Exercise 1.3

Show that the span of $\mathbf{v}_1, \ldots, \mathbf{v}_m$ is a subspace.

Exercise 1.6

Prove that the null space of $\mathbb{T}$ is a subspace.

Exercise 1.9

Let $\mathbf{z}$ be in the null space of $\begin{bmatrix}\mathbf{v}_1 & \cdots & \mathbf{v}_m\end{bmatrix}$. Given any vector of scalars $\mathbf{b}$, show that the linear combination of $\mathbf{v}_1, \ldots, \mathbf{v}_m$ produced by the entries of $\mathbf{b} + \mathbf{z}$ is exactly the same as that produced by $\mathbf{b}$.

Exercise 1.17

Let $\mathcal{F}$ be a field. Find the dimension of $\mathcal{F}^m$ as defined in Section 1.2.

Let $\mathbf{b}_1$ and $\mathbf{b}_2$ be in the null space. Given any scalars a_1, a_2, consider the vector of scalars $a_1\mathbf{b}_1 + a_2\mathbf{b}_2$.

$$\begin{aligned}\mathbb{T}(a_1\mathbf{b}_1 + a_2\mathbf{b}_2) &= a_1\underbrace{\mathbb{T}\mathbf{b}_1}_{\mathbf{0}} + a_2\underbrace{\mathbb{T}\mathbf{b}_2}_{\mathbf{0}} \\ &= \mathbf{0}\end{aligned}$$

Since $a_1\mathbf{b}_1 + a_2\mathbf{b}_2$ is also mapped to $\mathbf{0}$, it's in the null space as well; the null space therefore satisfies the definition of a subspace.

Consider two vectors in the span, say $\begin{bmatrix}\mathbf{v}_1 & \cdots & \mathbf{v}_m\end{bmatrix}\mathbf{b}_1$ and $\begin{bmatrix}\mathbf{v}_1 & \cdots & \mathbf{v}_m\end{bmatrix}\mathbf{b}_2$. For a pair of scalars a_1, a_2 the linear combination

$$a_1\begin{bmatrix}\mathbf{v}_1 & \cdots & \mathbf{v}_m\end{bmatrix}\mathbf{b}_1 + a_2\begin{bmatrix}\mathbf{v}_1 & \cdots & \mathbf{v}_m\end{bmatrix}\mathbf{b}_2 = \begin{bmatrix}\mathbf{v}_1 & \cdots & \mathbf{v}_m\end{bmatrix}(a_1\mathbf{b}_1 + a_2\mathbf{b}_2)$$

is also in the span, so the span satisfies the definition of a subspace.

Consider the n vectors $\mathbf{e}_1 := (1, 0, \ldots, 0), \ldots, \mathbf{e}_m := (0, \ldots, 0, 1)$. A given vector $(c_1, \ldots, c_m) \in \mathcal{F}^m$ has the unique representation $c_1\mathbf{e}_1 + \ldots + c_m\mathbf{e}_m$ with respect to these vectors, so they comprise a basis (known as the *standard basis*). This tells us that the dimension of $\mathcal{F}^m$ is m.

With $\mathbf{z}$ in the null space, the $\mathbf{b} + \mathbf{z}$ linear combination results in

$$\begin{aligned}\begin{bmatrix}\mathbf{v}_1 & \cdots & \mathbf{v}_m\end{bmatrix}(\mathbf{b} + \mathbf{z}) &= \begin{bmatrix}\mathbf{v}_1 & \cdots & \mathbf{v}_m\end{bmatrix}\mathbf{b} + \underbrace{\begin{bmatrix}\mathbf{v}_1 & \cdots & \mathbf{v}_m\end{bmatrix}\mathbf{z}}_{\mathbf{0}} \\ &= \begin{bmatrix}\mathbf{v}_1 & \cdots & \mathbf{v}_m\end{bmatrix}\mathbf{b}.\end{aligned}$$

Exercise 1.18

Suppose $\begin{bmatrix}\mathbf{v}_1 & \cdots & \mathbf{v}_m\end{bmatrix}$ and $\begin{bmatrix}\mathbf{w}_1 & \cdots & \mathbf{w}_m\end{bmatrix}$ have the exact same behavior on a basis $B = \{\mathbf{b}_1, \ldots, \mathbf{b}_m\}$ for the vector space of scalar coefficients, that is, $\begin{bmatrix}\mathbf{v}_1 & \cdots & \mathbf{v}_m\end{bmatrix}\mathbf{b}_j = \begin{bmatrix}\mathbf{w}_1 & \cdots & \mathbf{w}_m\end{bmatrix}\mathbf{b}_j$ for every $j \in \{1, \ldots, m\}$. Show that $\mathbf{v}_j$ must equal $\mathbf{w}_j$ for every $j \in \{1, \ldots, m\}$.

Exercise 1.19

Let λ be an eigenvalue for $\mathbb{T}$. Show that the *eigenspace* of λ is a *subspace*.

Exercise 1.20

Suppose $\mathbb{T}$ has eigenvalues $\lambda_1, \ldots, \lambda_m$ with corresponding eigenvectors $\mathbf{q}_1, \ldots, \mathbf{q}_m$. Let a be a non-zero scalar. Identify eigenvalues and eigenvectors of $a\mathbb{T}$, i.e. the function that maps any vector $\mathbf{v}$ to a times $\mathbb{T}\mathbf{v}$.

Exercise 1.21

Explain why any linear operator that has 0 as an eigenvalue doesn't have an inverse function.

Suppose that $\mathbf{q}_1$ and $\mathbf{q}_2$ are both in the eigenspace. For any scalars a_1, a_2,

$$\begin{aligned}\mathbb{T}(a_1\mathbf{q}_1 + a_2\mathbf{q}_2) &= a_1\mathbb{T}\mathbf{q}_1 + a_2\mathbb{T}\mathbf{q}_2 \\ &= a_1\lambda\mathbf{q}_1 + a_2\lambda\mathbf{q}_2 \\ &= \lambda(a_1\mathbf{q}_1 + a_2\mathbf{q}_2)\end{aligned}$$

which confirms that $a_1\mathbf{q}_1 + a_2\mathbf{q}_2$ is also an eigenvector for $\mathbb{T}$ with eigenvalue λ.

We'll first show that $\begin{bmatrix}\mathbf{v}_1 & \cdots & \mathbf{v}_m\end{bmatrix}$ and $\begin{bmatrix}\mathbf{w}_1 & \cdots & \mathbf{w}_m\end{bmatrix}$ must have the exact same behavior on every vector in $\mathcal{F}^m$ (where $\mathcal{F}$ is the scalar field) by representing an arbitrary vector $\mathbf{x}$ with respect to U. Letting $\mathbf{x} = a_1\mathbf{b}_1 + \ldots + a_m\mathbf{b}_m$,

$$\begin{aligned}\begin{bmatrix}\mathbf{v}_1 & \cdots & \mathbf{v}_m\end{bmatrix}\mathbf{x} &= \begin{bmatrix}\mathbf{v}_1 & \cdots & \mathbf{v}_m\end{bmatrix}(a_1\mathbf{b}_1 + \ldots + a_m\mathbf{b}_m) \\ &= a_1\begin{bmatrix}\mathbf{v}_1 & \cdots & \mathbf{v}_m\end{bmatrix}\mathbf{b}_1 + \ldots + a_m\begin{bmatrix}\mathbf{v}_1 & \cdots & \mathbf{v}_m\end{bmatrix}\mathbf{b}_m \\ &= a_1\begin{bmatrix}\mathbf{w}_1 & \cdots & \mathbf{w}_m\end{bmatrix}\mathbf{b}_1 + \ldots + a_m\begin{bmatrix}\mathbf{w}_1 & \cdots & \mathbf{w}_m\end{bmatrix}\mathbf{b}_m \\ &= \begin{bmatrix}\mathbf{w}_1 & \cdots & \mathbf{w}_m\end{bmatrix}(a_1\mathbf{b}_1 + \ldots + a_m\mathbf{b}_m) \\ &= \begin{bmatrix}\mathbf{w}_1 & \cdots & \mathbf{w}_m\end{bmatrix}\mathbf{x}.\end{aligned}$$

In particular, the fact that $\begin{bmatrix}\mathbf{v}_1 & \cdots & \mathbf{v}_m\end{bmatrix}$ and $\begin{bmatrix}\mathbf{w}_1 & \cdots & \mathbf{w}_m\end{bmatrix}$ map $(1, 0, \ldots, 0)$ to the same vector means that $\mathbf{v}_1$ must equal $\mathbf{w}_1$. Such an argument holds for every *column* in turn.

The corresponding eigenspace is a subspace (with dimension at least 1) that the linear operator maps to $\mathbf{0}$. Because it maps multiple elements of its domain to the same value, it can't be invertible.

Consider the action of $a\mathbb{T}$ on $\mathbf{q}_j$.

$$\begin{aligned}[a\mathbb{T}](\mathbf{q}_j) &= a(\mathbb{T}\mathbf{q}_j) \\ &= a\lambda_j\mathbf{q}_j\end{aligned}$$

So $\mathbf{q}_1, \ldots, \mathbf{q}_m$ remain eigenvectors, and their eigenvalues are $a\lambda_1, \ldots, a\lambda_m$. Furthermore, no additional eigenvectors for $a\mathbb{T}$ are introduced because clearly they would also have been eigenvectors for $\mathbb{T}$.

Exercise 1.24

Let $\mathbb{T}$ be a linear operator that has non-zero eigenvalues $\lambda_1, \ldots, \lambda_n$ with eigenvectors $\mathbf{q}_1, \ldots, \mathbf{q}_n$. Suppose $\mathbb{T}$ is invertible. Show that $\mathbb{T}^{-1}$ also has $\mathbf{q}_1, \ldots, \mathbf{q}_n$ as eigenvectors, and find the corresponding eigenvalues.

Exercise 1.27

Show that if $\mathbf{y}$ is orthogonal to every one of $\mathbf{v}_1, \ldots, \mathbf{v}_m$, then it is also orthogonal to every vector in their span.

Exercise 1.31

Justify the Pythagorean identity extended to m orthogonal vectors $\mathbf{v}_1, \ldots, \mathbf{v}_m$:

$$\|\mathbf{v}_1 + \ldots + \mathbf{v}_m\|^2 = \|\mathbf{v}_1\|^2 + \ldots + \|\mathbf{v}_m\|^2.$$

Exercise 1.32

Given a non-zero vector $\mathbf{v}$, find the norm of $\frac{1}{\|\mathbf{v}\|}\mathbf{v}$.

Let $b_1\mathbf{v}_1 + \ldots b_m\mathbf{v}_m$ represent an arbitrary vector in the span. By linearity of inner products, its inner product with $\mathbf{y}$ is

$$\begin{aligned}\langle b_1\mathbf{v}_1 + \ldots b_m\mathbf{v}_m, \mathbf{y}\rangle &= b_1 \underbrace{\langle \mathbf{v}_1, \mathbf{y}\rangle}_{0} + \ldots + b_m \underbrace{\langle \mathbf{v}_m, \mathbf{y}\rangle}_{0} \\ &= 0\end{aligned}$$

because $\mathbf{y}$ is orthogonal to each of the basis vectors.

Consider the behavior of the inverse on $\mathbf{q}_j$. We know that the inverse is supposed to undo the behavior of $\mathbb{T}$, so $\mathbb{T}^{-1}\mathbb{T}\mathbf{q}_j$ should equal $\mathbf{q}_j$.

$$\begin{aligned}\mathbb{T}^{-1}\mathbb{T}\mathbf{q}_j &= \mathbb{T}^{-1}(\lambda_j\mathbf{q}_j) \\ &= \lambda_j\mathbb{T}^{-1}\mathbf{q}_j\end{aligned}$$

For $\lambda_j\mathbb{T}^{-1}\mathbf{q}_j$ to equal $\mathbf{q}_j$, we can see that $\mathbf{q}_j$ must be an eigenvector of $\mathbb{T}^{-1}$ with eigenvalue $1/\lambda_j$. Thus $\mathbb{T}^{-1}$ has eigenvalues $1/\lambda_1, \ldots, 1/\lambda_n$ with eigenvectors $\mathbf{q}_1, \ldots, \mathbf{q}_n$.

Using Exercise 1.28 and the fact that norms are non-negative,

$$\begin{aligned}\left\| \tfrac{1}{\|\mathbf{v}\|}\mathbf{v} \right\| &= \tfrac{1}{\|\mathbf{v}\|}\|\mathbf{v}\| \\ &= 1.\end{aligned}$$

$\mathbf{v}_1$ is orthogonal to $\mathbf{v}_2 + \ldots + \mathbf{v}_m$, so by the Pythagorean identity

$$\|\mathbf{v}_1 + \ldots + \mathbf{v}_m\|^2 = \|\mathbf{v}_1\|^2 + \|\mathbf{v}_2 + \ldots + \mathbf{v}_m\|^2.$$

This logic can be applied repeatedly to bring out one vector at a time leading to the desired result. (For a more formal argument, one can invoke *induction*.)

Exercise 1.33

Given a unit vector $\mathbf{u}$, find a unique representation of the vector $\mathbf{y}$ as the sum of a vector in the span of $\mathbf{u}$ and a vector orthogonal to the span of $\mathbf{u}$.

Exercise 1.34

Given a non-zero vector $\mathbf{v}$, find a unique representation of the vector $\mathbf{y}$ as the sum of a vector in the span of $\mathbf{v}$ and a vector orthogonal to the span of $\mathbf{v}$.

Exercise 1.35

Let $\{\mathbf{u}_1, \ldots, \mathbf{u}_m\}$ be an orthonormal basis for $\mathcal{V}$. Find a unique representation of $\mathbf{y} \in \mathcal{V}$ as a linear combination of the basis vectors.

Exercise 1.36

Let $\{\mathbf{u}_1, \ldots, \mathbf{u}_m\}$ be an orthonormal basis for a real vector space $\mathcal{V}$. Show that the inner product between $\mathbf{x}$ and $\mathbf{y}$ equals the sum of the product of their squared coordinates with respect to $\mathbf{u}_1, \ldots, \mathbf{u}_m$:

$$\langle \mathbf{x}, \mathbf{y} \rangle = \sum_i (\langle \mathbf{x}, \mathbf{u}_i \rangle \langle \mathbf{y}, \mathbf{u}_i \rangle).$$

A vector is in the span of $\mathbf{v}$ if and only if it's in the span of the unit vector $\frac{\mathbf{v}}{\|\mathbf{v}\|}$. Likewise, a vector is orthogonal to the span of $\mathbf{v}$ if and only if it's orthogonal to the unit vector $\frac{\mathbf{v}}{\|\mathbf{v}\|}$. Based on our solution to Exercise 1.33 the part in the span of $\mathbf{v}$ must be

$$\left\langle \mathbf{y}, \frac{\mathbf{v}}{\|\mathbf{v}\|} \right\rangle \frac{\mathbf{v}}{\|\mathbf{v}\|} = \frac{\langle \mathbf{y}, \mathbf{v} \rangle}{\|\mathbf{v}\|^2} \mathbf{v}.$$

Thus the desired representation of $\mathbf{y}$ is

$$\mathbf{y} = \underbrace{\frac{\langle \mathbf{y}, \mathbf{v} \rangle}{\|\mathbf{v}\|^2} \mathbf{v}}_{\in \text{span}\{\mathbf{v}\}} + \underbrace{\left(\mathbf{y} - \frac{\langle \mathbf{y}, \mathbf{v} \rangle}{\|\mathbf{v}\|^2} \mathbf{v} \right)}_{\perp \text{span}\{\mathbf{v}\}}.$$

We'll explicitly construct the desired vector in the span of $\mathbf{u}$. The vector we seek must equal $\hat{b}\,\mathbf{u}$ for some scalar $\hat{b}$. Based on the trivial identity $\mathbf{y} = \hat{b}\,\mathbf{u} + (\mathbf{y} - \hat{b}\,\mathbf{u})$, we see that we need the second vector $\mathbf{y} - \hat{b}\,\mathbf{u}$ to be orthogonal to $\mathbf{u}$. Its inner product with $\mathbf{u}$ is

$$\langle \mathbf{y} - \hat{b}\,\mathbf{u}, \mathbf{u} \rangle = \langle \mathbf{y}, \mathbf{u} \rangle - \hat{b} \underbrace{\langle \mathbf{u}, \mathbf{u} \rangle}_{\|\mathbf{u}\|^2 = 1}$$

which is zero precisely when $\hat{b} = \langle \mathbf{y}, \mathbf{u} \rangle$. Therefore, $\mathbf{y}$ can be represented as the sum of $\langle \mathbf{y}, \mathbf{u} \rangle \mathbf{u}$ which is in the span of $\mathbf{u}$ and $(\mathbf{y} - \langle \mathbf{y}, \mathbf{u} \rangle \mathbf{u})$ which is orthogonal to the span of $\mathbf{u}$.

