Exercise 1.1

Let $\mathbf{0}$ be the zero vector, and let a be a scalar. Show that $a\mathbf{0} = \mathbf{0}$.

Exercise 1.2

Let $\mathbf{v}$ be a vector, and let 0 be the zero scalar. Show that $0\mathbf{v}$ equals the zero vector.

Exercise 1.3

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

Exercise 1.4

Show that the span of $\mathbf{v}_1, \ldots, \mathbf{v}_m$ is the same as the span of $\mathbf{v}_1 + a_1\mathbf{v}_m, \ldots, \mathbf{v}_{m-1} + a_{m-1}\mathbf{v}_m, \mathbf{v}_m$ for any scalars $a_1, \ldots, a_{m-1}$.

The zero scalar satisfies $a-a=0$ for every a in the scalar field; in particular, $0-0=0$. We make this substitution, then distribute and invoke the fact that $\mathbf{v}-\mathbf{v}$ equals the zero vector $\mathbf{0}$ for any vector $\mathbf{v}$.

$$\begin{aligned} 0\mathbf{v} &= (0-0)\mathbf{v} \\ &= 0\mathbf{v} - 0\mathbf{v} \\ &= \mathbf{0} \end{aligned}$$

By definiton, a vector space has the property that $\mathbf{v}-\mathbf{v}=\mathbf{0}$ for any vector $\mathbf{v}$. In particular, $\mathbf{0}-\mathbf{0}=\mathbf{0}$.

$$\begin{aligned} a\mathbf{0} &= a(\mathbf{0}-\mathbf{0}) \\ &= a\mathbf{0} - a\mathbf{0} \\ &= \mathbf{0} \end{aligned}$$

We distributed scalar multiplication then used the $\mathbf{v}-\mathbf{v}=\mathbf{0}$ property again.

We'll show that an arbitrary linear combination of $\mathbf{v}_1, \ldots, \mathbf{v}_m$ can also be represented as a linear combination of the altered vectors by adding and subtracting the appropriately scaled versions of $\mathbf{v}_m$.

$$\begin{aligned} &b_1\mathbf{v}_1 + \ldots + b_{m-1}\mathbf{v}_{m-1} + b_m\mathbf{v}_m \\ &\quad = b_1(\mathbf{v}_1 + a_1\mathbf{v}_m) + \ldots + b_{m-1}(\mathbf{v}_{m-1} + a_{m-1}\mathbf{v}_m) + (b_m - b_1a_1 - \ldots - b_{m-1}a_{m-1})\mathbf{v}_m \end{aligned}$$

Similarly, an arbitrary linear combination of the altered vectors becomes a linear combination of $\mathbf{v}_1, \ldots, \mathbf{v}_m$ by distributing the coefficients and regrouping the terms.

$$\begin{aligned} &b_1(\mathbf{v}_1 + a_1\mathbf{v}_1) + \ldots + b_{m-1}(\mathbf{v}_{m-1} + a_{m-1}\mathbf{v}_m) + b_m\mathbf{v}_m \\ &\quad = b_1\mathbf{v}_1 + \ldots + b_{m-1}\mathbf{v}_{m-1} + (b_m + b_1a_1 + \ldots + b_{m-1}a_{m-1})\mathbf{v}_m \end{aligned}$$

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.

Exercise 1.5

Let $\mathcal{F}$ be a field, and suppose $\mathbb{T}$ is a linear operator from $\mathcal{F}^m$ to $\mathcal{V}$. Show that $\mathbb{T}$ can be represented as a mapping of $\mathbf{b} \in \mathcal{F}^m$ to $\begin{bmatrix}\mathbf{v}_1 & \cdots & \mathbf{v}_m\end{bmatrix}\mathbf{b}$ for some $\mathbf{v}_1, \ldots, \mathbf{v}_m$.

Exercise 1.6

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

Exercise 1.7

Show that if $\mathbf{v}_1, \ldots, \mathbf{v}_m$ are linearly independent then $(b_1, \ldots, b_m) = (0, \ldots, 0)$ is the only vector of scalars for which $b_1\mathbf{v}_1 + \ldots + b_m\mathbf{v}_m = \mathbf{0}$.

Exercise 1.8

Show that if $(b_1, \ldots, b_m) = (0, \ldots, 0)$ is the only vector of scalars for which $b_1\mathbf{v}_1 + \ldots + b_m\mathbf{v}_m = \mathbf{0}$, then $\mathbf{v}_1, \ldots, \mathbf{v}_m$ are linearly independent.

