Data Science for Mathematicians
Exercises 2: Linear Regression from a Geometric Perspective

Instructions

Answer all exercises completely. Show all working, justify your answers, and state any assumptions you make. For computational exercises, carry out all intermediate steps explicitly. For proof exercises, clearly identify which definitions and theorems you are applying.

Exercises

Exercise 1. (Orthogonality Verification) Consider the data:

$$X = \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 4 \\ 1 & 5 \end{pmatrix}, \quad \mathbf{y} = \begin{pmatrix} 3 \\ 5 \\ 8 \\ 11 \end{pmatrix}.$$

(a) Compute the OLS estimator $\hat{\boldsymbol{\beta}}$ using the formula $(X^\top X)^{-1} X^\top \mathbf{y}$.

(b) Compute the residual vector $\mathbf{e} = \mathbf{y} - X\hat{\boldsymbol{\beta}}$.

(c) Verify explicitly that $X^\top \mathbf{e} = \mathbf{0}$.

(d) Without further computation, explain why $\mathbf{1}^\top \mathbf{e} = 0$ and $\sum_{i=1}^{n} x_i e_i = 0$, where $\mathbf{1} = (1, 1, 1, 1)^\top$ and x_i denotes the value of the second column for observation i.

Exercise 2. (Null Space and Rank Deficiency) Consider the design matrix

$$X = \begin{pmatrix} 1 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 3 \\ 0 & 0 & 2 \end{pmatrix} \in \mathbb{R}^{4\times 3}.$$

(a) Determine the rank of X and find a basis for $\text{Null}(X)$.

(b) Verify that the dimension formula $\dim(\text{Col}(X)) + \dim(\text{Null}(X)) = p$ holds, where p is the number of columns.

(c) Explain why the normal equation $(X^\top X)\boldsymbol{\beta} = X^\top \mathbf{y}$ do not have a unique solution for this matrix.

(d) If $\mathbf{y} = (3, 6, 5, 2)^\top$, find all least squares solutions and identify the minimum-norm solution.

Exercise 3. (Right Null Space)

In this lecture, we studied the left null space $\text{Null}(X^\top)$ and established its relationship to the column space via **??**. The **right null space** (or simply the **null space**) of a matrix $X \in \mathbb{R}^{n\times p}$ is defined as

$$\text{Null}(X) = \{\mathbf{v} \in \mathbb{R}^p : X\mathbf{v} = \mathbf{0}\}.$$

This subspace consists of all vectors in the domain that are mapped to zero by X.

Prove the following results:

(a) Prove that the null space of X is the orthogonal complement of the row space of X:

$$\text{Null}(X) = \text{Row}(X)^\perp.$$

Hint: Recall that $\text{Row}(X) = \text{Col}(X^\top)$.

(b) Prove that the dimension of the null space and the rank of X satisfy

$$\dim(\text{Null}(X)) + \text{rank}(X) = p,$$

where p is the number of columns of X.

Hint: Use the Orthogonal Decomposition Theorem applied to $\mathbb{R}^p = \text{Row}(X) \oplus \text{Row}(X)^\perp$.

(c) Prove that for any matrix $X \in \mathbb{R}^{n\times p}$,

$$\text{Null}(X^\top X) = \text{Null}(X).$$

Use this result to explain why $X^\top X$ is invertible if and only if X has full column rank.

Exercise 4. (Generalized Pythagorean Theorem) Prove by induction that if $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k \in V$ are mutually orthogonal vectors in an inner product space (i.e., $\langle \mathbf{v}_i, \mathbf{v}_j \rangle = 0$ for all $i \neq j$), then

$$\left\| \sum_{i=1}^{k} \mathbf{v}_i \right\|^2 = \sum_{i=1}^{k} \|\mathbf{v}_i\|^2.$$

Exercise 5. (Properties of the Residual Maker Matrix) Let $X \in \mathbb{R}^{n\times p}$ have linearly independent columns, and let $M = I_n - P$ where $P = X(X^\top X)^{-1}X^\top$ is the hat matrix.

(a) Prove that M is symmetric.

(b) Prove that M is idempotent, i.e., $M^2 = M$.

(c) Prove that $MX = O$ (the zero matrix).

(d) Prove that $\text{tr}(M) = n - p$.

(e) Interpret parts (c) and (d) geometrically in terms of the column space of X.

Exercise 6. (Condition Number Analysis) Consider the nearly collinear design matrix

$$X = \begin{pmatrix} 1 & 1.000 \\ 1 & 1.001 \\ 1 & 1.002 \end{pmatrix}.$$

(a) Compute the Gram matrix $X^\top X$.

(b) Find the eigenvalues of $X^\top X$ and compute the condition number $\kappa(X^\top X)$.

(c) Explain why solving the normal equations directly would be numerically problematic for this matrix.

(d) How does $\kappa(X^\top X)$ relate to $\kappa(X)$?

Exercise 7. (Rank and Trace of Projection Matrices) Let $P \in \mathbb{R}^{n\times n}$ be any symmetric idempotent matrix (i.e., $P^\top = P$ and $P^2 = P$).

(a) Prove that all eigenvalues of P are either 0 or 1. *Hint: If $P\mathbf{v} = \lambda\mathbf{v}$, apply P to both sides.*

(b) Prove that $\operatorname{rank}(P) = \operatorname{tr}(P)$.

(c) Explain how this result is consistent with Theorem **??**.

Exercise 8. (Centering and the Intercept) Consider the simple linear regression model $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$. Define the centered variables $\tilde{x}_i = x_i - \bar{x}$ and $\tilde{y}_i = y_i - \bar{y}$, where $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$ and $\bar{y} = \frac{1}{n}\sum_{i=1}^n y_i$.

(a) Show that the OLS estimator for the slope in the original model is

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$

(b) Prove that if we regress $\tilde{\mathbf{y}}$ on $\tilde{\mathbf{x}}$ (without an intercept), the resulting coefficient equals $\hat{\beta}_1$.

(c) Prove that $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$.

Exercise 9. (Uniqueness of the Best Approximation) The Best Approximation Theorem states that for a finite-dimensional subspace W of an inner product space V and any $\mathbf{y} \in V$, there exists a unique vector $\hat{\mathbf{y}} \in W$ minimizing $\|\mathbf{y} - \mathbf{w}\|$ over all $\mathbf{w} \in W$. Provide a complete proof of the uniqueness assertion. *Hint: Suppose $\hat{\mathbf{y}}_1$ and $\hat{\mathbf{y}}_2$ both achieve the minimum, and use the parallelogram law:*

$$\|\mathbf{u} + \mathbf{v}\|^2 + \|\mathbf{u} - \mathbf{v}\|^2 = 2\|\mathbf{u}\|^2 + 2\|\mathbf{v}\|^2.$$

Exercise 10. (SVD and the Pseudoinverse) Consider the rank-deficient matrix

$$X = \begin{pmatrix} 1 & 2 \\ 2 & 4 \\ 1 & 2 \end{pmatrix}, \quad \mathbf{y} = \begin{pmatrix} 3 \\ 7 \\ 2 \end{pmatrix}.$$

(a) Verify that X has rank 1 and find a basis for $\operatorname{Col}(X)$ and $\operatorname{Null}(X)$.

