$.

- Symmetry: $\mathbf{Q}^\top = (\mathbf{I} - \mathbf{P_X})^\top = \mathbf{I}^\top - \mathbf{P_X}^\top = \mathbf{I} - \mathbf{P_X} = \mathbf{Q}$. (Since $\mathbf{P_X}$ is symmetric).
- Idempotency: $\mathbf{Q}^2 = (\mathbf{I} - \mathbf{P_X})(\mathbf{I} - \mathbf{P_X}) = \mathbf{I} - \mathbf{P_X} - \mathbf{P_X} + \mathbf{P_X}^2 = \mathbf{I} - 2\mathbf{P_X} + \mathbf{P_X} = \mathbf{I} - \mathbf{P_X} = \mathbf{Q}$. (Since $\mathbf{P_X}$ is idempotent).

So $\mathbf{Q}$ is an orthogonal projection matrix. Onto which space does it project? Let $\boldsymbol{v} \in \mathbb{R}^n$. We know $\boldsymbol{v} = \mathbf{P_X}\boldsymbol{v} + (\mathbf{I} - \mathbf{P_X})\boldsymbol{v} = \boldsymbol{u} + \boldsymbol{w}$. Here $\boldsymbol{u} = \mathbf{P_X}\boldsymbol{v} \in \mathrm{Im}(\mathbf{X})$ (by Prop 1.7.5). We need to show that $\boldsymbol{w} = (\mathbf{I} - \mathbf{P_X})\boldsymbol{v}$ lies in $\mathrm{Im}(\mathbf{X})^\perp$. To show this, we must show $\boldsymbol{w}$ is orthogonal to any vector in $\mathrm{Im}(\mathbf{X})$. Any vector in $\mathrm{Im}(\mathbf{X})$ can be written as $\mathbf{X}\boldsymbol{a}$ for some $\boldsymbol{a}$. We check the dot product:

$$(\mathbf{X}\boldsymbol{a})^\top \boldsymbol{w} = (\mathbf{X}\boldsymbol{a})^\top (\mathbf{I} - \mathbf{P_X})\boldsymbol{v} = \boldsymbol{a}^\top \mathbf{X}^\top (\mathbf{I} - \mathbf{P_X})\boldsymbol{v}$$

By Prop 1.7.4, $\mathbf{X}^\top(\mathbf{I} - \mathbf{P_X}) = \mathbf{0}$. So the dot product is $\boldsymbol{a}^\top \mathbf{0} \boldsymbol{v} = 0$. Thus, $\boldsymbol{w} = (\mathbf{I} - \mathbf{P_X})\boldsymbol{v}$ is orthogonal to all vectors in $\mathrm{Im}(\mathbf{X})$, meaning $\boldsymbol{w} \in \mathrm{Im}(\mathbf{X})^\perp$. Since $\mathbf{Q}$ is an orthogonal projection matrix and its image is contained in $\mathrm{Im}(\mathbf{X})^\perp$, it must be the projection onto $\mathrm{Im}(\mathbf{X})^\perp$. □

**Proposition 1.11.** *Let $\mathbf{Q}$ be an $n \times n$ matrix of rank $m \le n$ which is symmetric ($\mathbf{Q}^\top = \mathbf{Q}$) and idempotent ($\mathbf{Q}^2 = \mathbf{Q}$). Then $\mathbf{Q}$ is the orthogonal projection matrix onto its own image, i.e., $\mathbf{Q} = \mathbf{P}_M$ where $M := \mathrm{Im}(\mathbf{Q})$.*

*Proof.* Exercise. (Hint: Show that any vector $\boldsymbol{v}$ can be written as $\mathbf{Q}\boldsymbol{v} + (\mathbf{I} - \mathbf{Q})\boldsymbol{v}$, show $\mathbf{Q}\boldsymbol{v} \in \mathrm{Im}(\mathbf{Q})$ and $(\mathbf{I} - \mathbf{Q})\boldsymbol{v} \in \mathrm{Im}(\mathbf{Q})^\perp$.) □

## 2 Connections to Linear Algebra and OLS

Let's explore some further connections.

**Question 2.1.** 1. Assume $\mathbf{A}$ is a square matrix. Prove that $\text{Im}(\mathbf{A}^\top) = \text{Ker}(\mathbf{A})^\perp$. (This is a fundamental theorem of linear algebra).

2. Claim (without proof): A matrix $\mathbf{A}$ is non-diagonalizable if and only if there exists at least one eigenvalue $\lambda_i$ of $\mathbf{A}$ for which the minimal polynomial has a factor $(x - \lambda_i)^k$ with $k \geq 2$. Equivalently, for some $k \geq 2$, $(A - \lambda_i I)^k = 0$ but $(A - \lambda_i I)^{k-1} \neq 0$. Use this claim to argue that if $\mathbf{A}$ is a symmetric matrix, then it must be diagonalizable.

3. Use these results to show that $\mathbf{Q} = \mathbf{I} - \mathbf{P_X}$ is the projection matrix onto the orthogonal complement of the column space of $\mathbf{X}$, i.e., $\text{Im}(\mathbf{X})^\perp$. Write the spectral decomposition of $\mathbf{Q}$ in terms of the eigenvalues and eigenvectors of $\mathbf{P_X}$.

4. Deduce from this that the Ordinary Least Squares (OLS) estimator $\hat{\boldsymbol{\beta}}_{OLS}$ minimizes the squared norm $\|\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta}\|^2$. Also argue that as the rank of $\mathbf{X}$ increases (e.g., by adding more linearly independent predictors), this minimum norm value generally decreases (or stays the same).

*Solution Sketches.* 1. **Proof of** $\text{Im}(\mathbf{A}^\top) = \text{Ker}(\mathbf{A})^\perp$: We need to show two inclusions: (a) $\text{Im}(\mathbf{A}^\top) \subseteq \text{Ker}(\mathbf{A})^\perp$: Let $\boldsymbol{y} \in \text{Im}(\mathbf{A}^\top)$. Then $\boldsymbol{y} = \mathbf{A}^\top \boldsymbol{x}$ for some $\boldsymbol{x}$. Let $\boldsymbol{z} \in \text{Ker}(\mathbf{A})$, meaning $\mathbf{A}\boldsymbol{z} = \mathbf{0}$. We need to show $\boldsymbol{y}$ is orthogonal to $\boldsymbol{z}$. Consider their dot product: $\boldsymbol{y}^\top \boldsymbol{z} = (\mathbf{A}^\top \boldsymbol{x})^\top \boldsymbol{z} = \boldsymbol{x}^\top \mathbf{A}\boldsymbol{z} = \boldsymbol{x}^\top(\mathbf{0}) = 0$. Since this holds for any $\boldsymbol{z} \in \text{Ker}(\mathbf{A})$, we have $\boldsymbol{y} \in \text{Ker}(\mathbf{A})^\perp$. (b) $\text{Ker}(\mathbf{A})^\perp \subseteq \text{Im}(\mathbf{A}^\top)$: This relies on the dimension theorem: $\dim(\text{Im}(\mathbf{A}^\top)) = \text{rank}(\mathbf{A}^\top) = \text{rank}(\mathbf{A})$. Also, $\dim(\text{Ker}(\mathbf{A})) + \text{rank}(\mathbf{A}) = n$ (where $\mathbf{A}$ is $n \times n$). Furthermore, $\dim(\text{Ker}(\mathbf{A})^\perp) = n - \dim(\text{Ker}(\mathbf{A}))$. Combining these gives $\dim(\text{Im}(\mathbf{A}^\top)) = \dim(\text{Ker}(\mathbf{A})^\perp)$. Since we already showed $\text{Im}(\mathbf{A}^\top)$ is a subspace of $\text{Ker}(\mathbf{A})^\perp$, and they have the same dimension, they must be equal.

2. **Symmetric matrices are diagonalizable**: The Spectral Theorem for symmetric matrices states they are always diagonalizable (over $\mathbb{R}$) with an orthonormal basis of eigenvectors. The provided claim gives a condition for non-diagonalizability related to the minimal polynomial. For a symmetric matrix $\mathbf{A}$, its minimal polynomial has only distinct linear factors $(x - \lambda_i)$, meaning $k = 1$ for all eigenvalues. Thus, the condition for non-diagonalizability is never met. Alternatively, one can show directly that if $(\mathbf{A} - \lambda\mathbf{I})^2 \boldsymbol{v} = \mathbf{0}$, then for symmetric $\mathbf{A}$, it must be that $(\mathbf{A} - \lambda\mathbf{I})\boldsymbol{v} = \mathbf{0}$, preventing $k \geq 2$.