We'll use the orthonormal basis representation (Exercise 1.35) to expand $\mathbf{y}$ use linearity of inner products.

$$\begin{aligned}
\langle \mathbf{x}, \mathbf{y} \rangle &= \langle \mathbf{x}, \langle \mathbf{y}, \mathbf{u}_1 \rangle \mathbf{u}_1 + \ldots + \langle \mathbf{y}, \mathbf{u}_m \rangle \mathbf{u}_m \rangle \\
&= \langle \mathbf{x}, \langle \mathbf{y}, \mathbf{u}_1 \rangle \mathbf{u}_1 \rangle + \ldots + \langle \mathbf{x}, \langle \mathbf{y}, \mathbf{u}_m \rangle \mathbf{u}_m \rangle \\
&= \langle \mathbf{y}, \mathbf{u}_1 \rangle \langle \mathbf{x}, \mathbf{u}_1 \rangle + \ldots + \langle \mathbf{y}, \mathbf{u}_m \rangle \langle \mathbf{x}, \mathbf{u}_m \rangle.
\end{aligned}$$

The correct coefficients can be readily determined thanks to the orthogonality of the terms:

$$\mathbf{y} = \underbrace{\hat{b}_1 \mathbf{u}_1}_{\in \text{span}\{\mathbf{u}_1\}} + \underbrace{\hat{b}_2 \mathbf{u}_2 + \ldots + \hat{b}_m \mathbf{u}_m}_{\perp \text{span}\{\mathbf{u}_1\}}.$$

By comparison to Exercise 1.33, the first term has to be $\langle \mathbf{y}, \mathbf{u}_1 \rangle \mathbf{u}_1$, so its coefficient has to be $\hat{b}_1 = \langle \mathbf{y}, \mathbf{u}_1 \rangle$. By reasoning similarly for each of the basis vectors, we conclude that $\mathbf{y}$ must have the unique representation

$$\mathbf{y} = \langle \mathbf{y}, \mathbf{u}_1 \rangle \mathbf{u}_1 + \ldots + \langle \mathbf{y}, \mathbf{u}_m \rangle \mathbf{u}_m.$$

Exercise 1.37

Let $\{\mathbf{u}_1, \ldots, \mathbf{u}_m\}$ be an orthonormal basis for a real vector space $\mathcal{V}$, and let $\mathbf{y} \in \mathcal{V}$. Consider the *approximation* $\hat{\mathbf{y}} := \langle \mathbf{y}, \mathbf{u}_1 \rangle \mathbf{u}_1 + \ldots + \langle \mathbf{y}, \mathbf{u}_k \rangle \mathbf{u}_k$ with $k \leq m$. Use Parseval's identity to derive a simple formula for the squared norm of $\mathbf{y} - \hat{\mathbf{y}}$, which we might call the *squared approximation error*.

Exercise 1.38

Let $\{\mathbf{u}_1, \ldots, \mathbf{u}_m\}$ be an orthonormal basis for a real vector space $\mathcal{V}$, and let $\mathbf{y} \in \mathcal{V}$. Explain which term in the representation $\langle \mathbf{y}, \mathbf{u}_1 \rangle \mathbf{u}_1 + \ldots + \langle \mathbf{y}, \mathbf{u}_m \rangle \mathbf{u}_m$ best approximates $\mathbf{y}$ in the sense that it results in the smallest approximation error $\|\mathbf{y} - \langle \mathbf{y}, \mathbf{u}_j \rangle \mathbf{u}_j\|$.

Exercise 1.39

Given a subspace $\mathcal{S}$, show that $\mathcal{S}^{\perp}$ is also a subspace.

Exercise 1.40

Let $\hat{\mathbf{y}}$ be the orthogonal projection of $\mathbf{y}$ onto $\mathcal{S}$. Use the Pythagorean identity to show that the vector in $\mathcal{S}$ that is closest to $\mathbf{y}$ is $\hat{\mathbf{y}}$.

Based on Exercise 1.37, the squared approximation error $\|\mathbf{y} - \langle \mathbf{y}, \mathbf{u}_j \rangle \mathbf{u}_j\|^2$ is equal to the sum of the squares of the other coefficients $\sum_{i \neq j} \langle \mathbf{y}, \mathbf{u}_i \rangle^2$. Therefore, the approximation error is minimized if we use the term with the largest squared coefficient.

Representing $\mathbf{y}$ with respect to the orthonormal basis, we find that the difference between the vectors is

$$\begin{aligned}\mathbf{y} - \hat{\mathbf{y}} &= (\langle \mathbf{y}, \mathbf{u}_1 \rangle \mathbf{u}_1 + \ldots + \langle \mathbf{y}, \mathbf{u}_m \rangle \mathbf{u}_m) - (\langle \mathbf{y}, \mathbf{u}_1 \rangle \mathbf{u}_1 + \ldots + \langle \mathbf{y}, \mathbf{u}_k \rangle \mathbf{u}_k) \\ &= \langle \mathbf{y}, \mathbf{u}_{k+1} \rangle \mathbf{u}_{k+1} + \ldots + \langle \mathbf{y}, \mathbf{u}_m \rangle \mathbf{u}_m.\end{aligned}$$

Its squared norm is the sum of its squared coordinates, so

$$\|\mathbf{y} - \hat{\mathbf{y}}\|^2 = \langle \mathbf{y}, \mathbf{u}_{k+1} \rangle^2 + \ldots + \langle \mathbf{y}, \mathbf{u}_m \rangle^2.$$

Let $\mathbf{v}$ be an arbitrary vector in $\mathcal{S}$. Realizing that $\hat{\mathbf{y}} - \mathbf{v}$ is in $\mathcal{S}$ and that $\mathbf{y} - \hat{\mathbf{y}}$ is orthogonal to $\mathcal{S}$, we observe a right triangle (Figure 1.3) with sides $\mathbf{y} - \mathbf{v}$, $\hat{\mathbf{y}} - \mathbf{v}$, and $\mathbf{y} - \hat{\mathbf{y}}$. By the Pythagorean identity,

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

The first term on the right doesn't depend on the choice of $\mathbf{v}$, so the quantity is uniquely minimized by choosing $\mathbf{v}$ equal to $\hat{\mathbf{y}}$ to make the second term zero.

Let $\mathbf{v}_1, \mathbf{v}_2 \in \mathcal{S}^\perp$, and let b_1 and b_2 be scalars. We need to show that the linear combination $b_1\mathbf{v}_1 + b_2\mathbf{v}_2$ is also in $\mathcal{S}^\perp$. Letting $\mathbf{w}$ be an arbitrary vector in $\mathcal{S}$,

$$\begin{aligned}\langle b_1\mathbf{v}_1 + b_2\mathbf{v}_2, \mathbf{w} \rangle &= b_1 \underbrace{\langle \mathbf{v}_1, \mathbf{w} \rangle}_{0} + b_2 \underbrace{\langle \mathbf{v}_2, \mathbf{w} \rangle}_{0} \\ &= 0.\end{aligned}$$

Exercise 1.41

Let $\mathcal{S}_1$ and $\mathcal{S}_2$ be subspaces that are orthogonal to each other, and let $\mathcal{S}$ be the span of their union. If $\hat{\mathbf{y}}_1$ and $\hat{\mathbf{y}}_2$ are the orthogonal projections of $\mathbf{y}$ onto $\mathcal{S}_1$ and $\mathcal{S}_2$, show that the orthogonal projection of $\mathbf{y}$ onto $\mathcal{S}$ is $\hat{\mathbf{y}}_1 + \hat{\mathbf{y}}_2$.

Exercise 1.42

Let $\mathcal{S}$ be a subspace of $\mathcal{V}$, and let $\mathbf{u}_1, \ldots, \mathbf{u}_m$ comprise an orthonormal basis for $\mathcal{S}$. Given any $\mathbf{y} \in \mathcal{V}$, show that $\hat{\mathbf{y}} := \langle \mathbf{y}, \mathbf{u}_1 \rangle \mathbf{u}_1 + \ldots + \langle \mathbf{y}, \mathbf{u}_m \rangle \mathbf{u}_m$ is the orthogonal projection of $\mathbf{y}$ onto $\mathcal{S}$.

Exercise 1.43

Suppose $\hat{\mathbf{y}}_1$ and $\hat{\mathbf{y}}_2$ are the orthogonal projections of $\mathbf{y}_1$ and $\mathbf{y}_2$ onto $\mathcal{S}$. With scalars a_1 and a_2, find the orthogonal projection of $a_1\mathbf{y}_1 + a_2\mathbf{y}_2$ onto $\mathcal{S}$.

Exercise 1.44

Let $\hat{\mathbf{y}}$ be the orthogonal projection of $\mathbf{y}$ onto $\mathcal{S}$. How do we know that $\mathbf{y} - \hat{\mathbf{y}}$ is the orthogonal projection of $\mathbf{y}$ onto $\mathcal{S}^{\perp}$?

From Exercise 1.41, we understand that the orthogonal projection of onto $\mathcal{S}$ equals the sum of its orthogonal projections onto the spans of the orthonormal basis vectors. The representations of these orthogonal projections as $\langle \mathbf{y}, \mathbf{u}_1 \rangle \mathbf{u}_1, \ldots, \langle \mathbf{y}, \mathbf{u}_m \rangle \mathbf{u}_m$ comes from Exercise 1.33.

For an arbitrary $\mathbf{v} \in \mathcal{S}$, we need to establish that

$$\mathbf{y} - (\hat{\mathbf{y}}_1 + \hat{\mathbf{y}}_2) \perp \mathbf{v}.$$

Every vector in the span of $\mathcal{S}_1 \cup \mathcal{S}_2$ can be represented as the sum of a vector in $\mathcal{S}_1$ and a vector in $\mathcal{S}_2$. Making use of this fact, we let $\mathbf{v} = \mathbf{v}_1 + \mathbf{v}_2$ with $\mathbf{v}_1 \in \mathcal{S}_1$ and $\mathbf{v}_2 \in \mathcal{S}_2$.

$$\begin{aligned}
\langle \mathbf{v}, \mathbf{y} - (\hat{\mathbf{y}}_1 + \hat{\mathbf{y}}_2) \rangle &= \langle \mathbf{v}_1 + \mathbf{v}_2, \mathbf{y} - \hat{\mathbf{y}}_1 - \hat{\mathbf{y}}_2 \rangle \\
&= \langle \mathbf{v}_1, \mathbf{y} - \hat{\mathbf{y}}_1 - \hat{\mathbf{y}}_2 \rangle + \langle \mathbf{v}_2, \mathbf{y} - \hat{\mathbf{y}}_1 - \hat{\mathbf{y}}_2 \rangle \\
&= \underbrace{\langle \mathbf{v}_1, \mathbf{y} - \hat{\mathbf{y}}_1 \rangle}_{0} + \underbrace{\langle \mathbf{v}_2, \mathbf{y} - \hat{\mathbf{y}}_2 \rangle}_{0} \\
&= 0
\end{aligned}$$

We know that $\mathbf{y} = \hat{\mathbf{y}} + (\mathbf{y} - \hat{\mathbf{y}})$ with $\hat{\mathbf{y}} \in \mathcal{S}$ and $\mathbf{y} - \hat{\mathbf{y}} \in \mathcal{S}^\perp$ by definition of orthogonal projection. Of course, by definition of orthogonal complement, $\hat{\mathbf{y}} \perp \mathcal{S}^\perp$, so that same representation shows that $\mathbf{y} - \hat{\mathbf{y}}$ is the orthogonal projection of $\mathbf{y}$ onto $\mathcal{S}^\perp$.

We can write out each vector in terms of its orthogonal projections onto $\mathcal{S}$ and $\mathcal{S}^\perp$, then regroup the terms.

$$\begin{aligned}
a_1\mathbf{y}_1 + a_2\mathbf{y}_2 &= a_1[\hat{\mathbf{y}}_1 + (\mathbf{y}_1 - \hat{\mathbf{y}}_1)] + a_2[\hat{\mathbf{y}}_2 + (\mathbf{y}_2 - \hat{\mathbf{y}}_2)] \\
&= \underbrace{(a_1\hat{\mathbf{y}}_1 + a_2\hat{\mathbf{y}}_2)}_{\in \mathcal{S}} + \underbrace{[a_1(\mathbf{y}_1 - \hat{\mathbf{y}}_1) + a_2(\mathbf{y}_2 - \hat{\mathbf{y}}_2)]}_{\perp \mathcal{S}}
\end{aligned}$$

This shows that $a_1\hat{\mathbf{y}}_1 + a_2\hat{\mathbf{y}}_2$ is the orthogonal projection of $a_1\mathbf{y}_1 + a_2\mathbf{y}_2$ onto $\mathcal{S}$. In other words, the orthogonal projection of a linear combination is the linear combination of the orthogonal projections.

Exercise 1.46

Let $\mathbb{H}$ be an orthogonal projection operator onto $\mathcal{S}$. Show that every vector in $\mathcal{S}$ is an eigenvector of $\mathbb{H}$.

Exercise 1.47

Let $\mathbb{H}$ be the orthogonal projection operator onto $\mathcal{S}$. Show that every vector in $\mathcal{S}^{\perp}$ is an eigenvector of $\mathbb{H}$.

Exercise 1.48

Show that every orthogonal projection operator is idempotent.

Exercise 1.54

Show that $\mathbb{M}^T\mathbb{M}$ is symmetric.

If $\mathbf{v} \perp \mathcal{S}$, then clearly $\mathbf{v} = \mathbf{0} + \mathbf{v}$ is the unique representation of $\mathbf{v}$ as the sum of a vector in $\mathcal{S}$ and a vector orthogonal to $\mathcal{S}$. Therefore $\mathbb{H}\mathbf{v} = \mathbf{0}$, which means that $\mathbf{v}$ is an eigenvector with eigenvalue 0.

If $\mathbf{v}$ is in $\mathcal{S}$, then clearly $\mathbf{v} = \mathbf{v} + \mathbf{0}$ is the unique representation of $\mathbf{v}$ as the sum of a vector in $\mathcal{S}$ and a vector orthogonal to $\mathcal{S}$. Therefore $\mathbb{H}\mathbf{v} = \mathbf{v}$, which means that $\mathbf{v}$ is an eigenvector with eigenvalue 1.