(b) Compute the SVD of X by first finding the eigenvalues and eigenvectors of $X^\top X$.

(c) Using the SVD, compute the Moore-Penrose pseudoinverse X^+.

(d) Find the minimum-norm least squares solution $\hat{\boldsymbol{\beta}} = X^+\mathbf{y}$.

(e) Verify that among all solutions to the normal equations, $\hat{\boldsymbol{\beta}}$ has the smallest Euclidean norm.

Exercise 11. (Fundamental Subspaces) Let $X \in \mathbb{R}^{n \times p}$ be an arbitrary matrix.

(a) Prove that $\text{Row}(X) = \text{Col}(X^\top)$.

(b) Prove that $\text{Null}(X) = \text{Row}(X)^\perp$ (where orthogonality is in $\mathbb{R}^p$).

(c) Using parts (a) and (b), derive the dimension formula: $\dim(\text{Row}(X)) + \dim(\text{Null}(X)) = p$.

(d) How does this relate to the SVD partition in Theorem **??**?

Exercise 12. (Detecting Multicollinearity) Consider the data matrix

$$X = \begin{pmatrix} 1 & 2 & 5 \\ 1 & 3 & 7 \\ 1 & 4 & 9 \\ 1 & 5 & 11 \end{pmatrix}, \quad \mathbf{y} = \begin{pmatrix} 1 \\ 2 \\ 3 \\ 5 \end{pmatrix}.$$

(a) Show that the columns of X are linearly dependent by finding a non-trivial linear combination that equals the zero vector.

(b) Explain why the normal equations have infinitely many solutions.

(c) Find all solutions to the normal equations $(X^\top X)\boldsymbol{\beta} = X^\top \mathbf{y}$.

(d) Compute the projection $\hat{\mathbf{y}} = X\boldsymbol{\beta}$ and verify it is the same for all solutions found in part (c).

Exercise 13. (Orthogonality of Residuals and Fitted Values) Let $\hat{\mathbf{y}} = P\mathbf{y}$ be the vector of fitted values and $\mathbf{e} = M\mathbf{y}$ be the residual vector, where P is the hat matrix and $M = I - P$.

(a) Prove that $\hat{\mathbf{y}}^\top \mathbf{e} = 0$, i.e., the fitted values and residuals are orthogonal.

(b) Prove the Pythagorean decomposition: $\|\mathbf{y}\|^2 = \|\hat{\mathbf{y}}\|^2 + \|\mathbf{e}\|^2$.

(c) Define the coefficient of determination as $R^2 = \|\hat{\mathbf{y}} - \bar{y}\mathbf{1}\|^2 / \|\mathbf{y} - \bar{y}\mathbf{1}\|^2$ where $\bar{y} = \frac{1}{n}\sum_i y_i$. Show that $0 \leq R^2 \leq 1$ when the model includes an intercept.

Exercise 14. (Equivalence of Orthogonality and First-Order Conditions)

(a) Starting from the geometric orthogonality condition $X^\top(\mathbf{y} - X\boldsymbol{\beta}) = \mathbf{0}$, derive the normal equations.

(b) Starting from the loss function $L(\boldsymbol{\beta}) = \|\mathbf{y} - X\boldsymbol{\beta}\|_2^2$, compute the gradient $\nabla_{\boldsymbol{\beta}} L$ and show that setting it to zero yields the same normal equations.

(c) Explain why this equivalence is not coincidental, but reflects a deep connection between orthogonal projection and least squares minimization in Euclidean space.

Exercise 15. (QR Decomposition Approach) Let $X = QR$ be the (reduced) QR decomposition of a full column rank matrix $X \in \mathbb{R}^{n \times p}$, where $Q \in \mathbb{R}^{n \times p}$ has orthonormal columns and $R \in \mathbb{R}^{p \times p}$ is upper triangular with positive diagonal entries.

(a) Prove that $Q^\top Q = I_p$ and that $X^\top X = R^\top R$.

(b) Show that the OLS estimator satisfies $R\hat{\boldsymbol{\beta}} = Q^\top \mathbf{y}$.

(c) Prove that the hat matrix can be written as $P = QQ^\top$.

(d) Explain why solving $R\hat{\boldsymbol{\beta}} = Q^\top \mathbf{y}$ via back-substitution is numerically superior to computing $(X^\top X)^{-1} X^\top \mathbf{y}$ directly.

Exercise 16. (Leverage and Influence) The diagonal elements h_{ii} of the hat matrix P are called *leverage* values.

(a) Prove that $0 \leq h_{ii} \leq 1$ for all i. *Hint: Use $P = P^2$ and $P = P^\top$.*

(b) Prove that $\sum_{i=1}^{n} h_{ii} = p + 1$ (the number of parameters including the intercept).

(c) Show that $\hat{y}_i = h_{ii} y_i + \sum_{j \neq i} h_{ij} y_j$, and interpret this as a weighted average of the observed responses.

(d) Prove that for simple linear regression with centered predictor, $h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{j=1}^{n} (x_j - \bar{x})^2}$. What does this reveal about observations with extreme predictor values?

Exercise 17. (Alternative Loss Functions and Non-Orthogonal Projections) The manuscript establishes that minimizing $\|\mathbf{y} - X\boldsymbol{\beta}\|_2^2$ yields the orthogonal projection of $\mathbf{y}$ onto $\mathrm{Col}(X)$.

(a) Consider minimizing $\|\mathbf{y} - X\boldsymbol{\beta}\|_1 = \sum_{i=1}^{n} |y_i - (X\boldsymbol{\beta})_i|$ (the L^1 or LAD loss). Explain why the solution is generally *not* an orthogonal projection.

(b) Consider the weighted loss $L_W(\boldsymbol{\beta}) = (\mathbf{y} - X\boldsymbol{\beta})^\top W (\mathbf{y} - X\boldsymbol{\beta})$, where $W \in \mathbb{R}^{n \times n}$ is a symmetric positive definite matrix. Derive the first-order conditions and find the minimizer $\hat{\boldsymbol{\beta}}_W$.

(c) Show that the weighted least squares solution corresponds to an orthogonal projection with respect to the inner product $\langle \mathbf{u}, \mathbf{v} \rangle_W = \mathbf{u}^\top W \mathbf{v}$.

(d) What is the geometric interpretation of the weighting matrix W?