3. $\mathbf{Q} = \mathbf{I} - \mathbf{P_X}$ **projects onto** $\text{Im}(\mathbf{X})^\perp$: We already proved this in Proposition 1.10. Let $M = \text{Im}(\mathbf{X})$. We showed $\mathbf{Q} = \mathbf{I} - \mathbf{P_X}$ is the projection matrix onto $M^\perp$. **Spectral Decomposition**: Since $\mathbf{P_X}$ is symmetric, it is diagonalizable. Its eigenvalues are 1 (with multiplicity $r = \text{rank}(\mathbf{X})$) and 0 (with multiplicity $n - r$). Let $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_r\}$ be an orthonormal basis for the eigenspace of $\lambda = 1$ (which is $\text{Im}(\mathbf{X})$), and $\{\boldsymbol{v}_{r+1}, \ldots, \boldsymbol{v}_n\}$ be an orthonormal basis for the eigenspace of $\lambda = 0$ (which is $\text{Ker}(\mathbf{P_X}) = \text{Im}(\mathbf{X})^\perp$). The spectral decomposition of $\mathbf{P_X}$ is $\mathbf{P_X} = \sum_{i=1}^r 1 \cdot \boldsymbol{v}_i \boldsymbol{v}_i^\top + \sum_{i=r+1}^n 0 \cdot \boldsymbol{v}_i \boldsymbol{v}_i^\top = \sum_{i=1}^r \boldsymbol{v}_i \boldsymbol{v}_i^\top$. Now consider $\mathbf{Q} = \mathbf{I} - \mathbf{P_X}$. Since $\mathbf{I} = \sum_{i=1}^n \boldsymbol{v}_i \boldsymbol{v}_i^\top$ (using the full orthonormal basis of eigenvectors):

$$\mathbf{Q} = \sum_{i=1}^n \boldsymbol{v}_i \boldsymbol{v}_i^\top - \sum_{i=1}^r \boldsymbol{v}_i \boldsymbol{v}_i^\top = \sum_{i=r+1}^n \boldsymbol{v}_i \boldsymbol{v}_i^\top$$

This is the spectral decomposition of $\mathbf{Q}$. We can also see its eigenvalues: If $\mathbf{P_X}\boldsymbol{v} = \lambda\boldsymbol{v}$, then $\mathbf{Q}\boldsymbol{v} = (\mathbf{I} - \mathbf{P_X})\boldsymbol{v} = \boldsymbol{v} - \lambda\boldsymbol{v} = (1 - \lambda)\boldsymbol{v}$. So, if $\lambda = 1$ for $\mathbf{P_X}$, the eigenvalue for $\mathbf{Q}$ is $1 - 1 = 0$. If $\lambda = 0$ for $\mathbf{P_X}$, the eigenvalue for $\mathbf{Q}$ is $1 - 0 = 1$. Thus, $\mathbf{Q}$ has eigenvalue 1 with multiplicity $n - r$ (corresponding to the basis of $\text{Im}(\mathbf{X})^\perp$) and eigenvalue 0 with multiplicity $r$ (corresponding to the basis of $\text{Im}(\mathbf{X})$).

4. **OLS Minimization and Rank**: We want to minimize $S(\boldsymbol{\beta}) = \|\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta}\|^2$. The vector $\mathbf{X}\boldsymbol{\beta}$ is always in the column space $\operatorname{Im}(\mathbf{X})$. The problem asks to find the vector $\hat{\boldsymbol{Y}} = \mathbf{X}\boldsymbol{\beta}$ within $\operatorname{Im}(\mathbf{X})$ that is closest to $\boldsymbol{Y}$. From geometry (or the projection theorem), we know this closest vector is the orthogonal projection of $\boldsymbol{Y}$ onto $\operatorname{Im}(\mathbf{X})$. That is, $\hat{\boldsymbol{Y}} = \mathbf{P}_{\mathbf{X}}\boldsymbol{Y}$. So, we need $\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{P}_{\mathbf{X}}\boldsymbol{Y}$. Pre-multiplying by $\mathbf{X}^\top$: $\mathbf{X}^\top\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}^\top\mathbf{P}_{\mathbf{X}}\boldsymbol{Y} = \mathbf{X}^\top(\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top)\boldsymbol{Y} = (\mathbf{X}^\top\mathbf{X})(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\boldsymbol{Y} = \mathbf{X}^\top\boldsymbol{Y}$. This gives the normal equations $\mathbf{X}^\top\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}^\top\boldsymbol{Y}$, whose solution is $\hat{\boldsymbol{\beta}}_{OLS} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\boldsymbol{Y}$ (assuming full rank). The minimum value of the squared norm is $\left\|\boldsymbol{Y} - \hat{\boldsymbol{Y}}\right\|^2 = \|\boldsymbol{Y} - \mathbf{P}_{\mathbf{X}}\boldsymbol{Y}\|^2 = \|(\mathbf{I} - \mathbf{P}_{\mathbf{X}})\boldsymbol{Y}\|^2$. This is the squared norm of the projection of $\boldsymbol{Y}$ onto the orthogonal complement space $\operatorname{Im}(\mathbf{X})^\perp$. Now, suppose we augment $\mathbf{X}$ to $\mathbf{X}^*$ by adding more linearly independent columns, such that $\operatorname{Im}(\mathbf{X}) \subset \operatorname{Im}(\mathbf{X}^*)$. Let $M = \operatorname{Im}(\mathbf{X})$ and $M^* = \operatorname{Im}(\mathbf{X}^*)$. Since $M \subseteq M^*$, their orthogonal complements satisfy $(M^*)^\perp \subseteq M^\perp$. Projecting $\boldsymbol{Y}$ onto these complement spaces, the norm of the projection onto the smaller space $(M^*)^\perp$ must be less than or equal to the norm of the projection onto the larger space $M^\perp$. That is, $\|(\mathbf{I} - \mathbf{P}_{\mathbf{X}^*})\boldsymbol{Y}\|^2 \leq \|(\mathbf{I} - \mathbf{P}_{\mathbf{X}})\boldsymbol{Y}\|^2$. The residual sum of squares (the minimized norm) decreases (or stays the same) as we add more predictors (increase the rank of $\mathbf{X}$).

□

## 3 Expectation and Covariance of Random Vectors

We now shift focus to probability, specifically the properties of random vectors and matrices.

**Definition 3.1** (Expectation of a Random Matrix/Vector)**.** Let $\mathbf{Z}$ be an $n \times p$ random matrix, where each element $Z_{ij}$ is a random variable. The **expectation** of $\mathbf{Z}$ is the $n \times p$ matrix of expectations:

$$\mathbb{E}[\mathbf{Z}] = \begin{pmatrix} \mathbb{E}[Z_{11}] & \cdots & \mathbb{E}[Z_{1p}] \\ \vdots & \ddots & \vdots \\ \mathbb{E}[Z_{n1}] & \cdots & \mathbb{E}[Z_{np}] \end{pmatrix}$$

If $\mathbf{Z}$ is a random vector ($p = 1$), $\mathbb{E}[\boldsymbol{Z}]$ is the vector of the expectations of its components.

**Proposition 3.2** (Properties of Expectation)**.** *Let $\mathbf{Z}, \mathbf{W}$ be random matrices of compatible dimensions, and let $\mathbf{A}, \mathbf{B}, \mathbf{C}$ be constant (non-random, deterministic) matrices of compatible dimensions.*

1. *Linearity:* $\mathbb{E}[\mathbf{Z} + \mathbf{W}] = \mathbb{E}[\mathbf{Z}] + \mathbb{E}[\mathbf{W}]$
2. *Constant Multiplication:* $\mathbb{E}[\mathbf{AZB}] = \mathbf{A}\,\mathbb{E}[\mathbf{Z}]\mathbf{B}$
3. *Affine Transformation:* $\mathbb{E}[\mathbf{AZ} + \mathbf{C}] = \mathbf{A}\,\mathbb{E}[\mathbf{Z}] + \mathbf{C}$ *(Follows from 1 and 2)*

*Proof.* These follow directly from the linearity of the expectation operator applied element-wise. For example, $(\mathbb{E}[\mathbf{AZB}])_{ik} = \mathbb{E}[(\mathbf{AZB})_{ik}] = \mathbb{E}[\sum_j \sum_l A_{ij} Z_{jl} B_{lk}]$. By linearity of scalar expectation, this is $\sum_j \sum_l A_{ij}\,\mathbb{E}[Z_{jl}] B_{lk} = (\mathbf{A}\,\mathbb{E}[\mathbf{Z}]\mathbf{B})_{ik}$. □

**Definition 3.3** (Covariance and Variance Matrices)**.** Let $\boldsymbol{Z} \in \mathbb{R}^p$ and $\boldsymbol{W} \in \mathbb{R}^q$ be random vectors.