The transpose of a product of matrices is equal to the product of their transposes multiplied in the reverse order (Exercise 1.52). Thus

$$\begin{aligned}(\mathbb{M}^T\mathbb{M})^T &= (\mathbb{M})^T(\mathbb{M}^T)^T \\ &= \mathbb{M}^T\mathbb{M}.\end{aligned}$$

Let $\mathbb{H}$ be the orthogonal projection operator onto $\mathcal{S}$, and let $\hat{\mathbf{y}}$ be the orthogonal projection of $\mathbf{y}$ onto $\mathcal{S}$. Because $\hat{\mathbf{y}}$ is in $\mathcal{S}$, $\mathbb{H}$ maps it to itself.

$$\begin{aligned}[\mathbb{H} \circ \mathbb{H}]\mathbf{y} &= \mathbb{H}(\mathbb{H}\mathbf{y}) \\ &= \mathbb{H}\hat{\mathbf{y}} \\ &= \hat{\mathbf{y}}\end{aligned}$$

The action of $\mathbb{H} \circ \mathbb{H}$ is exactly the same as that of $\mathbb{H}$ on every vector, so they're the same operator.

Exercise 1.57

Let $\mathbf{q}_1, \ldots, \mathbf{q}_n$ be an orthonormal basis for $\mathbb{R}^n$. Show that $\mathbb{M}$ has the *spectral decomposition*

$$\mathbb{M} = \lambda_1 \mathbf{q}_1 \mathbf{q}_1^T + \ldots + \lambda_n \mathbf{q}_n \mathbf{q}_n^T$$

if and only if $\mathbf{q}_1, \ldots, \mathbf{q}_n$ are eigenvectors for $\mathbb{M}$ with eigenvalues $\lambda_1, \ldots, \lambda_n$.

Exercise 1.58

Let $\mathbb{M} \in \mathbb{R}^{n \times n}$ be a symmetric matrix with non-negative eigenvalues $\lambda_1, \ldots, \lambda_n$ and corresponding orthonormal eigenvectors $\mathbf{q}_1, \ldots, \mathbf{q}_n$. Show that the symmetric matrix that has eigenvalues $\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_n}$ with eigenvectors $\mathbf{q}_1, \ldots, \mathbf{q}_n$ is the *square root* of $\mathbb{M}$ (denoted $\mathbb{M}^{1/2}$) in the sense that $\mathbb{M}^{1/2}\mathbb{M}^{1/2} = \mathbb{M}$.

Exercise 1.59

Let $\mathbb{M}$ be a symmetric and invertible real matrix. Show that $\mathbb{M}^{-1}$ is also a symmetric real matrix.

Exercise 1.60

Let $\mathbb{M}$ be a symmetric real matrix. Show that the trace of $\mathbb{M}$ equals the sum of its eigenvalues.

Using a spectral decomposition, we multiply the proposed square root matrix by itself:

$$\begin{aligned}\mathbb{M}^{1/2}\mathbb{M}^{1/2} &= \mathbb{M}^{1/2}(\sqrt{\lambda_1}\mathbf{q}_1\mathbf{q}_1^T + \ldots + \sqrt{\lambda_n}\mathbf{q}_n\mathbf{q}_n^T) \\ &= \sqrt{\lambda_1}\underbrace{\mathbb{M}^{1/2}\mathbf{q}_1}_{\sqrt{\lambda_1}\mathbf{q}_1}\mathbf{q}_1^T + \ldots + \sqrt{\lambda_n}\underbrace{\mathbb{M}^{1/2}\mathbf{q}_n}_{\sqrt{\lambda_n}\mathbf{q}_n}\mathbf{q}_n^T \\ &= \lambda_1\mathbf{q}_1\mathbf{q}_1^T + \ldots + \lambda_n\mathbf{q}_n\mathbf{q}_n^T \\ &= \mathbb{M}.\end{aligned}$$

Let's figure out the behavior of $\lambda_1\mathbf{q}_1\mathbf{q}_1^T + \ldots + \lambda_n\mathbf{q}_n\mathbf{q}_n^T$ on the basis vectors.

$$\begin{aligned}(\lambda_1\mathbf{q}_1\mathbf{q}_1^T + \ldots + \lambda_n\mathbf{q}_n\mathbf{q}_n^T)\mathbf{q}_1 &= \lambda_1\mathbf{q}_1\underbrace{\mathbf{q}_1^T\mathbf{q}_1}_{\|\mathbf{q}_1\|^2=1} + \ldots + \lambda_n\mathbf{q}_n\underbrace{\mathbf{q}_n^T\mathbf{q}_1}_{0} \\ &= \lambda_1\mathbf{q}_1\end{aligned}$$

meaning $\mathbf{q}_1$ is also an eigenvector of this matrix with eigenvalue λ_1. Likewise for $\mathbf{q}_2, \ldots, \mathbf{q}_n$. By establishing that $\mathbb{M}$ and $\lambda_1\mathbf{q}_1\mathbf{q}_1^T + \ldots + \lambda_n\mathbf{q}_n\mathbf{q}_n^T$ behave the same on a basis, we see that they must be the same matrix by Exercise 1.18.

We'll use the matrix form of spectral decomposition $\mathbb{M} = \mathbb{Q}\mathbb{\Lambda}\mathbb{Q}^T$ and the *cyclic permutation* property of trace (Exercise 1.51).

$$\begin{aligned}\operatorname{tr}\mathbb{M} &= \operatorname{tr}(\mathbb{Q}\mathbb{\Lambda}\mathbb{Q}^T) \\ &= \operatorname{tr}(\underbrace{\mathbb{Q}^T\mathbb{Q}}_{\mathbb{I}_n}\mathbb{\Lambda}) \\ &= \operatorname{tr}\mathbb{\Lambda}\end{aligned}$$

Let $\lambda_1, \ldots, \lambda_n$ and $\mathbf{q}_1, \ldots, \mathbf{q}_n$ be eigenvalues and orthonormal eigenvectors of $\mathbb{M}$. Based on Exercise 1.24, we can deduce that $\mathbb{M}^{-1}$ has the spectral decomposition

$$\mathbb{M}^{-1} = \tfrac{1}{\lambda_1}\mathbf{q}_1\mathbf{q}_1^T + \ldots + \tfrac{1}{\lambda_n}\mathbf{q}_n\mathbf{q}_n^T.$$

Because $\mathbb{M}^{-1}$ is a linear combination of symmetric real matrices (see Exercise 1.54), it's clearly a symmetric real matrix as well.

Exercise 1.61

Let $\mathbb{M}$ be an $n \times m$ real matrix. How do you know that the number of terms in a singular value decomposition of $\mathbb{M}$ can't be more than $\min(n, m)$.

Exercise 1.62

Use a singular value decomposition for $\mathbb{M} \in \mathbb{R}^{n \times m}$ to find a spectral decomposition of $\mathbb{M}^T\mathbb{M}$.

Exercise 1.66

Provide a formula for the matrix that maps $\mathbf{y}$ to its orthogonal projection onto the span of the unit vector $\mathbf{u} \in \mathbb{R}^n$.

Exercise 1.67

Show that the trace of an orthogonal projection matrix equals the dimension of the subspace that it projects onto.

Writing $\mathbb{M} = \mathbb{U}\mathbb{S}\mathbb{V}^T$,

$$\begin{aligned}\mathbb{M}^T\mathbb{M} &= (\mathbb{U}\mathbb{S}\mathbb{V}^T)^T(\mathbb{U}\mathbb{S}\mathbb{V}^T) \\ &= \mathbb{V}\mathbb{S}^T \underbrace{\mathbb{U}^T\mathbb{U}}_{\mathbb{I}}\mathbb{S}\mathbb{V}^T \\ &= \mathbb{V}\mathbb{S}^2\mathbb{V}^T.\end{aligned}$$

By comparison to the matrix form of spectral decomposition, we see that $\mathbb{M}^T\mathbb{M}$ has eigenvalues equal to the squares of the singular values of $\mathbb{M}$, and the corresponding eigenvectors are the columns of $\mathbb{V}$.

The vectors $\mathbf{u}_1, \ldots, \mathbf{u}_d \in \mathbb{R}^n$ are linearly independent, so there can't be more than n of them. Likewise, the vectors $\mathbf{v}_1, \ldots, \mathbf{v}_d \in \mathbb{R}^m$ are linearly independent, so there can't be more than m of them.

From Exercise 1.60, we know that the trace of $\mathbb{H}$ equals the sum of its eigenvalues $\lambda_1, \ldots, \lambda_n$. Furthermore, because it's is an orthogonal projection matrix, we know that it yields the spectral decomposition

$$\mathbb{H} = (1)\mathbf{q}_1\mathbf{q}_1^T + \ldots + (1)\mathbf{q}_m\mathbf{q}_m^T + (0)\mathbf{q}_{m+1}\mathbf{q}_{m+1}^T + \ldots + (0)\mathbf{q}_n\mathbf{q}_n^T$$

where $\mathbf{q}_1, \ldots, \mathbf{q}_m$ are in the subspace that $\mathbb{H}$ projects onto and the rest are necessarily orthogonal to it. We see m terms with the eigenvalue 1 and the remaining terms with the eigenvalue 0, so their sum is m which is the dimension of the subspace that $\mathbb{H}$ projects onto.

The orthogonal projection of $\mathbf{y}$ onto the span of $\mathbf{u}$ is $\langle \mathbf{y}, \mathbf{u} \rangle \mathbf{u}$. By rewriting this as $\mathbf{u}\mathbf{u}^T\mathbf{y}$, we realize that the matrix $\mathbf{u}\mathbf{u}^T$ maps any vector to its orthogonal projection onto the span of $\mathbf{u}$.

Exercise 1.68

Let $\mathbb{M}$ be a matrix. Explain why the rank of the orthogonal projection matrix onto $C(\mathbb{M})$ must be exactly the same as the rank of $\mathbb{M}$.

Exercise 1.69

Let $\mathbb{M} \in \mathbb{R}^{n \times m}$ and $\mathbf{y} \in \mathbb{R}^n$. Explain why the *Normal equation*

$$\mathbb{M}^T \mathbb{M} \hat{\mathbf{b}} = \mathbb{M}^T \mathbf{y}$$

is satisfied by the coefficient vector $\hat{\mathbf{b}} \in \mathbb{R}^m$ if and only if $\mathbb{M}\hat{\mathbf{b}}$ is the orthogonal projection of $\mathbf{y}$ onto $C(\mathbb{M})$.

Exercise 1.70

Suppose $\mathbb{M} \in \mathbb{R}^{n \times m}$ has linearly independent columns. Provide a formula for the coefficient vector $\hat{\mathbf{b}}$ for which $\mathbb{M}\hat{\mathbf{b}}$ equals the orthogonal projection of $\mathbf{y} \in \mathbb{R}^n$ onto $C(\mathbb{M})$.

Exercise 1.71

Suppose $\mathbb{M} \in \mathbb{R}^{n \times m}$ has linearly independent columns. Provide a formula for the orthogonal projection matrix onto $C(\mathbb{M})$.

The orthogonal projection $\mathbb{M}\hat{\mathbf{b}}$ is the unique vector in $C(\mathbb{M})$ with the property that $\mathbf{y} - \mathbb{M}\hat{\mathbf{b}} \perp C(\mathbb{M})$. It is equivalent to check that $\mathbf{y} - \mathbb{M}\hat{\mathbf{b}}$ is orthogonal to every column $\mathbf{v}_1, \ldots, \mathbf{v}_m$ of $\mathbb{M}$. Equivalently the following quantity should be equal to the zero vector:

$$\begin{aligned}\mathbb{M}^T(\mathbf{y} - \mathbb{M}\hat{\mathbf{b}}) &= \begin{bmatrix} - & \mathbf{v}_1 & - \\ & \vdots & \\ - & \mathbf{v}_m & - \end{bmatrix} (\mathbf{y} - \mathbb{M}\hat{\mathbf{b}}) \\ &= \begin{bmatrix} \mathbf{v}_1^T(\mathbf{y} - \mathbb{M}\hat{\mathbf{b}}) \\ \vdots \\ \mathbf{v}_m^T(\mathbf{y} - \mathbb{M}\hat{\mathbf{b}}) \end{bmatrix}.\end{aligned}$$

Setting this vector $\mathbb{M}^T(\mathbf{y} - \mathbb{M}\hat{\mathbf{b}})$ equal to the zero vector results in the Normal equation.

The equality of ranks follows from the stronger observation that the orthogonal projection matrix must have the exact same column space as $\mathbb{M}$. Every vector in $C(\mathbb{M})$ gets mapped to itself by the orthogonal projection matrix, so its column space is at least as large as $C(\mathbb{M})$. However, the orthogonal projection of any vector onto $C(\mathbb{M})$ must by definition be in $C(\mathbb{M})$, so the orthogonal projection matrix cannot map any vector to a result outside of $C(\mathbb{M})$.

We've already derived in Exercise 1.70 a formula for the desired coefficient vector $\hat{\mathbf{b}} = (\mathbb{M}^T\mathbb{M})^{-1}\mathbb{M}^T\mathbf{y}$, so we simply plug this into $\mathbb{M}\hat{\mathbf{b}}$ to find the orthogonal projection of $\mathbf{y}$ onto $C(\mathbb{M})$.

$$\begin{aligned}\hat{\mathbf{y}} &= \mathbb{M}\hat{\mathbf{b}} \\ &= \mathbb{M}(\mathbb{M}^T\mathbb{M})^{-1}\mathbb{M}^T\mathbf{y}\end{aligned}$$

Therefore, we see that $\mathbf{y}$ is mapped to its orthogonal projection onto $C(\mathbb{M})$ by the matrix $\mathbb{M}(\mathbb{M}^T\mathbb{M})^{-1}\mathbb{M}^T$.

Because the columns are linearly independent, we know that $\mathbb{M}^T\mathbb{M}$ is invertible and thus the Normal equation

$$\mathbb{M}^T\mathbb{M}\hat{\mathbf{b}} = \mathbb{M}^T\mathbf{y}$$

is uniquely solved by $\hat{\mathbf{b}} = (\mathbb{M}^T\mathbb{M})^{-1}\mathbb{M}^T\mathbf{y}$.

Exercise 1.76

For a unit vector $\mathbf{u}$, express the quadratic form $\mathbf{u}^T\mathbb{M}\mathbf{u}$ as a weighted average of the eigenvalues of $\mathbb{M} \in \mathbb{R}^{n\times n}$.

Exercise 1.77

Identify a unit vector $\mathbf{u}$ that maximizes the quadratic form $\mathbf{u}^T\mathbb{M}\mathbf{u}$.

Exercise 1.78

Given any real matrix $\mathbb{M}$, show that $\mathbb{M}^T\mathbb{M}$ is positive semi-definite.

Exercise 1.79

Let $\mathbb{M}$ be a symmetric real matrix. Show that $\mathbb{M}$ is positive semi-definite if and only if its eigenvalues are all non-negative.

From Exercise 1.76, we know that the quadratic form equals a weighted average of the eigenvalues. This weighted average is maximized by placing all of the weight on the largest eigenvalue, that is, by letting $\mathbf{u}$ be a principal eigenvector. Such a choice of $\mathbf{u}$ makes $\mathbf{u}^T\mathbb{M}\mathbf{u}$ equal to the largest eigenvalue.