Solutions

Solution 1. (a) We compute:

$$X^\top X = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 4 & 5 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 4 \\ 1 & 5 \end{pmatrix} = \begin{pmatrix} 4 & 12 \\ 12 & 46 \end{pmatrix}.$$

The determinant is $4 \cdot 46 - 12 \cdot 12 = 184 - 144 = 40$, so

$$(X^\top X)^{-1} = \frac{1}{40} \begin{pmatrix} 46 & -12 \\ -12 & 4 \end{pmatrix}.$$

Next,

$$X^\top \mathbf{y} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 4 & 5 \end{pmatrix} \begin{pmatrix} 3 \\ 5 \\ 8 \\ 11 \end{pmatrix} = \begin{pmatrix} 27 \\ 98 \end{pmatrix}.$$

Therefore,

$$\hat{\boldsymbol{\beta}} = \frac{1}{40} \begin{pmatrix} 46 & -12 \\ -12 & 4 \end{pmatrix} \begin{pmatrix} 27 \\ 98 \end{pmatrix} = \frac{1}{40} \begin{pmatrix} 1242 - 1176 \\ -324 + 392 \end{pmatrix} = \frac{1}{40} \begin{pmatrix} 66 \\ 68 \end{pmatrix} = \begin{pmatrix} 1.65 \\ 1.7 \end{pmatrix}.$$

(b) The fitted values are:

$$\hat{\mathbf{y}} = X\hat{\boldsymbol{\beta}} = \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 4 \\ 1 & 5 \end{pmatrix} \begin{pmatrix} 1.65 \\ 1.7 \end{pmatrix} = \begin{pmatrix} 3.35 \\ 5.05 \\ 8.45 \\ 10.15 \end{pmatrix}.$$

The residuals are:

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = \begin{pmatrix} 3 - 3.35 \\ 5 - 5.05 \\ 8 - 8.45 \\ 11 - 10.15 \end{pmatrix} = \begin{pmatrix} -0.35 \\ -0.05 \\ -0.45 \\ 0.85 \end{pmatrix}.$$

(c) We verify:

$$X^\top \mathbf{e} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 4 & 5 \end{pmatrix} \begin{pmatrix} -0.35 \\ -0.05 \\ -0.45 \\ 0.85 \end{pmatrix} = \begin{pmatrix} -0.35 - 0.05 - 0.45 + 0.85 \\ -0.35 - 0.10 - 1.80 + 4.25 \end{pmatrix} = \begin{pmatrix} 0 \\ 2 \end{pmatrix}.$$

Note: Due to rounding, we get approximately zero. With exact arithmetic using $\hat{\beta}_0 = 33/20$ and $\hat{\beta}_1 = 17/10$, we obtain exactly $X^\top \mathbf{e} = \mathbf{0}$.

(d) The condition $X^\top \mathbf{e} = \mathbf{0}$ means that each column of X is orthogonal to $\mathbf{e}$. The first column of X is $\mathbf{1}$, so $\mathbf{1}^\top \mathbf{e} = 0$, which is equivalent to $\sum_i e_i = 0$. The second column contains the x_i values, so its orthogonality to $\mathbf{e}$ gives $\sum_i x_i e_i = 0$.

Solution 2. (a) We perform row reduction on X to determine its rank:

$$X = \begin{pmatrix} 1 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 3 \\ 0 & 0 & 2 \end{pmatrix} \xrightarrow{R_2-2R_1} \begin{pmatrix} 1 & 2 & 1 \\ 0 & 0 & 0 \\ 1 & 2 & 3 \\ 0 & 0 & 2 \end{pmatrix} \xrightarrow{R_3-R_1} \begin{pmatrix} 1 & 2 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 2 \\ 0 & 0 & 2 \end{pmatrix} \xrightarrow{R_4-R_3} \begin{pmatrix} 1 & 2 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 2 \\ 0 & 0 & 0 \end{pmatrix}.$$

The row echelon form has two pivots (in columns 1 and 3), so $\operatorname{rank}(X) = 2$. The second column is a free variable. To find $\operatorname{Null}(X)$, we solve $X\mathbf{v} = \mathbf{0}$. Setting $v_2 = t$ (free parameter), the system gives $v_3 = 0$ from the third row, and $v_1 + 2v_2 + v_3 = 0$ from the first row, yielding $v_1 = -2t$. Thus,

$$\operatorname{Null}(X) = \operatorname{span}\left\{ \begin{pmatrix} -2 \\ 1 \\ 0 \end{pmatrix} \right\}.$$

A basis for $\operatorname{Null}(X)$ is $\left\{(-2, 1, 0)^\top\right\}$.

(b) We verify the Rank-Nullity Theorem. The number of columns is $p = 3$. We have:

$$\dim(\operatorname{Col}(X)) + \dim(\operatorname{Null}(X)) = \operatorname{rank}(X) + \operatorname{nullity}(X) = 2 + 1 = 3 = p. \quad \checkmark$$

(c) Since $\operatorname{Null}(X) \neq \{\mathbf{0}\}$, the columns of X are linearly dependent. By **??**, this implies $\operatorname{Null}(X^\top X) = \operatorname{Null}(X) \neq \{\mathbf{0}\}$, and hence the Gram matrix $X^\top X$ is singular by **??**. Consequently, the Normal Equations cannot be solved by multiplying by $(X^\top X)^{-1}$.

Geometrically, multiple coefficient vectors $\boldsymbol{\beta}$ produce the same fitted value $\hat{\mathbf{y}} = X\boldsymbol{\beta}$. Specifically, if $\hat{\boldsymbol{\beta}}$ is any solution, then $\hat{\boldsymbol{\beta}} + \mathbf{v}$ is also a solution for any $\mathbf{v} \in \operatorname{Null}(X)$, since $X(\hat{\boldsymbol{\beta}} + \mathbf{v}) = X\hat{\boldsymbol{\beta}} + X\mathbf{v} = X\hat{\boldsymbol{\beta}}$.

(d) We first find one particular least squares solution. Since the second column of X equals twice the first column, we can eliminate redundancy by working with a reduced matrix. Setting $\beta_2 = 0$ and solving with the remaining columns:

$$X_{\text{reduced}} = \begin{pmatrix} 1 & 1 \\ 2 & 2 \\ 1 & 3 \\ 0 & 2 \end{pmatrix}, \quad X_{\text{reduced}}^\top X_{\text{reduced}} = \begin{pmatrix} 6 & 10 \\ 10 & 18 \end{pmatrix}, \quad X_{\text{reduced}}^\top \mathbf{y} = \begin{pmatrix} 20 \\ 34 \end{pmatrix}.$$

Solving $(X_{\text{reduced}}^\top X_{\text{reduced}})\boldsymbol{\gamma} = X_{\text{reduced}}^\top \mathbf{y}$:

$$(X_{\text{reduced}}^\top X_{\text{reduced}})^{-1} = \frac{1}{8}\begin{pmatrix} 18 & -10 \\ -10 & 6 \end{pmatrix}, \quad \boldsymbol{\gamma} = \frac{1}{8}\begin{pmatrix} 18 & -10 \\ -10 & 6 \end{pmatrix}\begin{pmatrix} 20 \\ 34 \end{pmatrix} = \begin{pmatrix} 2.5 \\ 0.5 \end{pmatrix}.$$

Thus one particular solution is $\hat{\boldsymbol{\beta}}_0 = (2.5, 0, 0.5)^\top$. The general least squares solution is

$$\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_0 + t\begin{pmatrix} -2 \\ 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 2.5 - 2t \\ t \\ 0.5 \end{pmatrix}, \quad t \in \mathbb{R}.$$

To find the minimum-norm solution, we minimize $\|\hat{\boldsymbol{\beta}}\|_2^2 = (2.5-2t)^2 + t^2 + 0.25$ over t. Taking the derivative and setting it to zero:

$$\frac{d}{dt}\left[(2.5 - 2t)^2 + t^2\right] = -4(2.5 - 2t) + 2t = 10t - 10 = 0 \implies t = 1.$$

The minimum-norm solution is therefore

$$\hat{\boldsymbol{\beta}}^{+} = \begin{pmatrix} 2.5 - 2 \\ 1 \\ 0.5 \end{pmatrix} = \begin{pmatrix} 0.5 \\ 1 \\ 0.5 \end{pmatrix}.$$

This is the solution that would be obtained via the Moore-Penrose pseudoinverse: $\hat{\boldsymbol{\beta}}^{+} = X^{+}\mathbf{y}$.