1. The **covariance matrix** between $\boldsymbol{Z}$ and $\boldsymbol{W}$ is the $p \times q$ matrix:

$$\operatorname{Cov}(\boldsymbol{Z}, \boldsymbol{W}) = \mathbb{E}\left[(\boldsymbol{Z} - \mathbb{E}[\boldsymbol{Z}])(\boldsymbol{W} - \mathbb{E}[\boldsymbol{W}])^\top\right]$$

The $(i, j)$-th element of this matrix is $\operatorname{Cov}(Z_i, W_j) = \mathbb{E}[(Z_i - \mathbb{E}[Z_i])(W_j - \mathbb{E}[W_j])]$.

2. The **variance-covariance matrix** (or simply variance matrix) of $\boldsymbol{Z}$ is the $p \times p$ matrix:

$$\mathrm{Var}(\boldsymbol{Z}) = \mathrm{Cov}(\boldsymbol{Z}, \boldsymbol{Z}) = \mathbb{E}\left[(\boldsymbol{Z} - \mathbb{E}[\boldsymbol{Z}])(\boldsymbol{Z} - \mathbb{E}[\boldsymbol{Z}])^\top\right]$$

**Claim 3.4.** *The $(i,j)$-th element of* $\mathrm{Var}(\boldsymbol{Z})$ *is* $\mathrm{Cov}(Z_i, Z_j)$*. In particular, the diagonal elements* $(\mathrm{Var}(\boldsymbol{Z}))_{ii}$ *are the variances* $\mathrm{Var}(Z_i)$*, and the off-diagonal elements* $(\mathrm{Var}(\boldsymbol{Z}))_{ij}$ *are the covariances* $\mathrm{Cov}(Z_i, Z_j)$*. Since* $\mathrm{Cov}(Z_i, Z_j) = \mathrm{Cov}(Z_j, Z_i)$*, the variance matrix* $\mathrm{Var}(\boldsymbol{Z})$ *is symmetric.*

*Proof.* Let $\boldsymbol{\mu} = \mathbb{E}[\boldsymbol{Z}]$.

$$\mathrm{Var}(\boldsymbol{Z}) = \mathbb{E}\left[(\boldsymbol{Z} - \boldsymbol{\mu})(\boldsymbol{Z} - \boldsymbol{\mu})^\top\right] = \mathbb{E}\left[\begin{pmatrix} Z_1 - \mu_1 \\ \vdots \\ Z_p - \mu_p \end{pmatrix} \begin{pmatrix} Z_1 - \mu_1 & \cdots & Z_p - \mu_p \end{pmatrix}\right]$$

$$= \mathbb{E}\left[\begin{pmatrix} (Z_1 - \mu_1)^2 & (Z_1 - \mu_1)(Z_2 - \mu_2) & \cdots & (Z_1 - \mu_1)(Z_p - \mu_p) \\ (Z_2 - \mu_1)(Z_1 - \mu_1) & (Z_2 - \mu_2)^2 & \cdots & (Z_2 - \mu_2)(Z_p - \mu_p) \\ \vdots & \vdots & \ddots & \vdots \\ (Z_p - \mu_p)(Z_1 - \mu_1) & (Z_p - \mu_p)(Z_2 - \mu_2) & \cdots & (Z_p - \mu_p)^2 \end{pmatrix}\right]$$

Taking the expectation inside the matrix element-wise:

$$(\mathrm{Var}(\boldsymbol{Z}))_{ij} = \mathbb{E}[(Z_i - \mu_i)(Z_j - \mu_j)] = \mathrm{Cov}(Z_i, Z_j)$$

The diagonal elements are $(\mathrm{Var}(\boldsymbol{Z}))_{ii} = \mathrm{Cov}(Z_i, Z_i) = \mathrm{Var}(Z_i)$. Since $\mathrm{Cov}(Z_i, Z_j) = \mathbb{E}[(Z_i - \mu_i)(Z_j - \mu_j)] = \mathbb{E}[(Z_j - \mu_j)(Z_i - \mu_i)] = \mathrm{Cov}(Z_j, Z_i)$, we have $(\mathrm{Var}(\boldsymbol{Z}))_{ij} = (\mathrm{Var}(\boldsymbol{Z}))_{ji}$, so the matrix is symmetric. □

**Proposition 3.5** (Properties of Covariance Matrices)**.** *Let* $\boldsymbol{Z}, \boldsymbol{W}, \boldsymbol{R}$ *be random vectors (with appropriate dimensions),* $\mathbf{A}, \mathbf{B}$ *be constant matrices, and* $\boldsymbol{a}$ *be a constant vector.*

1. $\mathrm{Cov}(\boldsymbol{Z}, \boldsymbol{W}) = \mathrm{Cov}(\boldsymbol{W}, \boldsymbol{Z})^\top$
2. $\mathrm{Cov}(\boldsymbol{Z} + \boldsymbol{R}, \boldsymbol{W}) = \mathrm{Cov}(\boldsymbol{Z}, \boldsymbol{W}) + \mathrm{Cov}(\boldsymbol{R}, \boldsymbol{W})$
3. $\mathrm{Cov}(\mathbf{A}\boldsymbol{Z}, \mathbf{B}\boldsymbol{W}) = \mathbf{A}\,\mathrm{Cov}(\boldsymbol{Z}, \boldsymbol{W})\mathbf{B}^\top$
4. $\mathrm{Var}(\mathbf{A}\boldsymbol{Z}) = \mathrm{Cov}(\mathbf{A}\boldsymbol{Z}, \mathbf{A}\boldsymbol{Z}) = \mathbf{A}\,\mathrm{Cov}(\boldsymbol{Z}, \boldsymbol{Z})\mathbf{A}^\top = \mathbf{A}\,\mathrm{Var}(\boldsymbol{Z})\mathbf{A}^\top$ *(from 3)*
5. $\mathrm{Var}(\boldsymbol{a}^\top \boldsymbol{Z}) = \boldsymbol{a}^\top\,\mathrm{Var}(\boldsymbol{Z})\boldsymbol{a}$ *(from 4, noting* $\boldsymbol{a}^\top$ *is* $1 \times p$ *and the result is* $1 \times 1$*)*
6. $\mathrm{Var}(\boldsymbol{Z})$ *is a symmetric positive semi-definite matrix. (Symmetry shown above. Positive semi-definiteness follows from 5, since* $\mathrm{Var}(\boldsymbol{a}^\top \boldsymbol{Z})$ *is the variance of a scalar random variable, which must be* $\geq 0$*. So,* $\boldsymbol{a}^\top\,\mathrm{Var}(\boldsymbol{Z})\boldsymbol{a} \geq 0$ *for all* $\boldsymbol{a}$*.)*

*Proof of Property 3.* Let $\mathbb{E}[\boldsymbol{Z}] = \boldsymbol{\mu}_Z$ and $\mathbb{E}[\boldsymbol{W}] = \boldsymbol{\mu}_W$. Then $\mathbb{E}[\mathbf{A}\boldsymbol{Z}] = \mathbf{A}\boldsymbol{\mu}_Z$ and $\mathbb{E}[\mathbf{B}\boldsymbol{W}] = \mathbf{B}\boldsymbol{\mu}_W$.

$$\begin{aligned}
\mathrm{Cov}(\mathbf{A}\boldsymbol{Z}, \mathbf{B}\boldsymbol{W}) &= \mathbb{E}\left[(\mathbf{A}\boldsymbol{Z} - \mathbf{A}\boldsymbol{\mu}_Z)(\mathbf{B}\boldsymbol{W} - \mathbf{B}\boldsymbol{\mu}_W)^\top\right] \\
&= \mathbb{E}\left[\mathbf{A}(\boldsymbol{Z} - \boldsymbol{\mu}_Z)(\mathbf{B}(\boldsymbol{W} - \boldsymbol{\mu}_W))^\top\right] \\
&= \mathbb{E}\left[\mathbf{A}(\boldsymbol{Z} - \boldsymbol{\mu}_Z)(\boldsymbol{W} - \boldsymbol{\mu}_W)^\top \mathbf{B}^\top\right] \\
&= \mathbf{A}\,\mathbb{E}\left[(\boldsymbol{Z} - \boldsymbol{\mu}_Z)(\boldsymbol{W} - \boldsymbol{\mu}_W)^\top\right] \mathbf{B}^\top \quad (\text{Using } \mathbb{E}[\mathbf{CUD}] = \mathbf{C}\,\mathbb{E}[\mathbf{U}]\mathbf{D}) \\
&= \mathbf{A}\,\mathrm{Cov}(\boldsymbol{Z}, \boldsymbol{W})\mathbf{B}^\top
\end{aligned}$$

The other properties can be proven similarly using the definitions and properties of expectation. □

**Question 3.6** (Bernoulli Example - Original from Page 4)**.** Let $X \sim \mathrm{Ber}(p)$ and $Y \sim \mathrm{Ber}(q)$ be Bernoulli random variables. Define $M = XY$. Suppose we know $M \sim \mathrm{Ber}(r)$. Note that $M = 1$ if and only if $X = 1$ and $Y = 1$, so $P(M = 1) = P(X = 1, Y = 1) = r$. Let $\boldsymbol{Z} = (X, Y, M)^\top$.