Let $\mathbf{q}_1, \ldots, \mathbf{q}_n$ be an orthonormal basis of eigenvectors for $\mathbb{M}$ with eigenvalues $\lambda_1, \ldots, \lambda_n$. We can represent $\mathbf{u}$ with respect to the eigenvector basis as $\langle \mathbf{u}, \mathbf{q}_1\rangle \mathbf{q}_1 + \ldots + \langle \mathbf{u}, \mathbf{q}_n\rangle \mathbf{q}_n$.

$$\begin{aligned}\mathbf{u}^T\mathbb{M}\mathbf{u} &= \mathbf{u}^T\mathbb{M}(\langle \mathbf{u}, \mathbf{q}_1\rangle \mathbf{q}_1 + \ldots + \langle \mathbf{u}, \mathbf{q}_n\rangle \mathbf{q}_n) \\ &= \mathbf{u}^T(\langle \mathbf{u}, \mathbf{q}_1\rangle \underbrace{\mathbb{M}\mathbf{q}_1}_{\lambda_1\mathbf{q}_1} + \ldots + \langle \mathbf{u}, \mathbf{q}_n\rangle \underbrace{\mathbb{M}\mathbf{q}_n}_{\lambda_n\mathbf{q}_n}) \\ &= \langle \mathbf{u}, \mathbf{q}_1\rangle^2\lambda_1 + \ldots + \langle \mathbf{u}, \mathbf{q}_n\rangle^2\lambda_n\end{aligned}$$

$\langle \mathbf{u}, \mathbf{q}_1\rangle, \ldots, \langle \mathbf{u}, \mathbf{q}_n\rangle$ provide the coordinates of $\mathbf{u}$ with respect to the basis $\mathbf{q}_1, \ldots, \mathbf{q}_n$. Because $\mathbf{u}$ is a unit vector, the sum of these squared coordinates has to be 1. Additionally, the squared coordinates are non-negative. Consequently, we've expressed $\mathbf{u}^T\mathbb{M}\mathbf{u}$ as a weighted average of the eigenvalues; the weights are the squared coordinates of $\mathbf{u}$ with respect to the eigenvector basis.

From our work in Exercise 1.77, we've seen how to express the quadratic form as a linear combination of the eigenvalues

$$\mathbf{v}^T\mathbb{M}\mathbf{v} = \langle \mathbf{v}, \mathbf{q}_1\rangle^2\lambda_1 + \ldots + \langle \mathbf{v}, \mathbf{q}_n\rangle^2\lambda_n.$$

If every eigenvalue is at least zero, then every term in this sum is non-negative so the quadratic form must be non-negative. Conversely, if λ_j is negative, then the quadratic form arising from $\mathbf{v} = \mathbf{q}_j$ is negative, as it equals λ_j.

Exercise 1.54 established that the matrix in question is symmetric. The quadratic form

$$\begin{aligned}\mathbf{v}^T(\mathbb{M}^T\mathbb{M})\mathbf{v} &= (\mathbf{v}^T\mathbb{M}^T)(\mathbb{M}\mathbf{v}) \\ &= (\mathbb{M}\mathbf{v})^T(\mathbb{M}\mathbf{v})\end{aligned}$$

equals the squared norm of the vector $\mathbb{M}\mathbf{v}$ which is non-negative.

Exercise 1.80

Let $\mathbb{H} \in \mathbb{R}^{n \times n}$ be an orthogonal projection matrix, and let $\mathbf{v} \in \mathbb{R}^n$. Show that the squared length of $\mathbb{H}\mathbf{v}$ equals the quadratic form $\mathbf{v}^T \mathbb{H} \mathbf{v}$.

Exercise 1.81

Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be the rows of a real matrix $\mathbb{X}$. Show that the quadratic form $\mathbf{u}^T(\frac{1}{n}\mathbb{X}^T\mathbb{X})\mathbf{u}$ is equal to the average of the squares of the coefficients of $\mathbf{x}_1, \ldots, \mathbf{x}_n$ with respect to $\mathbf{u}$.

Exercise 1.82

Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be the rows of the matrix $\mathbb{X}$. Show that $\frac{1}{n}\mathbb{X}^T\mathbb{X}$ is the matrix whose (j,k)-entry is the average of the product of the jth and kth coordinates of the vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n$.

Exercise 1.83

Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be the rows of a real matrix $\mathbb{X}$. Show that the average squared length $\frac{1}{n}\sum_i \|\mathbf{x}_i\|^2$ equals the sum of the eigenvalues of $\frac{1}{n}\mathbb{X}^T\mathbb{X}$.

We'll first express the quadratic form in terms of the squared norm of a vector.

$$\begin{aligned}\mathbf{u}^T(\tfrac{1}{n}\mathbb{X}^T\mathbb{X})\mathbf{u} &= \tfrac{1}{n}(\mathbb{X}\mathbf{u})^T(\mathbb{X}\mathbf{u}) \\ &= \tfrac{1}{n}\|\mathbb{X}\mathbf{u}\|^2\end{aligned}$$

The entries of the vector $\mathbb{X}\mathbf{u}$ are the coefficients of $\mathbf{x}_1, \ldots, \mathbf{x}_n$ with respect to $\mathbf{u}$. Its squared norm is the sum of its squared entries, so $\frac{1}{n}\|\mathbb{X}\mathbf{u}\|^2$ is the average of the squared coefficients.

Because $\mathbb{H}$ is symmetric and idempotent,

$$\begin{aligned}\|\mathbb{H}\mathbf{v}\|^2 &= (\mathbb{H}\mathbf{v})^T(\mathbb{H}\mathbf{v}) \\ &= \mathbf{v}^T\mathbb{H}^T\mathbb{H}\mathbf{v} \\ &= \mathbf{v}^T\mathbb{H}\mathbf{v}.\end{aligned}$$

By Parseval's identity, the squared norm equals the sum of the squared coordinates using any basis; let's consider the orthonormal eigenvectors $\mathbf{q}_1, \ldots, \mathbf{q}_m$ of $\frac{1}{n}\mathbb{X}^T\mathbb{X}$, with $\lambda_1, \ldots, \lambda_m$ denoting their eigenvalues. Recall that Exercise 1.81 allows us to rewrite the average of squared coefficients as a quadratic form.

$$\begin{aligned}\tfrac{1}{n}\sum_i \|\mathbf{x}_i\|^2 &= \tfrac{1}{n}\sum_i(\langle\mathbf{x}_i, \mathbf{q}_1\rangle^2 + \ldots + \langle\mathbf{x}_i, \mathbf{q}_m\rangle^2) \\ &= \tfrac{1}{n}\sum_i\langle\mathbf{x}_i, \mathbf{q}_1\rangle^2 + \ldots + \tfrac{1}{n}\sum_i\langle\mathbf{x}_i, \mathbf{q}_m\rangle^2 \\ &= \underbrace{\mathbf{q}_1^T(\tfrac{1}{n}\mathbb{X}^T\mathbb{X})\mathbf{q}_1}_{\lambda_1} + \ldots + \underbrace{\mathbf{q}_m^T(\tfrac{1}{n}\mathbb{X}^T\mathbb{X})\mathbf{q}_m}_{\lambda_m}\end{aligned}$$

Exercise 1.76 demonstrated that a quadratic form evaluated at a unit eigenvector equals the corresponding eigenvalue.

The product of the matrices

$$\mathbb{X}^T\mathbb{X} = \begin{bmatrix} | & & | \\ \mathbf{x}_1 & \cdots & \mathbf{x}_n \\ | & & | \end{bmatrix}\begin{bmatrix} - & \mathbf{x}_1 & - \\ & \vdots & \\ - & \mathbf{x}_n & - \end{bmatrix}$$

has as its (j,k)-entry the inner product of the jth row of $\mathbb{X}^T$ and the kth column of $\mathbb{X}$. With $x_{i,j}$ denoting the jth coordinate of $\mathbf{x}_i$, this inner product equals $\sum_i x_{i,j}x_{i,k}$. When multiplied by $1/n$, this entry is indeed the average of the products of the coordinates. By thinking about summing over the observations, $\frac{1}{n}\mathbb{X}^T\mathbb{X}$ can also be understood as an average of rank-1 matrices $\frac{1}{n}\mathbf{x}_i\mathbf{x}_i^T$.

Exercise 2.1

Show that the entries of $\mathbf{v} = (v_1, \ldots, v_n)$ have mean zero if and only if $\mathbf{v}$ is orthogonal to $\mathbf{1} = (1, \ldots, 1)$.

Exercise 2.2

Use the Pythagorean identity to decompose the average of the squared differences between the response values and $a \in \mathbb{R}$, that is $\frac{1}{n}\sum_i (y_i - a)^2$, into two terms, one of which is the empirical variance of $y_1, \ldots, y_n$.

Exercise 2.3

Is it possible for the *least-squares line*'s sum of squared residuals to be greater than the *least-squares point*'s sum of squared residuals?

Exercise 2.4

The variables picture provides us with a more specific answer to the question posed in Exercise 2.3. Use the Pythagorean identity to quantify the difference between the least-squares point's sum of squared residuals and the least-squares line's sum of squared residuals.

We can write $\sum_i (y_1 - a)^2$ as the squared norm $\|\mathbf{y} - a\mathbf{1}\|^2$. The vector $\mathbf{y} - a\mathbf{1}$ is the hypotenuse of the right triangle whose other two sides are $\mathbf{y} - \bar{y}\mathbf{1}$ and $\bar{y}\mathbf{1} - a\mathbf{1}$. By the Pythagorean identity,

$$\begin{aligned}
\tfrac{1}{n}\sum_i (y_i - a)^2 &= \tfrac{1}{n}\|\mathbf{y} - a\mathbf{1}\|^2 \\
&= \tfrac{1}{n}[\|\mathbf{y} - \bar{y}\mathbf{1}\|^2 + \|\bar{y}\mathbf{1} - a\mathbf{1}\|^2] \\
&= \tfrac{1}{n}\left[\sum_i (y_i - \bar{y})^2 + n(\bar{y} - a)^2\right] \\
&= \tfrac{1}{n}\sum_i (y_i - \bar{y})^2 + (\bar{y} - a)^2.
\end{aligned}$$

The average of the entries is proportional to the inner product of $\mathbf{v}$ with $\mathbf{1}$.

$$\tfrac{1}{n}\sum_i v_i = \tfrac{1}{n}\langle \mathbf{v}, \mathbf{1}\rangle$$

So the average is zero if and only if the inner product is zero.

Because $\bar{y}\mathbf{1}$ is in the span of $\mathbf{1}$ and $\mathbf{x}$, we see that the least-squares line's residual vector $\mathbf{y} - \hat{\mathbf{y}}$ must be orthogonal to $\hat{\mathbf{y}} - \bar{y}\mathbf{1}$. Invoking the Pythagorean identity,

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

The least-squares point's sum of squared residuals is larger than the least-squares line's sum of squared residuals by $\|\hat{\mathbf{y}} - \bar{y}\mathbf{1}\|^2$.

The set of possible prediction functions corresponding to lines $\{f(x) = a + bx : a, b \in \mathbb{R}\}$ is strictly larger than the set of possible prediction functions corresponding to points $\{f(x) = a : a \in \mathbb{R}\}$. A line predicts every response value by the same number if its slope is zero. By definition, the least-squares line will use a slope of zero if and only if that leads to the smallest possible sum of squared residuals, in which case its sum of squared residuals would be equal to that of the least-squares point.

Exercise 2.6

Let $\mathbf{y} \in \mathbb{R}^n$ be a response variable and $\mathbf{x} \in \mathbb{R}^n$ be an explanatory variable. Consider fitting the response variable using quadratic functions of the explanatory variable:

$$\{f_{a,b,c}(x) = a + bx + cx^2 : a, b, c \in \mathbb{R}\}.$$

Show that the set of possible prediction vectors is a subspace of $\mathbb{R}^n$.

Exercise 2.7

Let $\mathbf{y} \in \mathbb{R}^n$ be a response variable vector and $\mathbf{x} \in \mathbb{R}^n$ be an explanatory variable vector. Consider predicting the response variable by using quadratic functions of the explanatory variable:

$$\{f_{a,b,c}(x) = a + bx + cx^2 : a, b, c \in \mathbb{R}\}.$$

Explain how to find the coefficients $(\hat{a}, \hat{b}, \hat{c})$ of the quadratic function that minimizes the sum of squared residuals.

Exercise 2.8

Let $\hat{\mathbf{y}}$ be the orthogonal projection of $\mathbf{y}$ onto $C(\mathbb{M})$. Explain why $(\mathbb{M}^T\mathbb{M})^-\mathbb{M}^T\hat{\mathbf{y}}$ must be equal to $(\mathbb{M}^T\mathbb{M})^-\mathbb{M}^T\mathbf{y}$.

Exercise 2.9

Suppose $\hat{\mathbf{y}}$ is the orthogonal projection of $\mathbf{y}$ onto $\mathcal{S}$, $\check{\mathbf{y}}$ is the orthogonal projection of $\mathbf{y}$ onto $\mathcal{S}_0 \subseteq \mathcal{S}$, and that $\mathbf{1} \in \mathcal{S}_0$. Explain why

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

With $\mathbf{x}^2$ representing the vector of squared explanatory values, we can use the design matrix

$$\mathbb{M} := \begin{bmatrix} | & | & | \\ \mathbf{1} & \mathbf{x} & \mathbf{x}^2 \\ | & | & | \end{bmatrix}.$$

According to Theorem 2.4, the least-squares coefficients are $(\hat{a}, \hat{b}, \hat{c}) = (\mathbb{M}^T\mathbb{M})^-\mathbb{M}^T\mathbf{y}$.

Let $f_{a,b,c}(\mathbf{x})$ denote the vector of predictions $(f_{a,b,c}(x_1), \ldots, f_{a,b,c}(x_n))$. With $\mathbf{x}^2$ representing the vector of squared explanatory values, an arbitrary linear combination of two arbitrary vectors of predicted values is

$$\begin{aligned} \alpha_1 f_{a_1,b_1,c_1}(\mathbf{x}) + \alpha_2 f_{a_2,b_2,c_2}(\mathbf{x}) &= \alpha_1(a_1\mathbf{1} + b_1\mathbf{x} + c_1\mathbf{x}^2) + \alpha_2(a_2\mathbf{1} + b_2\mathbf{x} + c_2\mathbf{x}^2) \\ &= (\alpha_1 a_1 + \alpha_2 a_2)\mathbf{1} + (\alpha_1 b_1 + \alpha_2 b_2)\mathbf{x} + (\alpha_1 c_1 + \alpha_2 c_2)\mathbf{x}^2 \\ &= f_{\alpha_1 a_1 + \alpha_2 a_2, \alpha_1 b_1 + \alpha_2 b_2, \alpha_1 c_1 + \alpha_2 c_2}(\mathbf{x}) \end{aligned}$$

which is another possible vector of predicted values that can be achieved using a quadratic function. In fact, we can see that the set of possible predictions is exactly the span of $\mathbf{1}$, $\mathbf{x}$, and $\mathbf{x}^2$.

The vector $\check{\mathbf{y}}$ is defined to be the orthogonal projection of $\mathbf{y}$ onto $\mathcal{S}_0$. However, it's also the orthogonal projection of $\hat{\mathbf{y}}$ onto $\mathcal{S}_0$ because according to Exercise 1.50, orthogonal projection onto $\mathcal{S}$ followed by orthogonal projection onto $\mathcal{S}_0$ lands you at the exact same vector that a single orthogonal projection onto $\mathcal{S}_0$ does. Likewise, $\bar{y}\mathbf{1}$ is the orthogonal projection of $\check{\mathbf{y}}$ onto $\mathbf{1}$. Invoke the ANOVA decomposition with $\hat{\mathbf{y}}$ playing the role of the response variable.

There's an intuitive explanation for this. You can think of $(\mathbb{M}^T\mathbb{M})^-\mathbb{M}^T$ as the matrix that maps any vector in $\mathbb{R}^n$ to the (minimum norm) coefficients of the columns of $\mathbb{M}$ that lead to the orthogonal projection of that vector onto $C(\mathbb{M})$. Because the orthogonal projection of $\hat{\mathbf{y}}$ onto $C(\mathbb{M})$ is exactly the same as the orthogonal projection of $\mathbf{y}$ onto $C(\mathbb{M})$ (namely, both are $\hat{\mathbf{y}}$), the coefficients leading to this orthogonal projection must be the same.

Exercise 3.6

Let $\mathbf{X}$ be a random vector and $\mathbf{v}$ be a non-random vector. Explain why $\mathbb{E}(\mathbf{X} + \mathbf{v}) = \mathbb{E}\mathbf{X} + \mathbf{v}$.