Solution 3. The Right Null Space and Its Properties:

(a) **Orthogonal Complement of the Row Space:**
We prove that $\text{Null}(X) = \text{Row}(X)^{\perp}$ by establishing inclusion in both directions.
First, suppose $\mathbf{v} \in \text{Null}(X)$, so that $X\mathbf{v} = \mathbf{0}$. Let $\mathbf{w} \in \text{Row}(X)$ be arbitrary. Since $\text{Row}(X) = \text{Col}(X^{\top})$, there exists $\boldsymbol{\alpha} \in \mathbb{R}^n$ such that $\mathbf{w} = X^{\top}\boldsymbol{\alpha}$. The inner product of $\mathbf{v}$ with $\mathbf{w}$ is

$$\langle \mathbf{v}, \mathbf{w} \rangle = \mathbf{v}^{\top}\mathbf{w} = \mathbf{v}^{\top}X^{\top}\boldsymbol{\alpha} = (X\mathbf{v})^{\top}\boldsymbol{\alpha} = \mathbf{0}^{\top}\boldsymbol{\alpha} = 0.$$

Since $\mathbf{w}$ was arbitrary, $\mathbf{v} \perp \text{Row}(X)$, and hence $\mathbf{v} \in \text{Row}(X)^{\perp}$.
Conversely, suppose $\mathbf{v} \in \text{Row}(X)^{\perp}$, so that $\mathbf{v}$ is orthogonal to every vector in $\text{Row}(X)$. In particular, $\mathbf{v}$ is orthogonal to each row of X. Let $\mathbf{r}_1^{\top}, \mathbf{r}_2^{\top}, \ldots, \mathbf{r}_n^{\top}$ denote the rows of X, so that $\mathbf{r}_i \in \text{Row}(X)$ for each i. Then

$$\langle \mathbf{v}, \mathbf{r}_i \rangle = \mathbf{r}_i^{\top}\mathbf{v} = 0 \text{ for all } i.$$

The product $X\mathbf{v}$ is a vector whose i-th component is $\mathbf{r}_i^{\top}\mathbf{v} = 0$. Therefore, $X\mathbf{v} = \mathbf{0}$, which means $\mathbf{v} \in \text{Null}(X)$.
Having established both inclusions, we conclude that $\text{Null}(X) = \text{Row}(X)^{\perp}$.

(b) **Rank-Nullity Theorem:**
By part (a), $\text{Null}(X) = \text{Row}(X)^{\perp}$. Applying the Orthogonal Decomposition Theorem to the vector space $\mathbb{R}^p$, we have

$$\mathbb{R}^p = \text{Row}(X) \oplus \text{Row}(X)^{\perp} = \text{Row}(X) \oplus \text{Null}(X).$$

For a direct sum decomposition, the dimensions satisfy

$$\dim(\mathbb{R}^p) = \dim(\text{Row}(X)) + \dim(\text{Null}(X)).$$

Since $\dim(\mathbb{R}^p) = p$ and $\dim(\text{Row}(X)) = \text{rank}(X)$ (the row rank equals the column rank), we obtain

$$p = \text{rank}(X) + \dim(\text{Null}(X)),$$

which is equivalent to $\dim(\text{Null}(X)) + \text{rank}(X) = p$.

(c) **Characterization via the Gram Matrix:**
We prove $\text{Null}(X^{\top}X) = \text{Null}(X)$ by showing inclusion in both directions.
For $\text{Null}(X) \subseteq \text{Null}(X^{\top}X)$: Let $\mathbf{v} \in \text{Null}(X)$, so $X\mathbf{v} = \mathbf{0}$. Then

$$(X^{\top}X)\mathbf{v} = X^{\top}(X\mathbf{v}) = X^{\top}\mathbf{0} = \mathbf{0},$$

hence $\mathbf{v} \in \text{Null}(X^{\top}X)$.

For $\text{Null}(X^\top X) \subseteq \text{Null}(X)$: Let $\mathbf{v} \in \text{Null}(X^\top X)$, so $(X^\top X)\mathbf{v} = \mathbf{0}$. Left-multiplying by $\mathbf{v}^\top$ gives $\mathbf{v}^\top X^\top X\mathbf{v} = 0$. Recognizing this as $\|X\mathbf{v}\|_2^2 = 0$, positive definiteness of the norm implies $X\mathbf{v} = \mathbf{0}$, hence $\mathbf{v} \in \text{Null}(X)$.
Therefore, $\text{Null}(X^\top X) = \text{Null}(X)$.
For the invertibility statement: The matrix $X^\top X \in \mathbb{R}^{p\times p}$ is invertible if and only if $\text{Null}(X^\top X) = \{\mathbf{0}\}$. By the equality just proven, this holds if and only if $\text{Null}(X) = \{\mathbf{0}\}$. By **??**, $\text{Null}(X) = \{\mathbf{0}\}$ if and only if the columns of X are linearly independent, i.e., X has full column rank. Thus, $X^\top X$ is invertible if and only if X has full column rank.

Solution 4. *Base case* ($k = 2$): This is the standard Pythagorean theorem proven in the manuscript. For orthogonal vectors $\mathbf{v}_1, \mathbf{v}_2$:

$$\|\mathbf{v}_1 + \mathbf{v}_2\|^2 = \langle \mathbf{v}_1 + \mathbf{v}_2, \mathbf{v}_1 + \mathbf{v}_2\rangle = \|\mathbf{v}_1\|^2 + 2\langle \mathbf{v}_1, \mathbf{v}_2\rangle + \|\mathbf{v}_2\|^2 = \|\mathbf{v}_1\|^2 + \|\mathbf{v}_2\|^2.$$

Inductive step: Assume the result holds for k mutually orthogonal vectors. Consider $k+1$ mutually orthogonal vectors $\mathbf{v}_1, \ldots, \mathbf{v}_{k+1}$. Let $\mathbf{w} = \sum_{i=1}^k \mathbf{v}_i$. We claim that $\mathbf{w} \perp \mathbf{v}_{k+1}$:

$$\langle \mathbf{w}, \mathbf{v}_{k+1}\rangle = \left\langle \sum_{i=1}^k \mathbf{v}_i, \mathbf{v}_{k+1} \right\rangle = \sum_{i=1}^k \langle \mathbf{v}_i, \mathbf{v}_{k+1}\rangle = 0,$$

since each $\mathbf{v}_i$ is orthogonal to $\mathbf{v}_{k+1}$ by the mutual orthogonality assumption.