(a) Find the expectation vector $\mathbb{E}[\boldsymbol{Z}]$ and the covariance matrix $\mathrm{Var}(\boldsymbol{Z})$.

(b) Define the map $A : \mathbb{R}^3 \to \mathbb{R}$ by $A(\boldsymbol{v}) = 2v_1 - 3v_2 + 4v_3 + 7$. Is this map linear? Calculate the expectation and variance of the random variable $W = A(\boldsymbol{Z})$.

(c) Now assume $X$ and $Y$ are independent. Calculate the probability $P(X = 1, Y = 1, M = 1)$.

*Proof.* (a) **Expectation Vector** $\mathbb{E}[\boldsymbol{Z}]$: We need $\mathbb{E}[X]$, $\mathbb{E}[Y]$, and $\mathbb{E}[M]$. Since $X \sim \mathrm{Ber}(p)$, $\mathbb{E}[X] = p$. Since $Y \sim \mathrm{Ber}(q)$, $\mathbb{E}[Y] = q$. Since $M \sim \mathrm{Ber}(r)$, $\mathbb{E}[M] = r$. Therefore,

$$\mathbb{E}[\boldsymbol{Z}] = \begin{pmatrix} p \\ q \\ r \end{pmatrix}$$

**Covariance Matrix** $\mathrm{Var}(\boldsymbol{Z})$: This is a $3 \times 3$ symmetric matrix:

$$\mathrm{Var}(\boldsymbol{Z}) = \begin{pmatrix} \mathrm{Var}(X) & \mathrm{Cov}(X,Y) & \mathrm{Cov}(X,M) \\ \mathrm{Cov}(Y,X) & \mathrm{Var}(Y) & \mathrm{Cov}(Y,M) \\ \mathrm{Cov}(M,X) & \mathrm{Cov}(M,Y) & \mathrm{Var}(M) \end{pmatrix}$$

We calculate the components:

- $\mathrm{Var}(X) = p(1 - p)$
- $\mathrm{Var}(Y) = q(1 - q)$
- $\mathrm{Var}(M) = r(1 - r)$
- $\mathrm{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y] = \mathbb{E}[M] - pq = r - pq$.
- $\mathrm{Cov}(X, M) = \mathbb{E}[XM] - \mathbb{E}[X]\,\mathbb{E}[M]$. Since $X$ is 0 or 1, $X^2 = X$. $M = XY$. So $XM = X(XY) = X^2Y = XY = M$. Thus, $\mathrm{Cov}(X, M) = \mathbb{E}[M] - \mathbb{E}[X]\,\mathbb{E}[M] = r - pr = r(1 - p)$.
- $\mathrm{Cov}(Y, M) = \mathbb{E}[YM] - \mathbb{E}[Y]\,\mathbb{E}[M]$. Similarly, $Y^2 = Y$. $YM = Y(XY) = XY^2 = XY = M$. Thus, $\mathrm{Cov}(Y, M) = \mathbb{E}[M] - \mathbb{E}[Y]\,\mathbb{E}[M] = r - qr = r(1 - q)$.

Using symmetry ($\mathrm{Cov}(A, B) = \mathrm{Cov}(B, A)$), the covariance matrix is:

$$\mathrm{Var}(\boldsymbol{Z}) = \begin{pmatrix} p(1-p) & r - pq & r(1-p) \\ r - pq & q(1-q) & r(1-q) \\ r(1-p) & r(1-q) & r(1-r) \end{pmatrix}$$

(b) **Map** $A(\boldsymbol{v})$ **and** $W = A(\boldsymbol{Z})$: The map $A(\boldsymbol{v}) = 2v_1 - 3v_2 + 4v_3 + 7$ is an **affine map**, not strictly linear because of the constant term $+7$. A linear map requires $A(\mathbf{0}) = \mathbf{0}$. However, we can write $A(\boldsymbol{v}) = \boldsymbol{a}^\top\boldsymbol{v} + c$, where $\boldsymbol{a} = (2, -3, 4)^\top$ and $c = 7$. Let $W = A(\boldsymbol{Z}) = 2X - 3Y + 4M + 7$. **Expectation of W**: Using linearity of expectation:

$$\mathbb{E}[W] = \mathbb{E}[2X - 3Y + 4M + 7] = 2\,\mathbb{E}[X] - 3\,\mathbb{E}[Y] + 4\,\mathbb{E}[M] + 7 = 2p - 3q + 4r + 7$$

Alternatively, using the property $\mathbb{E}[\boldsymbol{a}^\top\boldsymbol{Z} + c] = \boldsymbol{a}^\top\,\mathbb{E}[\boldsymbol{Z}] + c$:

$$\mathbb{E}[W] = \begin{pmatrix} 2 & -3 & 4 \end{pmatrix} \begin{pmatrix} p \\ q \\ r \end{pmatrix} + 7 = (2p - 3q + 4r) + 7$$

**Variance of W**: The constant term does not affect variance: $\mathrm{Var}(W) = \mathrm{Var}(2X - 3Y + 4M)$. Using the property $\mathrm{Var}(\boldsymbol{a}^\top \boldsymbol{Z}) = \boldsymbol{a}^\top \mathrm{Var}(\boldsymbol{Z})\boldsymbol{a}$:

$$\begin{aligned}
\mathrm{Var}(W) &= \begin{pmatrix} 2 & -3 & 4 \end{pmatrix} \mathrm{Var}(\boldsymbol{Z}) \begin{pmatrix} 2 \\ -3 \\ 4 \end{pmatrix} \\
&= \begin{pmatrix} 2 & -3 & 4 \end{pmatrix} \begin{pmatrix} p(1-p) & r - pq & r(1-p) \\ r - pq & q(1-q) & r(1-q) \\ r(1-p) & r(1-q) & r(1-r) \end{pmatrix} \begin{pmatrix} 2 \\ -3 \\ 4 \end{pmatrix}
\end{aligned}$$

Expanding this matrix multiplication gives the variance. For example, the (1,1) term of the result is $4\,\mathrm{Var}(X) + 9\,\mathrm{Var}(Y) + 16\,\mathrm{Var}(M)$, plus cross-terms involving covariances: $2 \times (2)(-3)\,\mathrm{Cov}(X, Y) + 2 \times (2)(4)\,\mathrm{Cov}(X, M) + 2 \times (-3)(4)\,\mathrm{Cov}(Y, M)$.

$$\mathrm{Var}(W) = 4\,\mathrm{Var}(X) + 9\,\mathrm{Var}(Y) + 16\,\mathrm{Var}(M) - 12\,\mathrm{Cov}(X, Y) + 16\,\mathrm{Cov}(X, M) - 24\,\mathrm{Cov}(Y, M)$$

Substituting the expressions for variances and covariances gives the final answer in terms of $p, q, r$.