Exercise 3.7

Suppose $\mathbb{E}\mathbf{X} = \mathbf{0}$. Show that the coordinate of $\mathbf{X}$ with respect to $\mathbf{u}$ has expectation 0.

Exercise 3.8

Let $\mathbf{X}$ be a random vector that maps to a real vector space with an inner product. Show that the expected squared length of $\mathbf{X}$ equals sum of the expected squares of its coordinates with respect to any orthonormal basis $\mathbf{u}_1, \ldots, \mathbf{u}_m$.

Exercise 3.9

Let $\mathbf{X} = (X_1, \ldots, X_n)$ be a random vector, $\mathbf{v} = (v_1, \ldots, v_n)$ be a non-random vector, and $\mathbb{M}$ be an $n \times m$ matrix. Show that

$$\mathbb{E}(\mathbf{v} + \mathbb{M}\mathbf{X}) = \mathbf{v} + \mathbb{M}\mathbb{E}\mathbf{X}.$$

$$\mathbb{E}\langle \mathbf{X}, \mathbf{u}\rangle = \langle \underbrace{\mathbb{E}\mathbf{X}}_{\mathbf{0}}, \mathbf{u}\rangle$$
$$= 0$$

The random vector $\mathbf{X} + \mathbf{v}$ maps any ω to $\mathbf{X}(\omega) + \mathbf{v}$; we're justified in treating $\mathbf{v}$ as if it's the random vector that maps every element of the sample space to the vector $\mathbf{v}$. By property (iii), $\mathbb{E}(\mathbf{X} + \mathbf{v}) = \mathbb{E}\mathbf{X} + \mathbb{E}\mathbf{v}$, and by Exercise 3.5, $\mathbb{E}\mathbf{v} = \mathbf{v}$.

From Exercise 3.6, $\mathbb{E}(\mathbf{v} + \mathbb{M}\mathbf{X}) = \mathbf{v} + \mathbb{E}\mathbb{M}\mathbf{X}$. Let $\mathbf{m}_1, \ldots, \mathbf{m}_n$ be the rows of $\mathbb{M}$. Putting the expectation into each coordinate of the vector,

$$\mathbb{E}\mathbb{M}\mathbf{X} = \mathbb{E}\begin{bmatrix} \mathbf{m}_1^T\mathbf{X} \\ \vdots \\ \mathbf{m}_n^T\mathbf{X} \end{bmatrix}$$
$$= \begin{bmatrix} \mathbb{E}\mathbf{m}_1^T\mathbf{X} \\ \vdots \\ \mathbb{E}\mathbf{m}_n^T\mathbf{X} \end{bmatrix}$$
$$= \begin{bmatrix} \mathbf{m}_1^T\mathbb{E}\mathbf{X} \\ \vdots \\ \mathbf{m}_n^T\mathbb{E}\mathbf{X} \end{bmatrix}$$
$$= \mathbb{M}\mathbb{E}\mathbf{X}.$$

This is a simple consequence of Parseval's identity.

$$\mathbb{E}\|\mathbf{X}\|^2 = \mathbb{E}[\langle \mathbf{X}, \mathbf{u}_1\rangle^2 + \ldots + \langle \mathbf{X}, \mathbf{u}_m\rangle^2]$$
$$= \mathbb{E}\langle \mathbf{X}, \mathbf{u}_1\rangle^2 + \ldots + \mathbb{E}\langle \mathbf{X}, \mathbf{u}_m\rangle^2$$

Exercise 3.10

Suppose $\mathbf{X}$ is a discrete random vector with probability mass function p on $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$. Show that $\mathbb{E}\mathbf{X} = \sum_i \mathbf{x}_i p(\mathbf{x}_i)$.

Exercise 3.11

Let X be a discrete random variable whose possible values are the positive integers. In particular, suppose that $\mathbb{P}\{X = k\}$ is proportional to $1/k^2$ for $k \in \{1, 2, \ldots\}$. What's the expectation of X?

Exercise 3.15

Let $\mathbf{Y}$ be a random vector with expectation $\boldsymbol{\mu}$. Find the non-random vector $\mathbf{v}$ that minimizes $\mathbb{E}\|\mathbf{Y} - \mathbf{v}\|^2$.

Exercise 3.16

Explain how Exercise 2.2 is an instance of the bias-variance decomposition.

Recall that $\sum_{k=1}^{\infty} \frac{1}{k^2} = \pi^2/6$, so this distribution is well-defined. However, its expectation is

$$\begin{aligned}
\mathbb{E}X &= \sum_{k=1}^{\infty} k\mathbb{P}\{X = k\} \\
&= \sum_{k=1}^{\infty} k \tfrac{6}{\pi^2} \tfrac{1}{k^2} \\
&= \tfrac{6}{\pi^2} \sum_{k=1}^{\infty} \tfrac{1}{k} \\
&= \infty.
\end{aligned}$$

The random vector can be represented by the sum

$$\mathbf{X}(\omega) = \mathbf{x}_1 \mathbb{1}_{\mathbf{X}(\omega)=\mathbf{x}_1} + \ldots + \mathbf{x}_n \mathbb{1}_{\mathbf{X}(\omega)=\mathbf{x}_n}$$

Taking the expectation,

$$\begin{aligned}
\mathbb{E}\mathbf{X} &= \mathbb{E}[\mathbf{x}_1 \mathbb{1}_{\mathbf{X}=\mathbf{x}_1} + \ldots + \mathbf{x}_n \mathbb{1}_{\mathbf{X}=\mathbf{x}_n}] \\
&= \mathbf{x}_1 \underbrace{\mathbb{E}\mathbb{1}_{\mathbf{X}=\mathbf{x}_1}}_{p(\mathbf{x}_1)} + \ldots + \mathbf{x}_n \underbrace{\mathbb{E}\mathbb{1}_{\mathbf{X}=\mathbf{x}_n}}_{p(\mathbf{x}_n)}
\end{aligned}$$

by property (i) of the definition of expectation.

If the distribution of the random variable Y is the empirical distribution defined by $\mathbf{y} = (y_1, \ldots, y_n)$, then its expectation is $\bar{y}$. By the bias-variance decomposition,

$$\begin{aligned}
\mathbb{E}(Y - a)^2 &= (a - \mathbb{E}Y)^2 + \mathbb{E}(Y - \mathbb{E}Y)^2 \\
&\Updownarrow \\
\tfrac{1}{n}\sum_i (y_i - a)^2 &= (a - \bar{y})^2 + \tfrac{1}{n}\sum_i (y_i - \bar{y})^2.
\end{aligned}$$

By the bias-variance decomposition, the objective function equals $\|\mathbf{v} - \boldsymbol{\mu}\|^2 + \mathbb{E}\|\mathbf{Y} - \boldsymbol{\mu}\|^2$. The second term doesn't depend on $\mathbf{v}$, so we can minimize the sum by taking $\mathbf{v}$ to be $\boldsymbol{\mu}$ which makes the first term zero.

Exercise 3.17

Let $\mathbf{Y}$ be a random vector that is an *unbiased estimator* for $\boldsymbol{\theta}$, that is $\mathbb{E}\mathbf{Y} = \boldsymbol{\theta}$. If $\lambda \in \mathbb{R}$, express $\|\mathbb{E}(\lambda\mathbf{Y}) - \boldsymbol{\theta}\|^2$ (which can be thought of as the *squared bias* of the estimator $\lambda\mathbf{Y}$) in terms of λ and $\|\boldsymbol{\theta}\|^2$.

Exercise 3.18

Let $\mathbf{Y}$ be a random vector, and let $\lambda \in \mathbb{R}$. Express $\mathbb{E}\|\lambda\mathbf{Y} - \mathbb{E}(\lambda\mathbf{Y})\|^2$ in terms of λ and $\mathbb{E}\|\mathbf{Y} - \mathbb{E}\mathbf{Y}\|^2$.

Exercise 3.19

Let $\mathbf{Y}$ be a random vector that is an *unbiased estimator* for $\boldsymbol{\theta} \in \mathbb{R}^n$. Use the bias-variance decomposition along with your results from Exercises 3.17 and 3.18 to find an expression for $\lambda \in \mathbb{R}$ (in terms of $\|\boldsymbol{\theta}\|^2$ and $\mathbb{E}\|\mathbf{Y} - \boldsymbol{\theta}\|^2$) for which $\mathbb{E}\|\boldsymbol{\theta} - \lambda\mathbf{Y}\|^2$ is as small as possible.

Exercise 3.22

Let $\mathbf{Y}$ be an $\mathbb{R}^n$-valued random vector, and let $\mathbf{v} \in \mathbb{R}^n$. Use Exercise 3.21 to show that the covariance of $\mathbf{v} + \mathbf{Y}$ has the same covariance matrix as $\mathbf{Y}$.

Factoring out λ,

$$\begin{aligned}\mathbb{E}\|\lambda\mathbf{Y} - \mathbb{E}(\lambda\mathbf{Y})\|^2 &= \mathbb{E}\|\lambda(\mathbf{Y} - \mathbb{E}\mathbf{Y})\|^2 \\ &= \lambda^2\mathbb{E}\|\mathbf{Y} - \mathbb{E}\mathbf{Y}\|^2.\end{aligned}$$

$$\begin{aligned}\|\mathbb{E}(\lambda\mathbf{Y}) - \boldsymbol{\theta}\|^2 &= \|\lambda\underbrace{\mathbb{E}\mathbf{Y}}_{\boldsymbol{\theta}} - \boldsymbol{\theta}\|^2 \\ &= \|(\lambda - 1)\boldsymbol{\theta}\|^2 \\ &= (1 - \lambda)^2\|\boldsymbol{\theta}\|^2\end{aligned}$$

Note that the factor $(\lambda - 1)^2$ is equal to $(1 - \lambda)^2$ which is a bit more intuitive when $\lambda \in [0, 1]$.

$$\begin{aligned}\operatorname{cov}(\mathbf{v} + \mathbf{Y}) &= \mathbb{E}[(\mathbf{v} + \mathbf{Y} - \mathbb{E}(\mathbf{v} + \mathbf{Y}))(\mathbf{v} + \mathbf{Y} - \mathbb{E}(\mathbf{v} + \mathbf{Y}))^T] \\ &= \mathbb{E}[(\mathbf{Y} - \mathbb{E}\mathbf{Y})(\mathbf{Y} - \mathbb{E}\mathbf{Y})^T] \\ &= \operatorname{cov}\mathbf{Y}\end{aligned}$$

By the bias-variance decomposition and our previous results,

$$\begin{aligned}\mathbb{E}\|\boldsymbol{\theta} - \lambda\mathbf{Y}\|^2 &= \|\mathbb{E}(\lambda\mathbf{Y}) - \boldsymbol{\theta}\|^2 + \mathbb{E}\|\lambda\mathbf{Y} - \mathbb{E}(\lambda\mathbf{Y})\|^2 \\ &= (1 - \lambda)^2\|\boldsymbol{\theta}\|^2 + \lambda^2\mathbb{E}\|\mathbf{Y} - \mathbb{E}\mathbf{Y}\|^2.\end{aligned}$$

Taking the derivative with respect to λ, and setting it to zero, we get the critical λ^*:

$$(1 - \lambda^*)\|\boldsymbol{\theta}\|^2 = \lambda^*\mathbb{E}\|\mathbf{Y} - \mathbb{E}\mathbf{Y}\|^2$$

is solved by $\lambda^* = \frac{\|\boldsymbol{\theta}\|^2}{\|\boldsymbol{\theta}\|^2 + \mathbb{E}\|\mathbf{Y} - \mathbb{E}\mathbf{Y}\|^2}$. Realize of course that when estimating an unknown parameter $\boldsymbol{\theta}$, we can't actually calculate this optimal value.

Exercise 3.23

Let $\mathbf{Y}$ be a random vector with covariance matrix $\mathbb{C}$. Let $\mathbf{v}$ be a non-random vector, and let $\mathbb{M}$ be a real matrix. Show that the covariance of $\mathbf{v} + \mathbb{M}\mathbf{Y}$ is $\mathbb{M}\mathbb{C}\mathbb{M}^T$.

Exercise 3.24

Show that every covariance matrix is positive semi-definite.

Exercise 3.25

Show that $\mathbb{E}\|\mathbf{X} - \mathbb{E}\mathbf{X}\|^2 = \operatorname{tr}(\operatorname{cov}\mathbf{X})$.

Exercise 3.27

Let $\boldsymbol{\epsilon}$ be a random vector with expectation $\mathbf{0}$ and covariance matrix $\sigma^2\mathbb{I}$. Let $\mathbf{v}$ be a non-random vector, and let $\mathbb{H}$ be an orthogonal projection matrix. Find the covariance matrix of $\mathbb{H}(\mathbf{v} + \boldsymbol{\epsilon})$.

To satisfy the definition, we need to show that every quadratic form is non-negative. We'll use the covariance expression from Exercise 3.21 and consider its quadratic form for an arbitrary vector $\mathbf{v}$,

$$\begin{aligned}\mathbf{v}^T\mathbb{E}[(\mathbf{Y}-\mathbb{E}\mathbf{Y})(\mathbf{Y}-\mathbb{E}\mathbf{Y})^T]\mathbf{v} &= \mathbb{E}[\mathbf{v}^T(\mathbf{Y}-\mathbb{E}\mathbf{Y})(\mathbf{Y}-\mathbb{E}\mathbf{Y})^T]\mathbf{v}\\ &= \mathbb{E}[\mathbf{v}^T(\mathbf{Y}-\mathbb{E}\mathbf{Y})(\mathbf{Y}-\mathbb{E}\mathbf{Y})^T\mathbf{v}]\\ &= \mathbb{E}\langle\mathbf{Y}-\mathbb{E}\mathbf{Y},\mathbf{v}\rangle^2.\end{aligned}$$

The expectation of a non-negative random variable has to be non-negative.

By Exercise 3.22, $\operatorname{cov}(\mathbf{v}+\mathbb{M}\mathbf{Y}) = \operatorname{cov}\mathbb{M}\mathbf{Y}$.

$$\begin{aligned}\operatorname{cov}\mathbb{M}\mathbf{Y} &= \mathbb{E}[(\mathbb{M}\mathbf{Y}-\mathbb{E}\mathbb{M}\mathbf{Y})(\mathbb{M}\mathbf{Y}-\mathbb{E}\mathbb{M}\mathbf{Y})^T]\\ &= \mathbb{E}[(\mathbb{M}\mathbf{Y}-\mathbb{M}\mathbb{E}\mathbf{Y})(\mathbb{M}\mathbf{Y}-\mathbb{M}\mathbb{E}\mathbf{Y})^T]\\ &= \mathbb{E}[(\mathbb{M}(\mathbf{Y}-\mathbb{E}\mathbf{Y}))(\mathbb{M}(\mathbf{Y}-\mathbb{E}\mathbf{Y}))^T]\\ &= \mathbb{M}(\mathbb{E}[(\mathbf{Y}-\mathbb{E}\mathbf{Y})(\mathbf{Y}-\mathbb{E}\mathbf{Y})^T])\mathbb{M}^T\\ &= \mathbb{M}\mathbb{C}\mathbb{M}^T\end{aligned}$$

Distribute the matrix multiplication to get $\mathbb{H}\mathbf{v}+\mathbb{H}\boldsymbol{\epsilon}$. According to Exercise 3.23, the covariance is

$$\begin{aligned}\mathbb{H}(\sigma^2\mathbb{I})\mathbb{H}^T &= \sigma^2\mathbb{H}\mathbb{H}^T\\ &= \sigma^2\mathbb{H}\end{aligned}$$

by symmetry and idempotence of orthogonal projection matrices.

$$\begin{aligned}\mathbb{E}\|\mathbf{X}-\mathbb{E}\mathbf{X}\|^2 &= \mathbb{E}[(X_1-\mathbb{E}X_1)^2+\ldots+(X_n-\mathbb{E}X_n)^2]\\ &= \mathbb{E}(X_1-\mathbb{E}X_1)^2+\ldots+\mathbb{E}(X_n-\mathbb{E}X_n)^2\end{aligned}$$

These variances are the diagonals of the covariance matrix, so its trace is their sum.