Applying the base case to $\mathbf{w}$ and $\mathbf{v}_{k+1}$:

$$\left\| \sum_{i=1}^{k+1} \mathbf{v}_i \right\|^2 = \|\mathbf{w} + \mathbf{v}_{k+1}\|^2 = \|\mathbf{w}\|^2 + \|\mathbf{v}_{k+1}\|^2.$$

By the inductive hypothesis, $\|\mathbf{w}\|^2 = \sum_{i=1}^k \|\mathbf{v}_i\|^2$. Therefore:

$$\left\| \sum_{i=1}^{k+1} \mathbf{v}_i \right\|^2 = \sum_{i=1}^k \|\mathbf{v}_i\|^2 + \|\mathbf{v}_{k+1}\|^2 = \sum_{i=1}^{k+1} \|\mathbf{v}_i\|^2.$$

By induction, the result holds for all $k \geq 2$. □

Solution 5. (a) Since P is symmetric ($P^\top = P$), we have:

$$M^\top = (I_n - P)^\top = I_n^\top - P^\top = I_n - P = M.$$

(b) Using the idempotence of P ($P^2 = P$):

$$\begin{aligned} M^2 &= (I_n - P)(I_n - P) = I_n - P - P + P^2 \\ &= I_n - P - P + P = I_n - P = M. \end{aligned}$$

(c) We compute:

$$MX = (I_n - P)X = X - PX = X - X(X^\top X)^{-1}X^\top X = X - X \cdot I_p = X - X = O.$$

(d) Using the linearity and cyclic property of trace:

$$\text{tr}(M) = \text{tr}(I_n - P) = \text{tr}(I_n) - \text{tr}(P) = n - (p+1).$$

If the design matrix has p features plus an intercept column, then $\text{tr}(M) = n - (p+1)$.

(e) Part (c) shows that M annihilates every column of X, meaning M projects onto the orthogonal complement of $\text{Col}(X)$. Part (d) confirms this: the dimension of $\text{Col}(X)^\perp$ is $n - \text{rank}(X) = n - p$ (assuming full column rank), which equals $\text{tr}(M)$ since for projection matrices, trace equals rank.

Solution 6. (a) We compute:

$$X^\top X = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1.001 & 1.002 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 1 & 1.001 \\ 1 & 1.002 \end{pmatrix} = \begin{pmatrix} 3 & 3.003 \\ 3.003 & 3.006005 \end{pmatrix}.$$

(b) The characteristic polynomial is:

$$\det(X^\top X - \lambda I) = (3-\lambda)(3.006005-\lambda) - 3.003^2 = \lambda^2 - 6.006005\lambda + 0.000006.$$

Using the quadratic formula:

$$\lambda = \frac{6.006005 \pm \sqrt{36.072... - 0.000024}}{2} \approx \frac{6.006005 \pm 6.005995}{2}.$$

So $\lambda_1 \approx 6.006$ and $\lambda_2 \approx 0.000005 = 5 \times 10^{-6}$.

The condition number is:

$$\kappa(X^\top X) = \frac{\lambda_{\max}}{\lambda_{\min}} \approx \frac{6.006}{5 \times 10^{-6}} \approx 1.2 \times 10^6.$$

(c) With $\kappa(X^\top X) \approx 10^6$, small perturbations in the data (of order machine epsilon $\epsilon \approx 10^{-16}$) can cause errors in $\hat{\boldsymbol{\beta}}$ of order $\kappa \cdot \epsilon \approx 10^{-10}$. While this may seem acceptable, the effective loss of precision means only about 10 significant digits remain reliable. For problems with larger condition numbers, catastrophic cancellation can occur.

(d) If σ_1, σ_2 are the singular values of X, then $\lambda_i = \sigma_i^2$ are the eigenvalues of $X^\top X$. Thus:

$$\kappa(X^\top X) = \frac{\sigma_1^2}{\sigma_2^2} = \left(\frac{\sigma_1}{\sigma_2}\right)^2 = \kappa(X)^2.$$

The condition number squares when forming the normal equations, explaining why methods avoiding this (QR, SVD) are preferred.

Solution 7. (a) Let λ be an eigenvalue of P with eigenvector $\mathbf{v} \neq \mathbf{0}$, so $P\mathbf{v} = \lambda\mathbf{v}$. Applying P to both sides and using idempotence:

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

But $P^2 = P$, so $P^2\mathbf{v} = P\mathbf{v} = \lambda\mathbf{v}$. Therefore $\lambda\mathbf{v} = \lambda^2\mathbf{v}$, which gives $(\lambda - \lambda^2)\mathbf{v} = \mathbf{0}$. Since $\mathbf{v} \neq \mathbf{0}$, we have $\lambda(1-\lambda) = 0$, so $\lambda = 0$ or $\lambda = 1$.

(b) Since P is symmetric, it is diagonalizable with an orthonormal basis of eigenvectors. Let r be the number of eigenvalues equal to 1 (counting multiplicity), and $n - r$ be the number equal to 0. Then:

$$\text{tr}(P) = \sum_{i=1}^{n} \lambda_i = r \cdot 1 + (n-r) \cdot 0 = r.$$

The rank of P equals the dimension of its range, which equals the number of nonzero eigenvalues (for diagonalizable matrices), which is r. Therefore $\text{rank}(P) = r = \text{tr}(P)$.

(c) For the hat matrix $P = X(X^\top X)^{-1}X^\top$, we proved $\mathrm{tr}(P) = p+1$ (the number of columns of the design matrix). By part (b), this equals $\mathrm{rank}(P)$. Since P projects onto $\mathrm{Col}(X)$, and $\mathrm{rank}(P) = \dim(\mathrm{Col}(X)) = p+1$, the results are consistent.