(c) **Independence Assumption**: If $X$ and $Y$ are independent, then $P(X = 1, Y = 1) = P(X = 1)P(Y = 1) = pq$. We know $P(M = 1) = P(X = 1, Y = 1)$, so under independence, $r = pq$. The question asks for $P(X = 1, Y = 1, M = 1)$. Since $M = XY$, the event $\{M = 1\}$ is exactly the same as the event $\{X = 1 \text{ and } Y = 1\}$. Therefore, $P(X = 1, Y = 1, M = 1) = P(X = 1, Y = 1)$. Under independence, this probability is $pq$. □

**Question 3.7** (Variance Comparison - Original from Page 4). Let $\boldsymbol{Z}, \boldsymbol{W} \in \mathbb{R}^p$ be random vectors. Show that the following are equivalent:

1. For all constant vectors $\boldsymbol{v} \in \mathbb{R}^p$, $\mathrm{Var}(\boldsymbol{v}^\top \boldsymbol{Z}) \geq \mathrm{Var}(\boldsymbol{v}^\top \boldsymbol{W})$.
2. The matrix $\mathbf{B} := \mathrm{Var}(\boldsymbol{Z}) - \mathrm{Var}(\boldsymbol{W})$ is positive semi-definite (psd).
3. The matrix square root $\mathbf{B}^{1/2}$ exists. (Assuming $\mathbf{B}$ is symmetric, which it is since $\mathrm{Var}(\boldsymbol{Z})$ and $\mathrm{Var}(\boldsymbol{W})$ are symmetric).

*Proof.* We will show (1) $\iff$ (2). The equivalence (2) $\iff$ (3) is a standard result from linear algebra for symmetric matrices: a symmetric matrix is positive semi-definite if and only if its (unique, positive semi-definite) square root exists.

**(1) $\implies$ (2)**: Assume $\mathrm{Var}(\boldsymbol{v}^\top \boldsymbol{Z}) \geq \mathrm{Var}(\boldsymbol{v}^\top \boldsymbol{W})$ for all $\boldsymbol{v} \in \mathbb{R}^p$. Using Property 5 of covariance matrices, $\mathrm{Var}(\boldsymbol{v}^\top \boldsymbol{Z}) = \boldsymbol{v}^\top \mathrm{Var}(\boldsymbol{Z})\boldsymbol{v}$ and $\mathrm{Var}(\boldsymbol{v}^\top \boldsymbol{W}) = \boldsymbol{v}^\top \mathrm{Var}(\boldsymbol{W})\boldsymbol{v}$. The assumption becomes:

$$\boldsymbol{v}^\top \mathrm{Var}(\boldsymbol{Z})\boldsymbol{v} \geq \boldsymbol{v}^\top \mathrm{Var}(\boldsymbol{W})\boldsymbol{v} \quad \text{for all } \boldsymbol{v} \in \mathbb{R}^p$$

Rearranging gives:

$$\begin{gathered}
\boldsymbol{v}^\top \mathrm{Var}(\boldsymbol{Z})\boldsymbol{v} - \boldsymbol{v}^\top \mathrm{Var}(\boldsymbol{W})\boldsymbol{v} \geq 0 \\
\boldsymbol{v}^\top (\mathrm{Var}(\boldsymbol{Z}) - \mathrm{Var}(\boldsymbol{W}))\boldsymbol{v} \geq 0
\end{gathered}$$

Let $\mathbf{B} = \mathrm{Var}(\boldsymbol{Z}) - \mathrm{Var}(\boldsymbol{W})$. We have shown that $\boldsymbol{v}^\top \mathbf{B}\boldsymbol{v} \geq 0$ for all $\boldsymbol{v} \in \mathbb{R}^p$. This is the definition of the matrix $\mathbf{B}$ being positive semi-definite.

**(2) $\implies$ (1)**: Assume $\mathbf{B} = \mathrm{Var}(\boldsymbol{Z}) - \mathrm{Var}(\boldsymbol{W})$ is positive semi-definite. By definition, this means $\boldsymbol{v}^\top \mathbf{B}\boldsymbol{v} \geq 0$ for all $\boldsymbol{v} \in \mathbb{R}^p$. Substituting back $\mathbf{B}$:

$$\boldsymbol{v}^\top (\mathrm{Var}(\boldsymbol{Z}) - \mathrm{Var}(\boldsymbol{W}))\boldsymbol{v} \geq 0$$

$$\boldsymbol{v}^\top \mathrm{Var}(\boldsymbol{Z})\boldsymbol{v} - \boldsymbol{v}^\top \mathrm{Var}(\boldsymbol{W})\boldsymbol{v} \geq 0$$

Using Property 5 again:

$$\mathrm{Var}(\boldsymbol{v}^\top \boldsymbol{Z}) - \mathrm{Var}(\boldsymbol{v}^\top \boldsymbol{W}) \geq 0$$
$$\mathrm{Var}(\boldsymbol{v}^\top \boldsymbol{Z}) \geq \mathrm{Var}(\boldsymbol{v}^\top \boldsymbol{W})$$

This holds for all $\boldsymbol{v} \in \mathbb{R}^p$.

Thus, (1) and (2) are equivalent. As noted, (2) is equivalent to (3) for symmetric matrices.

□

# 4 The Linear Model

We now introduce the standard linear regression model.

**Definition 4.1** (Linear Model)**.** The relationship between a response variable $Y_i$ and a set of predictors $X_{i1}, \ldots, X_{ip}$ for observation $i$ ($i = 1, \ldots, n$) is modeled as:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + \epsilon_i = \sum_{j=0}^{p} \beta_j X_{ij} + \epsilon_i$$

where $X_{i0} = 1$ for all $i$. The terms $\epsilon_i$ are random errors, typically assumed to satisfy:

- Zero mean: $\mathbb{E}[\epsilon_i] = 0$ for all $i$.
- Constant variance (homoscedasticity): $\mathrm{Var}(\epsilon_i) = \sigma^2$ for all $i$.
- Uncorrelated errors: $\mathrm{Cov}(\epsilon_i, \epsilon_{i'}) = 0$ for $i \neq i'$.

These assumptions can be summarized as $\mathbb{E}[\boldsymbol{\epsilon}] = \mathbf{0}$ and $\mathrm{Cov}(\boldsymbol{\epsilon}) = \sigma^2 \mathbf{I}_n$, where $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^\top$.

The model parameters are the coefficients $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^\top$ and the error variance $\sigma^2$.

**Matrix Notation**: The model for all $n$ observations can be written compactly using matrices. Let:

$$\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & X_{11} & \cdots & X_{1p} \\ 1 & X_{21} & \cdots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & \cdots & X_{np} \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

Here, $\boldsymbol{Y}$ is $n \times 1$, $\mathbf{X}$ is $n \times (p+1)$, $\boldsymbol{\beta}$ is $(p+1) \times 1$, and $\boldsymbol{\epsilon}$ is $n \times 1$. The model becomes:

$$\boldsymbol{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \text{with } \mathbb{E}[\boldsymbol{\epsilon}] = \mathbf{0}, \quad \mathrm{Cov}(\boldsymbol{\epsilon}) = \sigma^2 \mathbf{I}_n$$

**Assumptions vs. Results**: It's crucial to distinguish between the assumptions we make about the model and the mathematical results derived from those assumptions or from estimation procedures.

Consider the following statements related to the linear model and the Ordinary Least Squares (OLS) estimator $\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \boldsymbol{Y}$. Identify each as primarily an assumption of the model or a mathematical result/definition.

1. $\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{b}} \|\boldsymbol{Y} - \mathbf{X}\boldsymbol{b}\|^2$. **Result/Definition**. This is the definition of the OLS estimator - it's the vector that minimizes the sum of squared residuals. It's derived from the principle of least squares, not assumed about the underlying reality.

2. $\mathbf{X}\boldsymbol{\beta} = \mathbb{E}[\boldsymbol{Y}|\mathbf{X}]$. **Assumption**. This assumes that the conditional expectation of the response vector $\boldsymbol{Y}$, given the predictors $\mathbf{X}$, is a linear function of the parameters $\boldsymbol{\beta}$, defined by the matrix $\mathbf{X}$. This is the core "linearity" assumption. If $\mathbf{X}$ is considered fixed/deterministic, this simplifies to $\mathbb{E}[\boldsymbol{Y}] = \mathbf{X}\boldsymbol{\beta}$.

3. $\mathbb{E}[\epsilon_i] = 0$. **Assumption**. This is a standard assumption about the error terms, implying that the model $\mathbf{X}\boldsymbol{\beta}$ correctly captures the systematic part of $\boldsymbol{Y}$ on average.