Exercise 3.28

Let $\mathbf{X}$ have expectation $\boldsymbol{\mu}_{\mathbf{X}}$ and $\mathbf{Y}$ have expectation $\boldsymbol{\mu}_{\mathbf{Y}}$. Show that the expected inner product between the centered vectors $\mathbf{X} - \boldsymbol{\mu}_{\mathbf{X}}$ and $\mathbf{Y} - \boldsymbol{\mu}_{\mathbf{Y}}$ is the same as the expected inner product when only one of them is centered.

Exercise 3.29

Use Exercise 3.28 to observe that

$$\langle \mathbf{x} - \bar{x}\mathbf{1}, \mathbf{y} \rangle = \langle \mathbf{x} - \bar{x}\mathbf{1}, \mathbf{y} - \bar{y}\mathbf{1} \rangle.$$

Exercise 3.30

Let $\mathbf{X}$ be a random vector mapping to a real vector space, and let $\mathbf{v}$ be a non-random vector. Show that the variance of the coordinate of $\mathbf{X}$ with respect to $\mathbf{u}$ is the same as the variance of the coordinate of $\mathbf{X} + \mathbf{v}$ with respect to $\mathbf{u}$.

Exercise 3.31

If $\mathbf{X}$ has expectation $\boldsymbol{\mu}$, find the expectation of the *centered* version $\mathbf{X} - \boldsymbol{\mu}$.

Let the *joint* distribution of (X, Y) be the empirical distribution of $(x_1, y_1), \ldots, (x_n, y_n)$.

$$\begin{aligned}
\langle \mathbf{x} - \bar{x}\mathbf{1}, \mathbf{y} \rangle &= n\tfrac{1}{n} \sum_i [(x_i - \bar{x}) y_i] \\
&= n\mathbb{E}[(X - \mathbb{E}X)Y] \\
&= n\mathbb{E}[(X - \mathbb{E}X)(Y - \mathbb{E}Y)] \\
&= n\tfrac{1}{n} \sum_i [(x_i - \bar{x})(y_i - \bar{y})] \\
&= \langle \mathbf{x} - \bar{x}\mathbf{1}, \mathbf{y} - \bar{y}\mathbf{1} \rangle
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}\langle \mathbf{X} - \boldsymbol{\mu}_\mathbf{X}, \mathbf{Y} - \boldsymbol{\mu}_\mathbf{Y} \rangle &= \mathbb{E}\langle \mathbf{X} - \boldsymbol{\mu}_\mathbf{X}, \mathbf{Y} \rangle - \mathbb{E}\langle \mathbf{X} - \boldsymbol{\mu}_\mathbf{X}, \boldsymbol{\mu}_\mathbf{Y} \rangle \\
&= \mathbb{E}\langle \mathbf{X} - \boldsymbol{\mu}_\mathbf{X}, \mathbf{Y} \rangle - \langle \underbrace{\mathbb{E}\mathbf{X} - \boldsymbol{\mu}_\mathbf{X}}_{\mathbf{0}}, \boldsymbol{\mu}_\mathbf{Y} \rangle \\
&= \mathbb{E}\langle \mathbf{X} - \boldsymbol{\mu}_\mathbf{X}, \mathbf{Y} \rangle
\end{aligned}$$

The same argument works for $\mathbf{Y} - \boldsymbol{\mu}_\mathbf{Y}$ if you keep Exercise 3.14 in mind.

$$\begin{aligned}
\mathbb{E}(\mathbf{X} - \boldsymbol{\mu}) &= \underbrace{\mathbb{E}\mathbf{X}}_{\boldsymbol{\mu}} - \boldsymbol{\mu} \\
&= \mathbf{0}
\end{aligned}$$

The difference between $\langle \mathbf{X} + \mathbf{v}, \mathbf{u} \rangle$ and $\langle \mathbf{X}, \mathbf{u} \rangle$ is $\langle \mathbf{v}, \mathbf{u} \rangle$ which is non-random. By Exercise 3.23, we can conclude that they must therefore have the same variance.

Exercise 3.32

Let $\mathbb{M}$ be a positive definite matrix. Based on Exercises 1.24 and 1.58, explain why the inverse of the square root of $\mathbb{M}$ is the same as the square root of the inverse of $\mathbb{M}$.

Exercise 3.33

Let $\mathbf{Y}$ have expectation $\boldsymbol{\mu}$ and covariance matrix $\mathbb{C}$. Find the expectation and covariance of $\mathbb{C}^{-1/2}(\mathbf{Y} - \boldsymbol{\mu})$.

Exercise 3.34

If $\mathbf{Y}$ has expectation $\boldsymbol{\mu}$ and a positive definite covariance matrix $\mathbb{C}$, find the expected squared Mahalanobis distance from $\mathbf{Y}$ to its own distribution.

Exercise 3.35

Let $\mathbb{H}$ be the orthogonal projection matrix onto a d-dimensional subspace $\mathcal{S} \subseteq \mathbb{R}^n$, and let $\mathbf{Y}$ be a random vector with covariance matrix $\sigma^2\mathbb{I}$. Show that

$$\mathbb{E}\|\mathbb{H}\mathbf{Y}\|^2 = d\sigma^2 + \|\mathbb{H}\boldsymbol{\mu}\|^2$$

The random vector $\mathbf{Y} - \boldsymbol{\mu}$ has expectation zero, so based on Exercise 3.9, $\mathbb{C}^{-1/2}(\mathbf{Y} - \boldsymbol{\mu})$ has expectation $\mathbb{C}^{-1/2}\mathbf{0} = \mathbf{0}$. For the covariance, we apply the formula from Exercise 3.23 to get

$$\begin{aligned}\operatorname{cov}[\mathbb{C}^{-1/2}(\mathbf{Y} - \boldsymbol{\mu})] &= \mathbb{C}^{-1/2}\mathbb{C}(\mathbb{C}^{-1/2})^T \\ &= \underbrace{\mathbb{C}^{-1/2}\mathbb{C}^{1/2}}_{\mathbb{I}}\underbrace{\mathbb{C}^{1/2}\mathbb{C}^{-1/2}}_{\mathbb{I}} \\ &= \mathbb{I}.\end{aligned}$$

To find the square root of a positive semi-definite matrix, you replace the eigenvalues by their square roots. To find the inverse of an invertible symmetric matrix, you replace the eigenvalues by their reciprocals. No matter which order you do these two operations in, you end up with the same matrix:

$$\frac{1}{\sqrt{\lambda_1}}\mathbf{q}_1\mathbf{q}_1^T + \ldots + \frac{1}{\sqrt{\lambda_n}}\mathbf{q}_n\mathbf{q}_n^T$$

where $\mathbf{q}_1, \ldots, \mathbf{q}_n$ are eigenvectors of $\mathbb{M}$ with eigenvalues $\lambda_1, \ldots, \lambda_n$.

By comparison to Equation 3.1, all that remains is to verify that the trace of $\mathbb{H}\sigma^2\mathbb{I}$ is $d\sigma^2$.

$$\begin{aligned}\operatorname{tr}[\mathbb{H}\sigma^2\mathbb{I}] &= \sigma^2 \operatorname{tr}\mathbb{H} \\ &= d\sigma^2\end{aligned}$$

because according to Exercise 1.67 the trace of an orthogonal projection matrix equals the dimension of the subspace that it projects onto.

Let $\mathbf{Z} := \mathbb{C}^{-1/2}(\mathbf{Y} - \boldsymbol{\mu})$ represent the standardized version of $\mathbf{Y}$, and let $(Z_1, \ldots, Z_n)$ represent its coordinates. Notice that the squared Mahalanobis distance from $\mathbf{Y}$ to its distribution is exactly the squared norm of the standardized version.

$$\begin{aligned}\mathbb{E}\|\mathbb{C}^{-1/2}[\mathbf{Y} - \boldsymbol{\mu}]\|^2 &= \mathbb{E}\|\mathbf{Z}\|^2 \\ &= \mathbb{E}Z_1^2 + \ldots + \mathbb{E}Z_n^2 \\ &= \underbrace{\operatorname{var} Z_1}_{1} + \ldots + \underbrace{\operatorname{var} Z_n}_{1} \\ &= n\end{aligned}$$

The expected squared Mahalanobis distance is the dimension of the vector space that $\mathbf{Y}$ inhabits.

Exercise 4.1

Suppose that $Y_1, \ldots, Y_n$ satisfy a location model

$$Y_i = \alpha + \epsilon_i.$$

Show that the least-squares point (Theorem 2.1) is an unbiased estimator for α.

Exercise 4.2

Suppose that $Y_1, \ldots, Y_n$ are uncorrelated and all have the same variance σ^2. What's the variance of the least-squares point?

Exercise 4.5

Suppose that a response variable satisfies a simple linear model of an explanatory variable and that it is predicted by the least-squares line. Which is larger: the sum of squared *errors* or the sum of squared *residuals*? Base your answer on the definition of the least-squares line, and explain.

Exercise 4.6

The variables picture provides us with a more specific answer to the question posed in Exercise 4.5. Use the Pythagorean identity to quantify the difference between the sum of squared errors and the sum of squared residuals.

The variance of a constant times a random variable equals the square of that constant times the variance of the random variable (Exercise 3.23). Furthermore, the variance of a sum of uncorrelated random variables equals the sum of the variances (Exercise 3.26).

$$\begin{aligned}\operatorname{var}\overline{Y} &= \operatorname{var}\left(\tfrac{1}{n}\sum_i Y_i\right)\\ &= \tfrac{1}{n^2}\sum_i \underbrace{\operatorname{var} Y_i}_{\sigma^2}\\ &= \tfrac{1}{n^2}(n\sigma^2)\\ &= \frac{\sigma^2}{n}.\end{aligned}$$

Remember that the *least-squares point* is simply the average of the response values. The expectation is

$$\begin{aligned}\mathbb{E}\overline{Y} &= \mathbb{E}(\tfrac{1}{n}\sum_i Y_i)\\ &= \tfrac{1}{n}\sum_i \underbrace{\mathbb{E}Y_i}_{\alpha}\\ &= \alpha.\end{aligned}$$

The error vector forms the hypotenuse of a right triangle whose other sides are $\widehat{\mathbf{Y}} - \mathbb{E}\mathbf{Y}$ and the residual vector $\mathbf{Y} - \widehat{\mathbf{Y}}$. Invoking the Pythagorean identity,

$$\|\boldsymbol{\epsilon}\|^2 = \|\mathbf{Y} - \widehat{\mathbf{Y}}\|^2 + \|\widehat{\mathbf{Y}} - \mathbb{E}\mathbf{Y}\|^2.$$

The sum of squared errors is larger than the sum of squared residuals by $\|\widehat{\mathbf{Y}} - \mathbb{E}\mathbf{Y}\|^2$.

The sum of squared errors is the sum of squared differences between the response values and the true line, while the sum of squared residuals is the sum of squared differences between the points and the least-squares line. The least-squares line is, by definition, the one with the smallest possible sum of squared differences from the points, so the sum of squared residuals can't possibly be larger than the sum of squared errors.

Exercise 4.7

Let $(x_1^{(1)}, \ldots, x_1^{(m)}), \ldots, (x_n^{(1)}, \ldots, x_n^{(m)}) \in \mathbb{R}^m$ be n observations of m explanatory variables, and suppose that the response variable $Y_1, \ldots, Y_n$ satisfies a multiple linear model

$$Y_i = \alpha + \beta_1(x^{(1)} - \bar{x}^{(1)}) + \ldots + \beta_m(x^{(m)} - \bar{x}^{(m)}) + \epsilon_i.$$

Assuming the explanatory variables' empirical covariance matrix Σ is full rank, show that the coefficients $\hat{\alpha}, \hat{\beta}_1, \ldots, \hat{\beta}_m$ in the least-squares hyperplane $y = \hat{\alpha} + \hat{\beta}_1(x^{(1)} - \bar{x}^{(1)}) + \ldots + \hat{\beta}_m(x^{(m)} - \bar{x}^{(m)})$ are unbiased estimators for $\alpha, \beta_1, \ldots, \beta_m$.

Exercise 4.8

Suppose $Y_1, \ldots, Y_n$ are uncorrelated and all have the same variance σ^2. With $(x_1^{(1)}, \ldots, x_1^{(m)}), \ldots, (x_n^{(1)}, \ldots, x_n^{(m)}) \in \mathbb{R}^m$ as n observations of m explanatory variables, what's the variance of $\hat{\alpha}$ and the covariance matrix of $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \ldots, \hat{\beta}_m)$ in the least-squares hyperplane $y = \hat{\alpha} + \hat{\beta}_1(x^{(1)} - \bar{x}^{(1)}) + \ldots + \hat{\beta}_m(x^{(m)} - \bar{x}^{(m)})$?

Exercise 4.10

Let $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^m$ be n observations of m explanatory variables, and suppose that the response variable $Y_1, \ldots, Y_n$ satisfies a linear model

$$Y_i = \gamma_1 g_1(\mathbf{x}_i) + \ldots + \gamma_d g_d(\mathbf{x}_i) + \epsilon_i.$$

Assuming the columns of the design matrix are linearly independent, show that the coefficients $\hat{\gamma}_1, \ldots, \hat{\gamma}_d$ in the least-squares linear fit $y = \hat{\gamma}_1 g_1(\mathbf{x}) + \ldots + \hat{\gamma}_d g_d(\mathbf{x})$ are unbiased estimators for $\gamma_1, \ldots, \gamma_d$.

Exercise 4.11

Suppose $Y_1, \ldots, Y_n$ are uncorrelated and all have the same variance σ^2. With $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^m$ as n observations of m explanatory variables, what's the covariance matrix of $\hat{\boldsymbol{\gamma}} = (\hat{\gamma}_1, \ldots, \hat{\gamma}_d)$ in the least-squares linear fit $y = \hat{\gamma}_1 g_1(\mathbf{x}) + \ldots + \hat{\gamma}_d g_d(\mathbf{x})$?

Remember that $\hat{\alpha} = \overline{Y}$ has the representation $\alpha + \frac{1}{n}\sum_i \epsilon_i$. Its variance once again works out to be $\frac{\sigma^2}{n}$. The covariance matrix of $\hat{\boldsymbol{\beta}}$ is

$$\begin{aligned}
\operatorname{cov}\hat{\boldsymbol{\beta}} &= \operatorname{cov}\Sigma^{-}\tfrac{1}{n}\widetilde{\mathbb{X}}^T\mathbf{Y} \\
&= \Sigma^{-}\tfrac{1}{n}\widetilde{\mathbb{X}}^T(\sigma^2\mathbb{I})[\Sigma^{-}\tfrac{1}{n}\widetilde{\mathbb{X}}^T]^T \\
&= \tfrac{\sigma^2}{n}\Sigma^{-}(\underbrace{\tfrac{1}{n}\widetilde{\mathbb{X}}^T\widetilde{\mathbb{X}}}_{\Sigma})\Sigma^{-} \\
&= \tfrac{\sigma^2}{n}\Sigma^{-}
\end{aligned}$$

by Exercise 1.75.