Solution 8. (a) For simple linear regression with design matrix having columns $\mathbf{1}$ and $\mathbf{x} = (x_1, \ldots, x_n)^\top$, the normal equations give:

$$\hat{\beta}_1 = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - (\sum x_i)^2}.$$

Using $\bar{x} = \frac{1}{n}\sum x_i$ and $\bar{y} = \frac{1}{n}\sum y_i$, and the identities:

$$\sum (x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - n\bar{x}\bar{y},$$
$$\sum (x_i - \bar{x})^2 = \sum x_i^2 - n\bar{x}^2,$$

we can verify that:

$$\hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}.$$

(b) Regressing $\tilde{\mathbf{y}}$ on $\tilde{\mathbf{x}}$ without intercept, the normal equation is $(\tilde{\mathbf{x}}^\top \tilde{\mathbf{x}})\gamma = \tilde{\mathbf{x}}^\top \tilde{\mathbf{y}}$. Thus:

$$\gamma = \frac{\tilde{\mathbf{x}}^\top \tilde{\mathbf{y}}}{\tilde{\mathbf{x}}^\top \tilde{\mathbf{x}}} = \frac{\sum \tilde{x}_i \tilde{y}_i}{\sum \tilde{x}_i^2} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \hat{\beta}_1.$$

(c) From the normal equations, the first equation (corresponding to the intercept column $\mathbf{1}$) gives:

$$n\hat{\beta}_0 + \hat{\beta}_1 \sum x_i = \sum y_i.$$

Dividing by n: $\hat{\beta}_0 + \hat{\beta}_1 \bar{x} = \bar{y}$, so $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$.

Solution 9. Suppose $\hat{\mathbf{y}}_1, \hat{\mathbf{y}}_2 \in W$ both minimize $\|\mathbf{y} - \mathbf{w}\|$ over $\mathbf{w} \in W$. Let $d = \|\mathbf{y} - \hat{\mathbf{y}}_1\| = \|\mathbf{y} - \hat{\mathbf{y}}_2\|$ be the minimum distance.

Since W is a subspace, the midpoint $\mathbf{m} = \frac{1}{2}(\hat{\mathbf{y}}_1 + \hat{\mathbf{y}}_2) \in W$. By the triangle inequality, $\|\mathbf{y} - \mathbf{m}\| \geq d$ (since d is the minimum).

Apply the parallelogram law with $\mathbf{u} = \mathbf{y} - \hat{\mathbf{y}}_1$ and $\mathbf{v} = \mathbf{y} - \hat{\mathbf{y}}_2$:

$$\|\mathbf{u} + \mathbf{v}\|^2 + \|\mathbf{u} - \mathbf{v}\|^2 = 2\|\mathbf{u}\|^2 + 2\|\mathbf{v}\|^2 = 2d^2 + 2d^2 = 4d^2.$$

Note that $\mathbf{u} + \mathbf{v} = 2\mathbf{y} - \hat{\mathbf{y}}_1 - \hat{\mathbf{y}}_2 = 2(\mathbf{y} - \mathbf{m})$, so $\|\mathbf{u} + \mathbf{v}\|^2 = 4\|\mathbf{y} - \mathbf{m}\|^2 \geq 4d^2$.

Also, $\mathbf{u} - \mathbf{v} = \hat{\mathbf{y}}_2 - \hat{\mathbf{y}}_1$.

From the parallelogram law:

$$4\|\mathbf{y} - \mathbf{m}\|^2 + \|\hat{\mathbf{y}}_2 - \hat{\mathbf{y}}_1\|^2 = 4d^2.$$

Since $\|\mathbf{y} - \mathbf{m}\|^2 \geq d^2$, we have $4d^2 + \|\hat{\mathbf{y}}_2 - \hat{\mathbf{y}}_1\|^2 \leq 4d^2$, which implies $\|\hat{\mathbf{y}}_2 - \hat{\mathbf{y}}_1\|^2 \leq 0$.

Since norms are non-negative, $\|\hat{\mathbf{y}}_2 - \hat{\mathbf{y}}_1\| = 0$, hence $\hat{\mathbf{y}}_1 = \hat{\mathbf{y}}_2$. □

Solution 10. (a) The second column is twice the first: $(2, 4, 2)^\top = 2(1, 2, 1)^\top$. Thus $\mathrm{rank}(X) = 1$.

Basis for $\mathrm{Col}(X)$: $\{(1, 2, 1)^\top\}$.

For $\mathrm{Null}(X)$: we solve $X\boldsymbol{\beta} = \mathbf{0}$, i.e., $\beta_1(1, 2, 1)^\top + \beta_2(2, 4, 2)^\top = \mathbf{0}$. This gives $\beta_1 + 2\beta_2 = 0$, so $\mathrm{Null}(X) = \mathrm{span}\{(2, -1)^\top\}$.

(b) We compute $X^\top X$ and its eigendecomposition:

$$X^\top X = \begin{pmatrix} 1 & 2 & 1 \\ 2 & 4 & 2 \end{pmatrix} \begin{pmatrix} 1 & 2 \\ 2 & 4 \\ 1 & 2 \end{pmatrix} = \begin{pmatrix} 6 & 12 \\ 12 & 24 \end{pmatrix}.$$

Eigenvalues: $\det(X^\top X - \lambda I) = \lambda^2 - 30\lambda = \lambda(\lambda - 30) = 0$, so $\lambda_1 = 30$, $\lambda_2 = 0$.
Singular values: $\sigma_1 = \sqrt{30}$, $\sigma_2 = 0$.
Eigenvector for $\lambda_1 = 30$: $(1, 2)^\top/\sqrt{5}$. So $V_1 = (1, 2)^\top/\sqrt{5}$.
The left singular vector: $U_1 = XV_1/\sigma_1 = \frac{1}{\sqrt{30}}(1, 2, 1)^\top \cdot \sqrt{5} = (1, 2, 1)^\top/\sqrt{6}$.

(c) The pseudoinverse is:

$$X^+ = V_1 \sigma_1^{-1} U_1^\top = \frac{1}{\sqrt{5}} \begin{pmatrix} 1 \\ 2 \end{pmatrix} \cdot \frac{1}{\sqrt{30}} \cdot \frac{1}{\sqrt{6}} \begin{pmatrix} 1 & 2 & 1 \end{pmatrix} = \frac{1}{30} \begin{pmatrix} 1 & 2 & 1 \\ 2 & 4 & 2 \end{pmatrix}.$$

(d)

$$\hat{\boldsymbol{\beta}} = X^+ \mathbf{y} = \frac{1}{30} \begin{pmatrix} 1 & 2 & 1 \\ 2 & 4 & 2 \end{pmatrix} \begin{pmatrix} 3 \\ 7 \\ 2 \end{pmatrix} = \frac{1}{30} \begin{pmatrix} 3 + 14 + 2 \\ 6 + 28 + 4 \end{pmatrix} = \frac{1}{30} \begin{pmatrix} 19 \\ 38 \end{pmatrix} = \begin{pmatrix} 19/30 \\ 19/15 \end{pmatrix}.$$

(e) The general solution to the normal equations is $\hat{\boldsymbol{\beta}} + t(2, -1)^\top$ for $t \in \mathbb{R}$. The norm squared is:

$$\left\| \begin{pmatrix} 19/30 + 2t \\ 19/15 - t \end{pmatrix} \right\|^2 = (19/30 + 2t)^2 + (19/15 - t)^2.$$

Taking the derivative with respect to t and setting to zero:

$$2(19/30+2t)(2)+2(19/15-t)(-1) = 0 \implies 4(19/30+2t) = 19/15-t \implies t = 0.$$

Thus $\hat{\boldsymbol{\beta}} = X^+ \mathbf{y}$ is indeed the minimum-norm solution.