4. $\mathbb{E}[\hat{\epsilon}_i] = 0$ (where $\hat{\boldsymbol{\epsilon}} = \boldsymbol{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}$ are the residuals). **Result**. Let $\mathbf{P}_\mathbf{X} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$. Then $\hat{\boldsymbol{\epsilon}} = \boldsymbol{Y} - \mathbf{P}_\mathbf{X}\boldsymbol{Y} = (\mathbf{I} - \mathbf{P}_\mathbf{X})\boldsymbol{Y}$. Assuming $\mathbb{E}[\boldsymbol{Y}|\mathbf{X}] = \mathbf{X}\boldsymbol{\beta}$ (or $\mathbb{E}[\boldsymbol{Y}] = \mathbf{X}\boldsymbol{\beta}$ if $\mathbf{X}$ is fixed), then $\mathbb{E}[\hat{\boldsymbol{\epsilon}}|\mathbf{X}] = \mathbb{E}[(\mathbf{I} - \mathbf{P}_\mathbf{X})\boldsymbol{Y}|\mathbf{X}] = (\mathbf{I} - \mathbf{P}_\mathbf{X})\mathbb{E}[\boldsymbol{Y}|\mathbf{X}] = (\mathbf{I} - \mathbf{P}_\mathbf{X})\mathbf{X}\boldsymbol{\beta}$. Since $\mathbf{P}_\mathbf{X}\mathbf{X} = \mathbf{X}$, this becomes $(\mathbf{X} - \mathbf{P}_\mathbf{X}\mathbf{X})\boldsymbol{\beta} = (\mathbf{X} - \mathbf{X})\boldsymbol{\beta} = \mathbf{0}$. Taking further expectation if $\mathbf{X}$ is random, $\mathbb{E}[\hat{\boldsymbol{\epsilon}}] = \mathbb{E}[\mathbb{E}[\hat{\boldsymbol{\epsilon}}|\mathbf{X}]] = \mathbb{E}[\mathbf{0}] = \mathbf{0}$. Note: The average of the residuals $\frac{1}{n}\sum \hat{\epsilon}_i$ is exactly zero if the model includes an intercept.

5. $\mathbf{X}^\top\hat{\boldsymbol{\epsilon}} = \mathbf{0}$. **Result**. This is a direct consequence of the normal equations used to derive $\hat{\boldsymbol{\beta}}$. $\mathbf{X}^\top(\boldsymbol{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = \mathbf{X}^\top\hat{\boldsymbol{\epsilon}} = \mathbf{0}$. Geometrically, it means the residual vector is orthogonal to the column space of $\mathbf{X}$.

6. $\text{Cov}(\boldsymbol{Y}) = \sigma^2\mathbf{I}$. **Depends/Result under fixed X**. If $\mathbf{X}$ is treated as a fixed, deterministic matrix, then $\boldsymbol{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\mathbf{X}\boldsymbol{\beta}$ is a constant vector. Then $\text{Cov}(\boldsymbol{Y}) = \text{Cov}(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}) = \text{Cov}(\boldsymbol{\epsilon}) = \sigma^2\mathbf{I}$ (using the assumption on $\text{Cov}(\boldsymbol{\epsilon})$). However, if $\mathbf{X}$ is random, $\text{Cov}(\boldsymbol{Y}) = \text{Var}(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon})$ is more complex. Often, the key assumption is about the conditional covariance: $\text{Cov}(\boldsymbol{Y}|\mathbf{X}) = \text{Cov}(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}|\mathbf{X}) = \text{Cov}(\boldsymbol{\epsilon}|\mathbf{X})$. If we assume $\boldsymbol{\epsilon}$ is independent of $\mathbf{X}$ (or just uncorrelated with mean zero conditional on $\mathbf{X}$), then $\text{Cov}(\boldsymbol{\epsilon}|\mathbf{X}) = \text{Cov}(\boldsymbol{\epsilon}) = \sigma^2\mathbf{I}$. So, $\text{Cov}(\boldsymbol{Y}|\mathbf{X}) = \sigma^2\mathbf{I}$ is often considered a consequence of the assumptions when $\mathbf{X}$ is random.

**Question 4.2** (Scenarios for the Linear Model - Original from Page 6)**.** For each scenario below, describe the distributions (or nature) of $\mathbf{X}$, $\boldsymbol{\epsilon}$, $\boldsymbol{Y}$, and $\boldsymbol{Y}|\mathbf{X}$. State which assumptions of the standard linear model ($\mathbb{E}[\boldsymbol{Y}|\mathbf{X}] = \mathbf{X}\boldsymbol{\beta}$, $\text{Cov}(\boldsymbol{Y}|\mathbf{X}) = \sigma^2\mathbf{I}$, errors often normal) hold. Assume $\boldsymbol{\beta}$ is a fixed, unknown vector and $\sigma^2$ is a fixed, unknown positive scalar. Let $Y_i = \mathbf{X}_i^\top\boldsymbol{\beta} + \epsilon_i$ where $\mathbf{X}_i^\top$ is the $i$-th row of $\mathbf{X}$ (potentially including the intercept '1').

1. $\mathbf{X}$ consists of pre-determined constants. $\epsilon_i \sim N(0, \sigma^2)$ i.i.d.

2. $X_i$ are i.i.d. random vectors, e.g., $X_i \sim N(\boldsymbol{\mu}_X, \Sigma_X)$. $\epsilon_i \sim N(0, \sigma^2)$ i.i.d., and $\boldsymbol{\epsilon}$ is independent of $\mathbf{X}$.

3. $X_i \sim U(-1, 1)$ i.i.d. (scalar predictor). Model is $Y_i = \beta_0 + \beta_1 X_i^2 + \epsilon_i$. $\epsilon_i \sim N(0, \sigma^2)$ i.i.d., and $\boldsymbol{\epsilon}$ is independent of $\mathbf{X}$.

4. $\mathbf{X}_{it}$ represents fixed measurements for individual $i$ at time $t$. $\epsilon_{it} \sim N(0, \sigma^2)$ i.i.d. (Panel data).

5. $X_i$ are i.i.d. random vectors $N(\mathbf{0}, \mathbf{I})$. $\epsilon_i \sim N(0, \sigma^2)$ i.i.d., and $\boldsymbol{\epsilon}$ is independent of $\mathbf{X}$. (This seems like a specific instance of case 2).

*Analysis of Scenarios.* 1. **Fixed X, Normal errors**:

- $\mathbf{X}$: Deterministic $n \times (p+1)$ matrix of constants.
- $\boldsymbol{\epsilon}$: Random vector $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I}_n)$.
- $\boldsymbol{Y}$: Random vector. $\boldsymbol{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$. Since $\mathbf{X}\boldsymbol{\beta}$ is constant, $\boldsymbol{Y}$ is a linear transformation of a normal vector, hence normal. $\mathbb{E}[\boldsymbol{Y}] = \mathbf{X}\boldsymbol{\beta} + \mathbb{E}[\boldsymbol{\epsilon}] = \mathbf{X}\boldsymbol{\beta}$. $\text{Cov}(\boldsymbol{Y}) = \text{Cov}(\boldsymbol{\epsilon}) = \sigma^2\mathbf{I}$. So $\boldsymbol{Y} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$.
- $\boldsymbol{Y}|\mathbf{X}$: Since $\mathbf{X}$ is fixed, conditioning on it doesn't change anything. $\boldsymbol{Y}|\mathbf{X} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$.
- Assumptions: All standard assumptions hold. Linearity $\mathbb{E}[\boldsymbol{Y}|\mathbf{X}] = \mathbf{X}\boldsymbol{\beta}$ holds. Homoscedasticity/Uncorrelated errors $\text{Cov}(\boldsymbol{Y}|\mathbf{X}) = \sigma^2\mathbf{I}$ holds. Errors are normal.

2. **Random X, Normal errors, Independent**:
   - $\mathbf{X}$: Random $n \times (p+1)$ matrix, rows $X_i^\top$ are i.i.d. $N(\boldsymbol{\mu}_X, \Sigma_X)$.
   - $\boldsymbol{\epsilon}$: Random vector $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I}_n)$, independent of $\mathbf{X}$.
   - $\boldsymbol{Y}$: Random vector $\boldsymbol{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$. Its marginal distribution is generally complex (not necessarily normal). $\mathbb{E}[\boldsymbol{Y}] = \mathbb{E}[\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}] = \mathbb{E}[\mathbf{X}]\boldsymbol{\beta} + \mathbb{E}[\boldsymbol{\epsilon}] =$ (matrix of $\mu_{X,j})\boldsymbol{\beta} + \mathbf{0}$. $\mathrm{Cov}(\boldsymbol{Y})$ involves variance of $\mathbf{X}\boldsymbol{\beta}$ and $\boldsymbol{\epsilon}$.
   - $\boldsymbol{Y}|\mathbf{X}$: Given a specific realization of $\mathbf{X}$, the matrix $\mathbf{X}$ is now fixed. $\boldsymbol{Y}|\mathbf{X} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}|\mathbf{X}$. Since $\boldsymbol{\epsilon}$ is independent of $\mathbf{X}$, its distribution doesn't change upon conditioning. So $\boldsymbol{\epsilon}|\mathbf{X} \sim N(\mathbf{0}, \sigma^2\mathbf{I}_n)$. Thus, $\boldsymbol{Y}|\mathbf{X} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$.
   - Assumptions: The key assumptions $\mathbb{E}[\boldsymbol{Y}|\mathbf{X}] = \mathbf{X}\boldsymbol{\beta}$ and $\mathrm{Cov}(\boldsymbol{Y}|\mathbf{X}) = \sigma^2\mathbf{I}$ hold. The errors are also normal. The difference from case 1 is that $\mathbf{X}$ is random. Inference is often done conditional on $\mathbf{X}$.