The *least-squares hyperplane* has $\hat{\alpha} = \overline{Y}$, which can be expressed as

$$\begin{aligned}
\overline{Y} &= \tfrac{1}{n}\sum_i Y_i \\
&= \tfrac{1}{n}\sum_i [\alpha + \beta_1(x_i^{(1)} - \bar{x}^{(1)}) + \ldots + \beta_m(x_i^{(m)} - \bar{x}^{(m)}) + \epsilon_i] \\
&= \alpha + \beta_1 \underbrace{\tfrac{1}{n}\sum_i (x_i^{(1)} - \bar{x}^{(1)})}_{0} + \ldots + \beta_m \underbrace{\tfrac{1}{n}\sum_i (x_i^{(m)} - \bar{x}^{(m)})}_{0} + \tfrac{1}{n}\sum_i \epsilon_i \\
&= \alpha + \tfrac{1}{n}\sum_i \epsilon_i.
\end{aligned}$$

Its expectation is $\mathbb{E}\overline{Y} = \alpha + \frac{1}{n}\sum_i \underbrace{\mathbb{E}\epsilon_i}_{0} = \alpha$. The vector of empirical covariances of $\mathbf{Y}$ with $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(m)}$ can be expressed as $\frac{1}{n}\widetilde{\mathbb{X}}^T\mathbf{Y}$ where $\widetilde{\mathbb{X}}$ is the centered version of the explanatory data matrix. Substituting this representation into the formula from Theorem 2.3,

$$\begin{aligned}
\mathbb{E}\hat{\boldsymbol{\beta}} &= \mathbb{E}\Sigma^{-1}\tfrac{1}{n}\widetilde{\mathbb{X}}^T\mathbf{Y} \\
&= \Sigma^{-1}\tfrac{1}{n}\widetilde{\mathbb{X}}^T\mathbb{E}\mathbf{Y} \\
&= \Sigma^{-1}\tfrac{1}{n}\widetilde{\mathbb{X}}^T(\alpha\mathbf{1} + \widetilde{\mathbb{X}}\boldsymbol{\beta}) \\
&= \Sigma^{-1}(\tfrac{\alpha}{n}\underbrace{\widetilde{\mathbb{X}}^T\mathbf{1}}_{0} + \underbrace{\tfrac{1}{n}\widetilde{\mathbb{X}}^T\widetilde{\mathbb{X}}}_{\Sigma}\boldsymbol{\beta}) \\
&= \Sigma^{-1}\Sigma\boldsymbol{\beta} \\
&= \boldsymbol{\beta}.
\end{aligned}$$

The covariance matrix of $\hat{\boldsymbol{\gamma}}$ is

$$\begin{aligned}
\operatorname{cov}\hat{\boldsymbol{\gamma}} &= \operatorname{cov}(\mathbb{M}^T\mathbb{M})^{-}\mathbb{M}^T\mathbf{Y} \\
&= (\mathbb{M}^T\mathbb{M})^{-}\mathbb{M}^T(\sigma^2\mathbb{I})[(\mathbb{M}^T\mathbb{M})^{-}\mathbb{M}^T]^T \\
&= \sigma^2(\mathbb{M}^T\mathbb{M})^{-}\mathbb{M}^T\mathbb{M}(\mathbb{M}^T\mathbb{M})^{-} \\
&= \sigma^2(\mathbb{M}^T\mathbb{M})^{-}.
\end{aligned}$$

by Exercise 1.75.

Let $\mathbb{M}$ represent the design matrix

$$\mathbb{M} := \begin{bmatrix} g_1(\mathbf{x}_1) & \cdots & g_d(\mathbf{x}_1) \\ \vdots & \ddots & \vdots \\ g_1(\mathbf{x}_n) & \cdots & g_d(\mathbf{x}_n) \end{bmatrix}.$$

The expectation of $\mathbf{Y} = \mathbb{M}\boldsymbol{\gamma} + \boldsymbol{\epsilon}$ is $\mathbb{M}\boldsymbol{\gamma}$. Using the formula for the least-squares coefficients provided in Theorem 2.4,

$$\begin{aligned}
\mathbb{E}\hat{\boldsymbol{\gamma}} &= \mathbb{E}(\mathbb{M}^T\mathbb{M})^{-1}\mathbb{M}^T\mathbf{Y} \\
&= (\mathbb{M}^T\mathbb{M})^{-1}\mathbb{M}^T\underbrace{\mathbb{E}\mathbf{Y}}_{\mathbb{M}\boldsymbol{\gamma}} \\
&= (\mathbb{M}^T\mathbb{M})^{-1}(\mathbb{M}^T\mathbb{M})\boldsymbol{\gamma} \\
&= \boldsymbol{\gamma}.
\end{aligned}$$

(We know that $\mathbb{M}^T\mathbb{M}$ is invertible because the columns of $\mathbb{M}$ are assumed to be linearly independent – see Exercise 1.63.)

Exercise 4.15

Suppose $\mathbf{Y} = \mathbb{M}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ with $\operatorname{cov} \boldsymbol{\epsilon} = \sigma^2 \mathbb{I}_n$, and let $\widehat{\mathbf{Y}}$ be the orthogonal projection of $\mathbf{Y}$ onto $C(\mathbb{M})$. Find $\mathbb{E}\|\mathbf{Y} - \widehat{\mathbf{Y}}\|^2$, the expected sum of squared residuals.

Exercise 4.17

Suppose $Y_1, \ldots, Y_n$ are uncorrelated and all have the same variance σ^2. Let $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^m$ be n observations of m explanatory variables, and assume their empirical covariance matrix Σ has full rank. Let $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \ldots, \hat{\beta}_m)$ be the coefficients of the explanatory variables in the least-squares hyperplane $y = \hat{\alpha} + \hat{\beta}_1(x^{(1)} - \bar{x}^{(1)}) + \ldots + \hat{\beta}_m(x^{(m)} - \bar{x}^{(m)})$. Find $\mathbb{E}\|\hat{\boldsymbol{\beta}} - \mathbb{E}\hat{\boldsymbol{\beta}}\|^2$ in terms of σ^2, n, and the eigenvalues of Σ.

Exercise 4.18

Based on Exercise 4.12, the Gauss-Markov theorem implies that the least-squares coefficient vector has the smallest possible expected squared estimation error among all random vectors that are both linear functions of the response and unbiased for its expectation. However, Equation 4.1 identified $a < 1$ such that a times the least-squares coefficients of the explanatory variables has smaller expected squared estimation error than the least-squares coefficient vector do; explain why this doesn't contradict the Gauss-Markov theorem.

Exercise 5.1

Find the probability density function for a standard Normal random vector on $\mathbb{R}^n$.

The "variance" of any random vector is the trace of its covariance matrix (Exercise 3.25).

$$\begin{aligned}\mathbb{E}\|\hat{\boldsymbol{\beta}} - \mathbb{E}\hat{\boldsymbol{\beta}}\|^2 &= \operatorname{tr}\operatorname{cov}\hat{\boldsymbol{\beta}} \\ &= \tfrac{\sigma^2}{n}\operatorname{tr}\Sigma^{-1} \\ &= \tfrac{\sigma^2}{n}(\lambda_1^{-1} + \ldots + \lambda_m^{-1})\end{aligned}$$

where $\lambda_1, \ldots, \lambda_m$ are the eigenvalues of Σ.

We'll let $\mathbb{H}$ be the orthogonal projection matrix onto $\mathbb{M}$, and use Exercise 3.35 along with Exercises 1.67 and 1.68.

$$\begin{aligned}\mathbb{E}\|\mathbf{Y} - \widehat{\mathbf{Y}}\|^2 &= \|(\mathbb{I} - \mathbb{H})\boldsymbol{\epsilon}\|^2 \\ &= \operatorname{tr}\left[(\mathbb{I} - \mathbb{H})\sigma^2\mathbb{I}\right] \\ &= \sigma^2(n - \operatorname{rank}\mathbb{M})\end{aligned}$$

Let $\mathbf{Z}$ be an $\mathbb{R}^n$-valued standard Normal random vector. By independence, its pdf equals the product of the individual pdfs of its coordinates $(Z_1, \ldots, Z_n)$.

$$\begin{aligned}f(\mathbf{z}) &= \prod_i \tfrac{1}{\sqrt{2\pi}} e^{-z_i^2/2} \\ &= \tfrac{1}{(2\pi)^{n/2}} e^{-(z_1^2 + \ldots + z_n^2)/2} \\ &= \tfrac{1}{(2\pi)^{n/2}} e^{-\|\mathbf{z}\|^2/2}\end{aligned}$$

Let's check the conditions of the Gauss-Markov theorem. It applies to linear functions of the response $\mathbf{Y}$ that are unbiased for $\mathbb{E}\mathbf{Y}$. Because $\hat{\boldsymbol{\beta}}$ is linear in $\mathbf{Y}$, so is $a\hat{\boldsymbol{\beta}}$. However, it's *biased*; its expectation is $a\boldsymbol{\beta} \neq \boldsymbol{\beta}$, so Gauss-Markov doesn't apply.

Exercise 5.3

Show that if $\mathbf{X}$ is a Normal random vector, then so is $\mathbb{M}\mathbf{X} + \mathbf{v}$ where $\mathbb{M}$ is a real matrix and $\mathbf{v}$ is a vector.

Exercise 5.6

Find the expectation of $W \sim \chi^2_k$.

Exercise 5.7

If $\mathbf{Y} \sim N(\boldsymbol{\mu}, \mathbb{C})$ is an $\mathbb{R}^n$-valued random vector, what's the distribution of the squared Mahalanobis distance of $\mathbf{Y}$ from its own distribution?

Exercise 5.8

Let $\mathbf{Z}$ be an $\mathbb{R}^n$-valued random vector with the standard Normal distribution, and let $\mathbb{H}$ be an orthogonal projection matrix. Find the distribution of $\|\mathbb{H}\mathbf{Z}\|^2$.

W can be represented as the squared norm of a standard Normal random vector. Its expectation is the same as the expected squared norm of *any* standardized random vector $\mathbf{Z}$ on $\mathbb{R}^k$:

$$\begin{aligned}\mathbb{E}\|\mathbf{Z}\|^2 &= \mathbb{E}(Z_1^2 + \ldots + Z_k^2) \\ &= \mathbb{E}Z_1^2 + \ldots + \mathbb{E}Z_k^2 \\ &= \underbrace{\operatorname{var} Z_1}_{1} + \ldots + \underbrace{\operatorname{var} Z_k}_{1} \\ &= k.\end{aligned}$$

With $\boldsymbol{\mu}$ and $\mathbb{C}$ representing the expectation and covariance of $\mathbf{X}$, the transformed random vector is

$$\begin{aligned}\mathbb{M}\mathbf{X} + \mathbf{v} &= \mathbb{M}(\mathbb{C}^{1/2}\mathbf{Z} + \boldsymbol{\mu}) + \mathbf{v} \\ &= [\mathbb{M}\mathbb{C}^{1/2}]\mathbf{Z} + [\mathbb{M}\boldsymbol{\mu} + \mathbf{v}]\end{aligned}$$

with $\mathbf{Z}$ standard Normal. This fits the definition of a Normal random vector.

Let $\mathbf{u}_1, \ldots, \mathbf{u}_n$ be an orthonormal basis with $\mathbf{u}_1, \ldots, \mathbf{u}_{\operatorname{rank}\mathbb{H}}$ spanning the space that $\mathbb{H}$ projects onto. Because the orthogonal projection is

$$\mathbb{H}\mathbf{Z} = \langle \mathbf{Z}, \mathbf{u}_1 \rangle \mathbf{u}_1 + \ldots + \langle \mathbf{Z}, \mathbf{u}_{\operatorname{rank}\mathbb{H}} \rangle \mathbf{u}_{\operatorname{rank}\mathbb{H}}$$

its squared length is the sum of its squared coordinates

$$\|\mathbb{H}\mathbf{Z}\|^2 = \langle \mathbf{Z}, \mathbf{u}_1 \rangle^2 + \ldots + \langle \mathbf{Z}, \mathbf{u}_{\operatorname{rank}\mathbb{H}} \rangle^2.$$

These coordinates are independent standard Normal random variables, according to the discussion in Section 5.1, so their sum of squares has distribution $\chi^2_{\operatorname{rank}\mathbb{H}}$.

Allow for degenerate distributions by using the approach described at the end of Section 3.5. Let $\mathbf{Z} := \mathbb{C}^{-1/2}(\mathbf{Y} - \boldsymbol{\mu}) \sim N(\mathbf{0}, \mathbb{I})$ represent the standardized version in $\mathbb{R}^{\operatorname{rank}\mathbb{C}}$. The squared Mahalanobis distance from $\mathbf{Y}$ to $N(\boldsymbol{\mu}, \mathbb{C})$ is

$$\begin{aligned}\|\mathbb{C}^{-1/2}[\mathbf{Y} - \boldsymbol{\mu}]\|^2 &= \|\mathbb{C}^{-1/2}[(\mathbb{C}^{1/2}\mathbf{Z} + \boldsymbol{\mu}) - \boldsymbol{\mu}]\|^2 \\ &= \|\mathbf{Z}\|^2 \\ &\sim \chi^2_{\operatorname{rank}\mathbb{C}}.\end{aligned}$$

Exercise 5.9

Let $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 \mathbb{I})$. If $\mathbb{H}$ is an orthogonal projection matrix and $\mathbf{u}$ is a unit vector orthogonal to $C(\mathbb{H})$, find the distribution of

$$\frac{\langle \boldsymbol{\epsilon}, \mathbf{u} \rangle}{\|\mathbb{H}\boldsymbol{\epsilon}\| / \sqrt{\operatorname{rank} \mathbb{H}}}.$$

Exercise 5.10

Let $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 \mathbb{I})$. If $\mathbb{H}$ is an orthogonal projection matrix and $\mathbf{u}$ is a unit vector orthogonal to $C(\mathbb{H})$, and $a \in \mathbb{R}$, find the distribution of

$$\frac{a + \langle \boldsymbol{\epsilon}, \mathbf{u} \rangle}{\|\mathbb{H}\boldsymbol{\epsilon}\| / \sqrt{\operatorname{rank} \mathbb{H}}}.$$

Exercise 5.11

Let $T \sim t_k$. What's the distribution of T^2?

Exercise 5.12

Let $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 \mathbb{I})$, and let $\mathbb{H}_1$ and $\mathbb{H}_2$ be orthogonal projection matrices onto two subspaces that are orthogonal to each other. Find the distribution of $\frac{\|\mathbb{H}_1\boldsymbol{\epsilon}\|^2 / \operatorname{rank} \mathbb{H}_1}{\|\mathbb{H}_2\boldsymbol{\epsilon}\|^2 / \operatorname{rank} \mathbb{H}_2}$.

First, we'll divide the numerator and the denominator by σ.

$$\frac{a + \langle \boldsymbol{\epsilon}, \mathbf{u} \rangle}{\|\mathbb{H}\boldsymbol{\epsilon}\|/\sqrt{\operatorname{rank} \mathbb{H}}} = \frac{a/\sigma + \langle (\boldsymbol{\epsilon}/\sigma), \mathbf{u} \rangle}{\|\mathbb{H}(\boldsymbol{\epsilon}/\sigma)\|/\sqrt{\operatorname{rank} \mathbb{H}}}$$

As in Exercise 5.9, the second term in the numerator is standard Normal, the denominator is $\chi^2_{\operatorname{rank} \mathbb{H}}$ divided by its degrees of freedom, and the numerator and denominator are independent. By the definition of non-central t-distributions, the ratio's distribution is $t_{\operatorname{rank} \mathbb{H}, a/\sigma}$.