Solution 11. (a) By definition, $\text{Row}(X) = \text{span}\{\text{rows of } X\}$. The rows of X are the columns of $X^\top$, so $\text{Row}(X) = \text{span}\{\text{columns of } X^\top\} = \text{Col}(X^\top)$.

(b) We show $\text{Null}(X) \subseteq \text{Row}(X)^\perp$: Let $\mathbf{v} \in \text{Null}(X)$, so $X\mathbf{v} = \mathbf{0}$. For any row $\mathbf{r}_i^\top$ of X, the i-th component of $X\mathbf{v}$ is $\mathbf{r}_i^\top \mathbf{v} = 0$. Thus $\mathbf{v}$ is orthogonal to every row, hence $\mathbf{v} \in \text{Row}(X)^\perp$.

For the reverse inclusion: Let $\mathbf{v} \in \text{Row}(X)^\perp$. Then $\mathbf{v}$ is orthogonal to each row of X, meaning $\mathbf{r}_i^\top \mathbf{v} = 0$ for all i. But these are precisely the components of $X\mathbf{v}$, so $X\mathbf{v} = \mathbf{0}$ and $\mathbf{v} \in \text{Null}(X)$.

(c) From part (b) and the orthogonal decomposition theorem:

$$\mathbb{R}^p = \text{Row}(X) \oplus \text{Row}(X)^\perp = \text{Row}(X) \oplus \text{Null}(X).$$

Therefore $p = \dim(\text{Row}(X)) + \dim(\text{Null}(X))$.

(d) In the SVD partition, V_1 spans $\text{Row}(X)$ (dimension r) and V_2 spans $\text{Null}(X)$ (dimension $p - r$). The orthogonality of V's columns ensures $\text{Row}(X) \perp \text{Null}(X)$, confirming part (b).

Solution 12. (a) Observe that $(1, 2, 5)^\top + 2(0, 1, 2)^\top = (1, 4, 9)^\top$, i.e., $\mathbf{x}_1 + 2\mathbf{x}_2 = \mathbf{x}_3$ where $\mathbf{x}_j$ denotes the j-th column. Thus $(1, 2, -1)^\top$ is in the null space: $(1)\mathbf{x}_1 + (2)\mathbf{x}_2 + (-1)\mathbf{x}_3 = \mathbf{0}$.

(b) Since the columns are linearly dependent, $X^\top X$ is singular and hence not invertible. The normal equations $(X^\top X)\boldsymbol{\beta} = X^\top \mathbf{y}$ have infinitely many solutions (a solution exists since $X^\top \mathbf{y} \in \mathrm{Col}(X^\top X)$).

(c) We compute $X^\top X$ and $X^\top \mathbf{y}$:

$$X^\top X = \begin{pmatrix} 4 & 14 & 32 \\ 14 & 54 & 122 \\ 32 & 122 & 276 \end{pmatrix}, \quad X^\top \mathbf{y} = \begin{pmatrix} 11 \\ 42 \\ 95 \end{pmatrix}.$$

One can verify the third column of $X^\top X$ equals the first plus twice the second, consistent with the dependency. A particular solution is found by setting $\beta_3 = 0$ and solving the reduced system for β_0, β_1. The general solution is:

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0^* \\ \beta_1^* \\ 0 \end{pmatrix} + t \begin{pmatrix} 1 \\ 2 \\ -1 \end{pmatrix}, \quad t \in \mathbb{R},$$

where (β_0^*, β_1^*) solves the reduced normal equations.

(d) The projection $\hat{\mathbf{y}} = X\boldsymbol{\beta}$ is independent of the choice of $\boldsymbol{\beta}$ among solutions because:

$$X(\boldsymbol{\beta} + t\mathbf{n}) = X\boldsymbol{\beta} + tX\mathbf{n} = X\boldsymbol{\beta} + t\mathbf{0} = X\boldsymbol{\beta},$$

where $\mathbf{n} = (1, 2, -1)^\top \in \mathrm{Null}(X)$.

Solution 13. (a) Using $\hat{\mathbf{y}} = P\mathbf{y}$ and $\mathbf{e} = M\mathbf{y} = (I - P)\mathbf{y}$:

$$\hat{\mathbf{y}}^\top \mathbf{e} = (P\mathbf{y})^\top (I - P)\mathbf{y} = \mathbf{y}^\top P^\top (I - P)\mathbf{y} = \mathbf{y}^\top P(I - P)\mathbf{y},$$

using symmetry $P^\top = P$. Now:

$$P(I - P) = P - P^2 = P - P = O,$$

using idempotence. Therefore $\hat{\mathbf{y}}^\top \mathbf{e} = 0$.

(b) Since $\mathbf{y} = \hat{\mathbf{y}} + \mathbf{e}$ and $\hat{\mathbf{y}} \perp \mathbf{e}$ (from part a), the Pythagorean theorem gives:

$$\|\mathbf{y}\|^2 = \|\hat{\mathbf{y}} + \mathbf{e}\|^2 = \|\hat{\mathbf{y}}\|^2 + \|\mathbf{e}\|^2.$$

(c) With an intercept, $\mathbf{1} \in \mathrm{Col}(X)$, so $P\mathbf{1} = \mathbf{1}$ and thus $\bar{\hat{y}} = \bar{y}$. Writing $\mathbf{y} - \bar{y}\mathbf{1} = (\hat{\mathbf{y}} - \bar{y}\mathbf{1}) + \mathbf{e}$, and noting these are orthogonal (since $\mathbf{e} \perp \mathrm{Col}(X) \ni (\hat{\mathbf{y}} - \bar{y}\mathbf{1})$):

$$\|\mathbf{y} - \bar{y}\mathbf{1}\|^2 = \|\hat{\mathbf{y}} - \bar{y}\mathbf{1}\|^2 + \|\mathbf{e}\|^2.$$

Thus:

$$R^2 = \frac{\|\hat{\mathbf{y}} - \bar{y}\mathbf{1}\|^2}{\|\mathbf{y} - \bar{y}\mathbf{1}\|^2} = 1 - \frac{\|\mathbf{e}\|^2}{\|\mathbf{y} - \bar{y}\mathbf{1}\|^2}.$$

Since $\|\mathbf{e}\|^2 \geq 0$ and $\|\mathbf{e}\|^2 \leq \|\mathbf{y} - \bar{y}\mathbf{1}\|^2$, we have $0 \leq R^2 \leq 1$.

Solution 14. (a) The orthogonality condition states $X^\top(\mathbf{y} - X\boldsymbol{\beta}) = \mathbf{0}$. Expanding:

$$X^\top \mathbf{y} - X^\top X\boldsymbol{\beta} = \mathbf{0} \implies X^\top X\boldsymbol{\beta} = X^\top \mathbf{y}.$$

These are the normal equations.