3. **Random X (Uniform), Model with $X^2$, Normal errors**:
   - $\mathbf{X}$: Here, the model is $Y_i = \beta_0 + \beta_1 X_i^2 + \epsilon_i$. We should define the design matrix relative to the model being fitted. Let $Z_i = X_i^2$. The effective design matrix $\mathbf{Z}$ has $i$-th row $(1, Z_i) = (1, X_i^2)$. $X_i \sim U(-1, 1)$ are i.i.d. Thus $\mathbf{Z}$ is a random matrix.
   - $\boldsymbol{\epsilon}$: Random vector $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I}_n)$, independent of $\mathbf{X}$ (and hence of $\mathbf{Z}$).
   - $\boldsymbol{Y}$: Random vector $\boldsymbol{Y} = \mathbf{Z}\boldsymbol{\beta}^* + \boldsymbol{\epsilon}$, where $\boldsymbol{\beta}^* = (\beta_0, \beta_1)^\top$. Marginal distribution is complex.
   - $\boldsymbol{Y}|\mathbf{X}$ (or $\boldsymbol{Y}|\mathbf{Z}$): Given $\mathbf{X}$ (and thus $\mathbf{Z}$), $\boldsymbol{Y}|\mathbf{Z} \sim N(\mathbf{Z}\boldsymbol{\beta}^*, \sigma^2\mathbf{I}_n)$.
   - Assumptions: If we define the model in terms of $Z_i = X_i^2$, then the assumptions $\mathbb{E}[\boldsymbol{Y}|\mathbf{Z}] = \mathbf{Z}\boldsymbol{\beta}^*$ and $\mathrm{Cov}(\boldsymbol{Y}|\mathbf{Z}) = \sigma^2\mathbf{I}$ hold. The errors are normal. Note the linearity is in the parameters $\beta_0, \beta_1$, even though the relationship with the original $X_i$ is quadratic.

4. **Fixed X (Panel), Normal errors**:
   - $\mathbf{X}$: Deterministic $N \times (p+1)$ matrix (where $N = \sum n_i$ or $N \times T$), containing fixed predictor values for individuals over time.
   - $\boldsymbol{\epsilon}$: Random vector $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I}_N)$. (Note: Panel data often has more complex error structures, e.g., correlation within individuals, but here i.i.d. is stated).
   - $\boldsymbol{Y}$: Random vector $\boldsymbol{Y} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_N)$.
   - $\boldsymbol{Y}|\mathbf{X}$: Same as $\boldsymbol{Y}$, $\boldsymbol{Y}|\mathbf{X} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_N)$.
   - Assumptions: Same as case 1. Standard assumptions hold under the stated i.i.d. error structure.

5. **Random X (Std Normal), Normal errors, Independent**:
   - This is a specific case of scenario 2, with $\boldsymbol{\mu}_X = \mathbf{0}$ and $\Sigma_X = \mathbf{I}$.
   - $\mathbf{X}$: Random matrix, rows $X_i^\top$ are i.i.d. $N(\mathbf{0}, \mathbf{I})$.
   - $\boldsymbol{\epsilon}$: Random vector $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I}_n)$, independent of $\mathbf{X}$.
   - $\boldsymbol{Y}$: Random vector. $\mathbb{E}[\boldsymbol{Y}] = \mathbb{E}[\mathbf{X}]\boldsymbol{\beta} + \mathbf{0} = \mathbf{0}\boldsymbol{\beta} + \mathbf{0} = \mathbf{0}$.
   - $\boldsymbol{Y}|\mathbf{X}$: $\boldsymbol{Y}|\mathbf{X} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$.
   - Assumptions: Same as case 2. The key conditional assumptions hold. $\mathbb{E}[\boldsymbol{Y}|\mathbf{X}] = \mathbf{X}\boldsymbol{\beta}$ and $\mathrm{Cov}(\boldsymbol{Y}|\mathbf{X}) = \sigma^2\mathbf{I}$. Errors are normal.

□

# 5 Miscellaneous Results and Proofs

**Question 5.1** (Rank-1 Projection Matrix - Original from Page 6)**.** Let $\boldsymbol{v} \in \mathbb{R}^n$ be a non-zero vector ($\boldsymbol{v} \neq \mathbf{0}$). Show that the matrix $\mathbf{P} = \frac{\boldsymbol{v}\boldsymbol{v}^\top}{\|\boldsymbol{v}\|^2}$ is an orthogonal projection matrix. What is its rank?

*Proof.* Recall $\|\boldsymbol{v}\|^2 = \boldsymbol{v}^\top\boldsymbol{v}$. The matrix is $\mathbf{P} = \frac{1}{\boldsymbol{v}^\top\boldsymbol{v}}\boldsymbol{v}\boldsymbol{v}^\top$.

1. **Symmetry**:

$$\mathbf{P}^\top = \left(\frac{1}{\boldsymbol{v}^\top\boldsymbol{v}}\boldsymbol{v}\boldsymbol{v}^\top\right)^\top = \frac{1}{\boldsymbol{v}^\top\boldsymbol{v}}(\boldsymbol{v}\boldsymbol{v}^\top)^\top = \frac{1}{\boldsymbol{v}^\top\boldsymbol{v}}(\boldsymbol{v}^\top)^\top\boldsymbol{v}^\top = \frac{1}{\boldsymbol{v}^\top\boldsymbol{v}}\boldsymbol{v}\boldsymbol{v}^\top = \mathbf{P}$$

   So $\mathbf{P}$ is symmetric.

2. **Idempotency**:

$$\begin{aligned}\mathbf{P}^2 &= \left(\frac{1}{\boldsymbol{v}^\top\boldsymbol{v}}\boldsymbol{v}\boldsymbol{v}^\top\right)\left(\frac{1}{\boldsymbol{v}^\top\boldsymbol{v}}\boldsymbol{v}\boldsymbol{v}^\top\right) \\ &= \frac{1}{(\boldsymbol{v}^\top\boldsymbol{v})^2}\boldsymbol{v}(\boldsymbol{v}^\top\boldsymbol{v})\boldsymbol{v}^\top \\ &= \frac{1}{(\boldsymbol{v}^\top\boldsymbol{v})^2}\boldsymbol{v}(\|\boldsymbol{v}\|^2)\boldsymbol{v}^\top \\ &= \frac{\|\boldsymbol{v}\|^2}{(\|\boldsymbol{v}\|^2)^2}\boldsymbol{v}\boldsymbol{v}^\top \\ &= \frac{1}{\|\boldsymbol{v}\|^2}\boldsymbol{v}\boldsymbol{v}^\top = \mathbf{P}\end{aligned}$$

   So $\mathbf{P}$ is idempotent.

Since $\mathbf{P}$ is symmetric and idempotent, it is an orthogonal projection matrix.

**Rank**: The matrix $\boldsymbol{v}\boldsymbol{v}^\top$ is an $n \times n$ matrix. Any column of this matrix is a multiple of $\boldsymbol{v}$. For example, column $j$ is $v_j\boldsymbol{v}$. Since $\boldsymbol{v} \neq \mathbf{0}$, the columns are non-zero (unless $v_j = 0$) and are all multiples of $\boldsymbol{v}$. Thus, the column space is spanned by the single non-zero vector $\boldsymbol{v}$. The dimension of the column space (the rank) is 1. Alternatively, the image of $\mathbf{P}$ is $\text{Im}(\mathbf{P}) = \{\mathbf{P}\boldsymbol{x} \mid \boldsymbol{x} \in \mathbb{R}^n\} = \{\frac{\boldsymbol{v}(\boldsymbol{v}^\top\boldsymbol{x})}{\|\boldsymbol{v}\|^2} \mid \boldsymbol{x} \in \mathbb{R}^n\}$. Since $\boldsymbol{v}^\top\boldsymbol{x}$ is a scalar, any vector in the image is a scalar multiple of $\boldsymbol{v}$. Thus $\text{Im}(\mathbf{P}) = \text{span}\{\boldsymbol{v}\}$. Since $\boldsymbol{v} \neq \mathbf{0}$, the dimension of this space is 1. So, $\text{rank}(\mathbf{P}) = 1$. It projects vectors onto the line spanned by $\boldsymbol{v}$. □

**Question 5.2** (Unbiased Sample Variance - Original from Page 6)**.** Let $Y_1, \ldots, Y_n$ be i.i.d. random variables with mean $\mu$ and variance $\sigma^2$ (parameters unknown). Show that the sample variance $S_n^2 = \frac{1}{n-1}\sum_{i=1}^n (Y_i - \bar{Y})^2$, where $\bar{Y} = \frac{1}{n}\sum Y_i$, is an unbiased estimator for $\sigma^2$, i.e., $\mathbb{E}[S_n^2] = \sigma^2$.