First, we'll divide the numerator and the denominator by σ to connect this ratio to the standard Normal random vector $\boldsymbol{\epsilon}/\sigma$.

$$\frac{\langle \boldsymbol{\epsilon}, \mathbf{u} \rangle}{\|\mathbb{H}\boldsymbol{\epsilon}\|/\sqrt{\operatorname{rank} \mathbb{H}}} = \frac{\langle (\boldsymbol{\epsilon}/\sigma), \mathbf{u} \rangle}{\|\mathbb{H}(\boldsymbol{\epsilon}/\sigma)\|/\sqrt{\operatorname{rank} \mathbb{H}}}$$

Let $\mathbf{u}_1, \ldots, \mathbf{u}_n$ be an orthonormal basis with $\mathbf{u}_1, \ldots, \mathbf{u}_{\operatorname{rank} \mathbb{H}}$ spanning $C(\mathbb{H})$ and $\mathbf{u}_{\operatorname{rank} \mathbb{H}+1}$ equal to $\mathbf{u}$. The numerator is simply the coordinate of $\boldsymbol{\epsilon}/\sigma$ with respect to $\mathbf{u}$, so it's a standard Normal random variable. From Exercise 5.8, $\|\mathbb{H}(\boldsymbol{\epsilon}/\sigma)\|^2 \sim \chi^2_{\operatorname{rank} \mathbb{H}}$. Because the numerator and the denominator are functions of distinct coordinates, they're independent of each other, so the random variable has the $t_{\operatorname{rank} \mathbb{H}}$ distribution.

We can divide both the numerator and the denominator by σ^2 to produce random variables whose distributions we know from Exercise 5.8.

$$\frac{\|\mathbb{H}_1\boldsymbol{\epsilon}\|^2/\operatorname{rank} \mathbb{H}_1}{\|\mathbb{H}_2\boldsymbol{\epsilon}\|^2/\operatorname{rank} \mathbb{H}_2} = \frac{\|\mathbb{H}_1(\boldsymbol{\epsilon}/\sigma)\|^2/\operatorname{rank} \mathbb{H}_1}{\|\mathbb{H}_2(\boldsymbol{\epsilon}/\sigma)\|^2/\operatorname{rank} \mathbb{H}_2}$$

The numerator is a $\chi^2_{\operatorname{rank} \mathbb{H}_1}$-distributed random variable divided by its degrees of freedom, while the denominator is a $\chi^2_{\operatorname{rank} \mathbb{H}_2}$-distributed random variable divided by its degrees of freedom. Because the subspaces are orthogonal, we know that the two orthogonal projections are independent of each other, allowing us to conclude that the ratio matches the definition of $f_{\operatorname{rank} \mathbb{H}_1, \operatorname{rank} \mathbb{H}_2}$.

From the definition of t_k, we can represent T using independent $Z \sim N(0, 1)$ and $V \sim \chi^2_k$.

$$\begin{aligned} T^2 &= \left(\frac{Z}{\sqrt{V/k}} \right)^2 \\ &= \frac{Z^2/1}{V/k} \end{aligned}$$

Because $Z^2 \sim \chi^2_1$, this expression matches the definition of the $f_{1,k}$ distribution.

Exercise 5.13

Let $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 \mathbb{I})$, and let $\mathbb{H}_1$ and $\mathbb{H}_2$ be orthogonal projection matrices onto two subspaces that are orthogonal to each other. Find the distribution of $\frac{\|\mathbf{v}+\mathbb{H}_1\boldsymbol{\epsilon}\|^2/\operatorname{rank}\mathbb{H}_1}{\|\mathbb{H}_2\boldsymbol{\epsilon}\|^2/\operatorname{rank}\mathbb{H}_2}$, where $\mathbf{v}$ is a non-random vector.

Exercise 6.1

Let $\mathbf{x}_i$ represent the explanatory value(s) of the ith observation. Consider modeling the response variable by

$$Y_i = f_\theta(\mathbf{x}_i) + \epsilon_i$$

with $\epsilon_1, \ldots, \epsilon_n \overset{iid}{\sim} N(0, \sigma^2)$ and $\theta \in \Theta$ indexing a set of possible functions. (Notice that this form is far more general than the linear model with iid Normal errors.) Show that the maximum likelihood estimator for θ is precisely the parameter value that minimizes the sum of squared residuals.

Exercise 6.2

Suppose $\mathbf{Y} = \mathbb{M}\boldsymbol{\gamma} + \boldsymbol{\epsilon}$ with error vector $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 \mathbb{I})$. If $\widehat{\mathbf{Y}}$ is the least-squares linear regression's prediction vector for design matrix $\mathbb{M}$, what's the distribution of $\|\mathbf{Y} - \widehat{\mathbf{Y}}\|^2/\sigma^2$?

Exercise 6.3

Let $\widetilde{\mathbb{X}} \in \mathbb{R}^{n \times m}$ be a centered explanatory data matrix with full rank. Assume

$$\mathbf{Y} = \alpha + \widetilde{\mathbb{X}}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

with $\epsilon_1, \ldots, \epsilon_n \overset{iid}{\sim} N(0, \sigma^2)$. Write the standardized version of the least-squares slope $\hat{\beta}_j$. What is its distribution?

The response values have distribution $Y_i \sim N(f_\theta(\mathbf{x}_i), \sigma^2)$ and are independent of each other. Because of independence, the overall likelihood $L(\theta; \mathbf{Y})$ is the product of the individual observations' likelihoods.

$$\begin{aligned} L(\theta; \mathbf{Y}) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(Y_i - f_\theta(\mathbf{x}_i))^2} \\ &= \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - f_\theta(\mathbf{x}_i))^2} \end{aligned}$$

The parameter θ only appears in the sum of squared residuals $\sum_{i=1}^{n}(Y_i - f_\theta(\mathbf{x}_i))^2$. The smaller the sum of squared residuals is, the larger the likelihood is, so the "least-squares parameter" is exactly the maximum likelihood estimator. Notice that this equivalence doesn't depend on the value of σ and that it holds even if σ is unknown.

Divide both the numerator and the denominator by σ^2.

$$\frac{\|\mathbf{v} + \mathbb{H}_1\boldsymbol{\epsilon}\|^2/\operatorname{rank}\mathbb{H}_1}{\|\mathbb{H}_2\boldsymbol{\epsilon}\|^2/\operatorname{rank}\mathbb{H}_2} = \frac{\|\frac{1}{\sigma}\mathbf{v} + \mathbb{H}_1(\boldsymbol{\epsilon}/\sigma)\|^2/\operatorname{rank}\mathbb{H}_1}{\|\mathbb{H}_2(\boldsymbol{\epsilon}/\sigma)\|^2/\operatorname{rank}\mathbb{H}_2}$$

As in Exercise 5.12, the denominator is $\chi^2_{\operatorname{rank}\mathbb{H}_2}$-distributed and is independent of the numerator. This time the numerator is a non-central χ^2 random variable divided by its degrees of freedom with non-centrality parameter $\|\frac{1}{\sigma}\mathbf{v}\|^2 = \|\mathbf{v}\|^2/\sigma^2$. Thus the ratio's distribution matches the definition of $f_{\operatorname{rank}\mathbb{H}_1, \operatorname{rank}\mathbb{H}_2, \|\mathbf{v}\|^2/\sigma^2}$.

The distribution of $\hat{\beta}_j$ is Normal because its a linear transformation of $\mathbf{Y}$ which is Normal. Its expectation equals the jth entry of $\mathbb{E}\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}$, and its variance equals the jth diagonal of $\operatorname{cov}\hat{\boldsymbol{\beta}} = \frac{\sigma^2}{n}\Sigma^{-1}$:

$$\hat{\beta}_j \sim N(\beta_j, \tfrac{\sigma^2}{n}\mathbb{\Sigma}^{-1}_{jj}).$$

The standardized version is

$$\frac{\hat{\beta}_j - \beta_j}{\frac{\sigma}{\sqrt{n}}\sqrt{\mathbb{\Sigma}^{-1}_{jj}}} \sim N(0, 1).$$

We saw in Chapter 4 that the least-squares residual vector $\mathbf{Y} - \widehat{\mathbf{Y}}$ is the orthogonal projection of $\boldsymbol{\epsilon}$ onto $C(\mathbb{M})^\perp$ which has dimension $n - \operatorname{rank}\mathbb{M}$. The standardized version $\boldsymbol{\epsilon}/\sigma$ is standard Normal, so according to Exercise 5.8,

$$\begin{aligned} \frac{\|\mathbf{Y} - \widehat{\mathbf{Y}}\|^2}{\sigma^2} &= \|(\mathbb{I} - \mathbb{H})(\boldsymbol{\epsilon}/\sigma)\|^2 \\ &\sim \chi^2_{n - \operatorname{rank}\mathbb{M}} \end{aligned}$$

where $\mathbb{H}$ represents the orthogonal projection matrix onto $C(\mathbb{M})$.

Exercise 6.4

Let $\widetilde{\mathbb{X}} \in \mathbb{R}^{n \times m}$ be a centered explanatory data matrix with full rank. Assume

$$\mathbf{Y} = \alpha + \widetilde{\mathbb{X}}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

with $\epsilon_1, \ldots, \epsilon_n \overset{iid}{\sim} N(0, \sigma^2)$. Devise a t-distributed random variable involving the least-squares slope $\hat{\beta}_j$.

Exercise 6.5

Let $\widetilde{\mathbb{X}} \in \mathbb{R}^{n \times m}$ be a centered explanatory data matrix with full rank. Assume

$$\mathbf{Y} = \alpha + \widetilde{\mathbb{X}}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

with $\epsilon_1, \ldots, \epsilon_n \overset{iid}{\sim} N(0, \sigma^2)$. Devise a 95% confidence interval for β_j.

Exercise 6.6

Let $\widetilde{\mathbb{X}} \in \mathbb{R}^{n \times m}$ be a centered explanatory data matrix with full rank. Assume

$$\mathbf{Y} = \alpha + \widetilde{\mathbb{X}}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

with $\epsilon_1, \ldots, \epsilon_n \overset{iid}{\sim} N(0, \sigma^2)$. Devise a test statistic T_j for the null hypothesis that $\beta_j = 0$.

Exercise 6.7

Section 6.3.7.2 described a test statistic (Equation 6.1) for the null hypothesis that all of the slopes in a multiple linear model are 0. Is it the same as the test statistic prescribed by Equation 6.3?

From Exercise 6.4, $\frac{\hat{\beta}_j - \beta_j}{\frac{\hat{\sigma}}{\sqrt{n}}\sqrt{\mathbb{\Sigma}^{-1}_{jj}}} \sim t_{n-m-1}$, so

$$\mathbb{P}\left\{ -\tau^{-1}_{n-m-1}(.975) \leq \frac{\beta_j - \hat{\beta}_j}{\frac{\hat{\sigma}}{\sqrt{n}}\sqrt{\mathbb{\Sigma}^{-1}_{jj}}} \leq \tau^{-1}_{n-m-1}(.975) \right\} = .95.$$

The event can be rewritten as

$$\hat{\beta}_j - \tau^{-1}_{n-m-1}(.975)\frac{\hat{\sigma}}{\sqrt{n}}\sqrt{\mathbb{\Sigma}^{-1}_{jj}} \leq \beta_j \leq \hat{\beta}_j + \tau^{-1}_{n-m-1}(.975)\frac{\hat{\sigma}}{\sqrt{n}}\sqrt{\mathbb{\Sigma}^{-1}_{jj}}$$

which means that $\hat{\beta}_j \pm \tau^{-1}_{n-m-1}(.975)\frac{\hat{\sigma}}{\sqrt{n}}\sqrt{\mathbb{\Sigma}^{-1}_{jj}}$ is a 95% confidence interval for β_j.

From Exercise 6.3, the standardized version is $\frac{\hat{\beta}_j - \beta_j}{\frac{\sigma}{\sqrt{n}}\sqrt{\mathbb{\Sigma}^{-1}_{jj}}} \sim N(0,1)$. Exercise 2.8 implies that $\hat{\boldsymbol{\beta}}$ is a function of $\mathbb{H}\boldsymbol{\epsilon}$, so the ratio trick allows us to substitute $\hat{\sigma}$ for σ to derive

$$\frac{\hat{\beta}_j - \beta_j}{\frac{\hat{\sigma}}{\sqrt{n}}\sqrt{\mathbb{\Sigma}^{-1}_{jj}}} \sim t_{n-m-1}.$$

The null hypothesis is that $\mathbb{E}\mathbf{Y}$ is in the span of $\mathbf{1}$, so the general approach (Equation 6.3) uses the test statistic

$$\frac{\|\widehat{\mathbf{Y}} - \overline{Y}\mathbf{1}\|^2/m}{\hat{\sigma}^2} \sim f_{m,n-m-1},$$

while Section 6.3.7.2 derived the test statistic

$$\frac{n\|\Sigma^{1/2}\hat{\boldsymbol{\beta}}\|^2/m}{\hat{\sigma}^2} \sim f_{m,n-m-1}.$$

Let's analyze the factor in which they appear to differ. Recall that for multiple linear regression the least-squares prediction vector can be expressed as

$$\widehat{\mathbf{Y}} = \overline{Y}\mathbf{1} + \tilde{\mathbb{X}}\hat{\boldsymbol{\beta}}$$

where $\tilde{\mathbb{X}}$ is the centered explanatory data matrix. Therefore,

$$\begin{aligned}\|\widehat{\mathbf{Y}} - \overline{Y}\mathbf{1}\|^2 &= \|\tilde{\mathbb{X}}\hat{\boldsymbol{\beta}}\|^2 \\ &= \hat{\boldsymbol{\beta}}^T\tilde{\mathbb{X}}^T\tilde{\mathbb{X}}\hat{\boldsymbol{\beta}}.\end{aligned}$$

And in the other test statistic,

$$\begin{aligned}n\|\Sigma^{1/2}\hat{\boldsymbol{\beta}}\|^2 &= n\hat{\boldsymbol{\beta}}^T \underbrace{\Sigma}_{\frac{1}{n}\tilde{\mathbb{X}}^T\tilde{\mathbb{X}}} \hat{\boldsymbol{\beta}} \\ &= \hat{\boldsymbol{\beta}}^T\tilde{\mathbb{X}}^T\tilde{\mathbb{X}}\hat{\boldsymbol{\beta}}\end{aligned}$$

so they turn out to be exactly the same.

From Exercise 6.4, $\frac{\hat{\beta}_j - \beta_j}{\frac{\hat{\sigma}}{\sqrt{n}}\sqrt{\mathbb{\Sigma}^{-1}_{jj}}} \sim t_{n-m-1}$. The assumption that $\beta_j = 0$ leads to the test statistic

$$\begin{aligned}T_j &:= \frac{\hat{\beta}_j}{\frac{\hat{\sigma}}{\sqrt{n}}\sqrt{\mathbb{\Sigma}^{-1}_{jj}}} \\ &\sim t_{n-m-1}.\end{aligned}$$

The significance probability is $2\tau_{n-m-1}(-|T_j|)$.

Exercise 6.8

Let $\mathbf{Y} = \alpha\mathbf{1} + \widetilde{\mathbb{X}}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ with $\widetilde{\mathbb{X}} \in \mathbb{R}^{n\times m}$ representing a centered data matrix. If $\widehat{\mathbf{Y}}$ is the least-squares prediction vector that comes from multiple linear regression, find the distribution of

$$\frac{\|\widehat{\mathbf{Y}} - \overline{Y}\mathbf{1}\|^2}{\hat{\sigma}^2}.$$

Based on the preceding discussion, the statistic in question has non-central f-distribution. $\widehat{\mathbf{Y}}$ is the projection onto an $(m+1)$-dimensional subspace, while $\overline{Y}\mathbf{1}$ is the projection onto a 1-dimensional subspace. Thus the numerator has m degrees of freedom, and the denominator has $n-m-1$ degrees of freedom. The non-centrality parameter is

$$\|(\alpha\mathbf{1}+\widetilde{\mathbb{X}}\boldsymbol{\beta})-(\alpha\mathbf{1})\|^2/\sigma^2 = \|\widetilde{\mathbb{X}}\boldsymbol{\beta}\|^2/\sigma^2.$$