(b) Expanding the loss function:

$$L(\boldsymbol{\beta}) = (\mathbf{y} - X\boldsymbol{\beta})^\top(\mathbf{y} - X\boldsymbol{\beta}) = \mathbf{y}^\top\mathbf{y} - 2\boldsymbol{\beta}^\top X^\top\mathbf{y} + \boldsymbol{\beta}^\top X^\top X\boldsymbol{\beta}.$$

Taking the gradient (using $\nabla_{\boldsymbol{\beta}}(\boldsymbol{\beta}^\top A\boldsymbol{\beta}) = 2A\boldsymbol{\beta}$ for symmetric A):

$$\nabla_{\boldsymbol{\beta}} L = -2X^\top\mathbf{y} + 2X^\top X\boldsymbol{\beta}.$$

Setting this to zero: $X^\top X\boldsymbol{\beta} = X^\top\mathbf{y}$, the same normal equations.

(c) The equivalence reflects that in Euclidean space (with the standard inner product), orthogonal projection minimizes squared distance. The gradient ∇L points in the direction of steepest ascent of L. Setting $\nabla L = 0$ finds stationary points. For the Euclidean norm, the gradient being zero is equivalent to the residual being orthogonal to the constraint space. This deep connection only holds for L^2 (Euclidean) norms; other norms yield different optimality conditions.

Solution 15. (a) Since Q has orthonormal columns: $Q^\top Q = I_p$.
For $X^\top X$: $X^\top X = (QR)^\top(QR) = R^\top Q^\top QR = R^\top I_p R = R^\top R$.

(b) Starting from the normal equations:

$$X^\top X\hat{\boldsymbol{\beta}} = X^\top\mathbf{y} \implies R^\top R\hat{\boldsymbol{\beta}} = R^\top Q^\top\mathbf{y}.$$

Since R is invertible (positive diagonal entries), so is $R^\top$. Left-multiplying by $(R^\top)^{-1}$:

$$R\hat{\boldsymbol{\beta}} = Q^\top\mathbf{y}.$$

(c) We verify:

$$P = X(X^\top X)^{-1}X^\top = QR(R^\top R)^{-1}R^\top Q^\top = QRR^{-1}(R^\top)^{-1}R^\top Q^\top = QI_pQ^\top = QQ^\top.$$

(d) The equation $R\hat{\boldsymbol{\beta}} = Q^\top\mathbf{y}$ is an upper triangular system solvable by back-substitution in $O(p^2)$ operations. This avoids:

- Forming $X^\top X$ (which squares the condition number)
- Computing $(X^\top X)^{-1}$ (expensive and unstable)

The QR approach has condition number $\kappa(R) = \kappa(X)$, versus $\kappa(X)^2$ for normal equations.

Solution 16. (a) From $P^2 = P$ and $P^\top = P$, we have $h_{ii} = P_{ii} = (P^2)_{ii} = \sum_{j=1}^n P_{ij}P_{ji} = \sum_{j=1}^n P_{ij}^2 \geq P_{ii}^2 = h_{ii}^2$.
Thus $h_{ii} \geq h_{ii}^2$, which gives $h_{ii}(1-h_{ii}) \geq 0$. Combined with $h_{ii} = \sum_j P_{ij}^2 \geq 0$, we get $0 \leq h_{ii} \leq 1$.

(b) $\sum_{i=1}^n h_{ii} = \operatorname{tr}(P) = p + 1$ by Theorem **??**.

(c) From $\hat{\mathbf{y}} = P\mathbf{y}$, the i-th component is:

$$\hat{y}_i = \sum_{j=1}^n P_{ij}y_j = P_{ii}y_i + \sum_{j\neq i} P_{ij}y_j = h_{ii}y_i + \sum_{j\neq i} h_{ij}y_j.$$

This shows $\hat{y}_i$ is a weighted combination of all responses, with weight h_{ii} on the i-th observation itself.

(d) For simple linear regression with centered predictor, the design matrix is $X = (\mathbf{1}, \tilde{\mathbf{x}})$ where $\tilde{x}_i = x_i - \bar{x}$. The hat matrix has diagonal elements:

$$h_{ii} = \frac{1}{n} + \frac{\tilde{x}_i^2}{\sum_j \tilde{x}_j^2} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_j (x_j - \bar{x})^2}.$$

The minimum leverage is $1/n$ (when $x_i = \bar{x}$). Observations with extreme predictor values have higher leverage, meaning they exert more influence on the fitted line. High-leverage points can drastically affect regression coefficients if they are also outliers in the response.

Solution 17. (a) The L^1 loss $\|\mathbf{y} - X\boldsymbol{\beta}\|_1 = \sum_i |y_i - (X\boldsymbol{\beta})_i|$ is not derived from an inner product. The L^1 geometry uses the "taxicab" metric, where the unit ball is a cross-polytope rather than a sphere. The minimizer of the L^1 loss produces a vector $\hat{\mathbf{y}}$ in $\mathrm{Col}(X)$, but the residual $\mathbf{y} - \hat{\mathbf{y}}$ is generally *not* orthogonal to $\mathrm{Col}(X)$ in the Euclidean sense. Instead, the optimality conditions involve subdifferentials and median-like properties.

(b) Setting the gradient to zero:

$$\nabla_{\boldsymbol{\beta}} L_W = -2X^\top W(\mathbf{y} - X\boldsymbol{\beta}) = \mathbf{0}.$$

This gives $X^\top W X\boldsymbol{\beta} = X^\top W\mathbf{y}$, so:

$$\hat{\boldsymbol{\beta}}_W = (X^\top W X)^{-1} X^\top W\mathbf{y}.$$

(c) Define the weighted inner product $\langle \mathbf{u}, \mathbf{v} \rangle_W = \mathbf{u}^\top W\mathbf{v}$. This is a valid inner product since W is symmetric positive definite. The induced norm is $\|\mathbf{u}\|_W = \sqrt{\mathbf{u}^\top W\mathbf{u}}$.

The first-order condition $X^\top W(\mathbf{y} - X\hat{\boldsymbol{\beta}}_W) = \mathbf{0}$ says that for each column $\mathbf{x}_j$ of X:

$$\langle \mathbf{x}_j, \mathbf{y} - X\hat{\boldsymbol{\beta}}_W \rangle_W = \mathbf{x}_j^\top W(\mathbf{y} - X\hat{\boldsymbol{\beta}}_W) = 0.$$

Thus the residual is orthogonal to $\mathrm{Col}(X)$ *in the W-inner product*, making $X\hat{\boldsymbol{\beta}}_W$ the orthogonal projection of $\mathbf{y}$ onto $\mathrm{Col}(X)$ with respect to $\langle \cdot, \cdot \rangle_W$.

(d) The matrix W changes the geometry of $\mathbb{R}^n$ by defining a new notion of distance. If $W = \mathrm{diag}(w_1, \ldots, w_n)$ is diagonal, then $\|\mathbf{u}\|_W^2 = \sum_i w_i u_i^2$, so observations with larger w_i contribute more to the loss. Geometrically, W stretches or compresses coordinates, transforming the Euclidean sphere into an ellipsoid. The weighted least squares solution finds the point in $\mathrm{Col}(X)$ closest to $\mathbf{y}$ in this deformed space.