*Proof.* We need to compute $\mathbb{E}[\sum_{i=1}^{n}(Y_i - \bar{Y})^2]$. Let's expand the term inside the sum:

$$\begin{aligned}
\sum_{i=1}^{n}(Y_i - \bar{Y})^2 &= \sum_{i=1}^{n}(Y_i - \mu - (\bar{Y} - \mu))^2 \\
&= \sum_{i=1}^{n}[(Y_i - \mu)^2 - 2(Y_i - \mu)(\bar{Y} - \mu) + (\bar{Y} - \mu)^2] \\
&= \sum_{i=1}^{n}(Y_i - \mu)^2 - 2(\bar{Y} - \mu)\sum_{i=1}^{n}(Y_i - \mu) + \sum_{i=1}^{n}(\bar{Y} - \mu)^2 \\
&= \sum_{i=1}^{n}(Y_i - \mu)^2 - 2(\bar{Y} - \mu)(n\bar{Y} - n\mu) + n(\bar{Y} - \mu)^2 \\
&= \sum_{i=1}^{n}(Y_i - \mu)^2 - 2n(\bar{Y} - \mu)^2 + n(\bar{Y} - \mu)^2 \\
&= \sum_{i=1}^{n}(Y_i - \mu)^2 - n(\bar{Y} - \mu)^2
\end{aligned}$$

Now, take the expectation:

$$\mathbb{E}\left[\sum_{i=1}^{n}(Y_i - \bar{Y})^2\right] = \mathbb{E}\left[\sum_{i=1}^{n}(Y_i - \mu)^2\right] - n\,\mathbb{E}\left[(\bar{Y} - \mu)^2\right]$$

By linearity of expectation:

$$= \sum_{i=1}^{n}\mathbb{E}[(Y_i - \mu)^2] - n\,\mathbb{E}[(\bar{Y} - \mu)^2]$$

We know $\mathbb{E}[(Y_i - \mu)^2] = \mathrm{Var}(Y_i) = \sigma^2$. And $\mathbb{E}[(\bar{Y} - \mu)^2] = \mathrm{Var}(\bar{Y})$. Since $Y_i$ are i.i.d., $\mathrm{Var}(\bar{Y}) = \mathrm{Var}(\frac{1}{n}\sum Y_i) = \frac{1}{n^2}\sum \mathrm{Var}(Y_i) = \frac{1}{n^2}(n\sigma^2) = \frac{\sigma^2}{n}$. Substituting these back:

$$\mathbb{E}\left[\sum_{i=1}^{n}(Y_i - \bar{Y})^2\right] = \sum_{i=1}^{n}\sigma^2 - n\left(\frac{\sigma^2}{n}\right) = n\sigma^2 - \sigma^2 = (n-1)\sigma^2$$

Therefore,

$$\mathbb{E}[S_n^2] = \mathbb{E}\left[\frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2\right] = \frac{1}{n-1}\mathbb{E}\left[\sum_{i=1}^{n}(Y_i - \bar{Y})^2\right] = \frac{1}{n-1}(n-1)\sigma^2 = \sigma^2$$

Thus, $S_n^2$ is an unbiased estimator for $\sigma^2$. The division by $n-1$ (Bessel's correction) is necessary for unbiasedness. □

**Theorem 5.3** (Distribution of Sample Variance under Normality)**.** *Assume $Y_1, \ldots, Y_n \sim N(\mu, \sigma^2)$ i.i.d. Then the random variable*

$$\frac{(n-1)S_n^2}{\sigma^2} = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{\sigma^2}$$

*follows a Chi-squared distribution with $n-1$ degrees of freedom, denoted $\chi^2_{n-1}$. Furthermore, $S_n^2$ is independent of $\bar{Y}$.*

*Proof.* This is a standard result, often proven using Cochran's Theorem or properties of orthogonal transformations (like the Helmert transformation) preserving normality and independence. The proof is beyond the scope of this summary but is fundamental in statistical inference (e.g., for t-tests and confidence intervals for $\mu$). □

**Question 5.4** (Expected Squared Norm - Original from Page 7)**.** Let $\boldsymbol{Z} \in \mathbb{R}^m$ be a random vector.

1. Show that $\mathbb{E}[\|\boldsymbol{Z}\|^2] = \operatorname{tr}(\mathbb{E}[\boldsymbol{Z}\boldsymbol{Z}^\top])$.
2. Deduce that if $\mathbb{E}[\boldsymbol{Z}] = \mathbf{0}$, then $\mathbb{E}[\|\boldsymbol{Z}\|^2] = \operatorname{tr}(\operatorname{Var}(\boldsymbol{Z}))$.

Justify each step.

*Proof.*

1. We start with the definition of the squared Euclidean norm:

$$\|\boldsymbol{Z}\|^2 = \sum_{i=1}^m Z_i^2$$

Taking the expectation, and using linearity of expectation:

$$\mathbb{E}[\|\boldsymbol{Z}\|^2] = \mathbb{E}\left[\sum_{i=1}^m Z_i^2\right] = \sum_{i=1}^m \mathbb{E}[Z_i^2]$$

Now consider the matrix $\mathbf{M} = \mathbb{E}[\boldsymbol{Z}\boldsymbol{Z}^\top]$. This is an $m \times m$ matrix. The $(i, j)$-th element of the random matrix $\boldsymbol{Z}\boldsymbol{Z}^\top$ is $Z_i Z_j$. The $(i, j)$-th element of $\mathbf{M}$ is $M_{ij} = \mathbb{E}[Z_i Z_j]$. The trace of $\mathbf{M}$ is the sum of its diagonal elements:

$$\operatorname{tr}(\mathbf{M}) = \operatorname{tr}(\mathbb{E}[\boldsymbol{Z}\boldsymbol{Z}^\top]) = \sum_{i=1}^m M_{ii} = \sum_{i=1}^m \mathbb{E}[Z_i Z_i] = \sum_{i=1}^m \mathbb{E}[Z_i^2]$$

Comparing the results, we see that:

$$\mathbb{E}[\|\boldsymbol{Z}\|^2] = \operatorname{tr}(\mathbb{E}[\boldsymbol{Z}\boldsymbol{Z}^\top])$$

2. Now, assume $\mathbb{E}[\boldsymbol{Z}] = \mathbf{0}$. Recall the definition of the variance-covariance matrix:

$$\operatorname{Var}(\boldsymbol{Z}) = \mathbb{E}[(\boldsymbol{Z} - \mathbb{E}[\boldsymbol{Z}])(\boldsymbol{Z} - \mathbb{E}[\boldsymbol{Z}])^\top]$$

If $\mathbb{E}[\boldsymbol{Z}] = \mathbf{0}$, this simplifies to:

$$\operatorname{Var}(\boldsymbol{Z}) = \mathbb{E}[(\boldsymbol{Z} - \mathbf{0})(\boldsymbol{Z} - \mathbf{0})^\top] = \mathbb{E}[\boldsymbol{Z}\boldsymbol{Z}^\top]$$

Substituting this into the result from part (1):

$$\mathbb{E}[\|\boldsymbol{Z}\|^2] = \operatorname{tr}(\mathbb{E}[\boldsymbol{Z}\boldsymbol{Z}^\top]) = \operatorname{tr}(\operatorname{Var}(\boldsymbol{Z}))$$

This relationship is useful, for instance, in calculating expected prediction errors. The expected squared norm of a zero-mean random vector is the sum of the variances of its components (since $\operatorname{tr}(\operatorname{Var}(\boldsymbol{Z})) = \sum \operatorname{Var}(Z_i)$).

□