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.

First, consider that $\mathbb{T}$ and $\begin{bmatrix}\mathbf{v}_1 & \cdots & \mathbf{v}_m\end{bmatrix}$ would have to agree on where to map the standard basis vectors. We see that $\begin{bmatrix}\mathbf{v}_1 & \cdots & \mathbf{v}_m\end{bmatrix}$ maps $\mathbf{e}_1 := (1, 0, \ldots, 0)$ to $\mathbf{v}_1$, so $\mathbf{v}_1$ must be $\mathbb{T}\mathbf{e}_1$. Likewise $\mathbf{v}_2$ would need to $\mathbb{T}\mathbf{e}_2$, and so on. Let's check that this proposal $\begin{bmatrix}\mathbb{T}\mathbf{e}_1 & \ldots & \mathbb{T}\mathbf{e}_m\end{bmatrix}\mathbf{b}$ is the same as $\mathbb{T}\mathbf{b}$ for an arbitrary vector $\mathbf{b}$.

$$\begin{aligned}\begin{bmatrix}\mathbb{T}\mathbf{e}_1 & \ldots & \mathbb{T}\mathbf{e}_m\end{bmatrix}\mathbf{b} &= \begin{bmatrix}\mathbb{T}\mathbf{e}_1 & \ldots & \mathbb{T}\mathbf{e}_m\end{bmatrix}(b_1\mathbf{e}_1 + \ldots + b_m\mathbf{e}_m) \\ &= b_1\begin{bmatrix}\mathbb{T}\mathbf{e}_1 & \ldots & \mathbb{T}\mathbf{e}_m\end{bmatrix}\mathbf{e}_1 + \ldots + b_m\begin{bmatrix}\mathbb{T}\mathbf{e}_1 & \ldots & \mathbb{T}\mathbf{e}_m\end{bmatrix}\mathbf{e}_m \\ &= b_1\mathbb{T}\mathbf{e}_1 + \ldots + b_m\mathbb{T}\mathbf{e}_m \\ &= \mathbb{T}(b_1\mathbf{e}_1 + \ldots + b_m\mathbf{e}_m) \\ &= \mathbb{T}\mathbf{b}\end{aligned}$$

Assume that $\mathbf{v}_1, \ldots, \mathbf{v}_m$ aren't linearly independent; in particular, and without loss of generality, assume $\mathbf{v}_m = c_1\mathbf{v}_1 + \ldots + c_{m-1}\mathbf{v}_{m-1}$ for some scalars $c_1, \ldots, c_{m-1}$. Then subtracting $\mathbf{v}_m$ from both sides, we see that

$$\mathbf{0} = c_1\mathbf{v}_1 + \ldots + c_{m-1}\mathbf{v}_{m-1} + (-1)\mathbf{v}_m$$

provides a linear combination of zero in which not all of the scalar coefficients are zeros.

Assume there exist scalars $b_1, \ldots, b_m$ for which $b_1\mathbf{v}_1 + \ldots + b_m\mathbf{v}_m = 0$ with $b_1 \neq 0$ (without loss of generality). Then the rearranged equation

$$\mathbf{v}_1 = \left(-\tfrac{b_2}{b_1}\right)\mathbf{v}_2 + \ldots + \left(-\tfrac{b_m}{b_1}\right)\mathbf{v}_m$$

shows that $\mathbf{v}_1$ is a linear combination of the other vectors, contradicting the assumption of linear independence.

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 1.10

Show that if $\begin{bmatrix}\mathbf{v}_1 & \cdots & \mathbf{v}_m\end{bmatrix}\mathbf{b} = \begin{bmatrix}\mathbf{v}_1 & \cdots & \mathbf{v}_m\end{bmatrix}\mathbf{c}$ then $\mathbf{c}$ must equal $\mathbf{b} + \mathbf{z}$ for some $\mathbf{z}$ in the null space of $\begin{bmatrix}\mathbf{v}_1 & \cdots & \mathbf{v}_m\end{bmatrix}$.

Exercise 1.11

Show that $\mathbf{v}_1, \ldots, \mathbf{v}_m$ are linearly independent if and only if $\mathbf{b} \neq \mathbf{c}$ implies $\begin{bmatrix}\mathbf{v}_1 & \cdots & \mathbf{v}_m\end{bmatrix}\mathbf{b} \neq \begin{bmatrix}\mathbf{v}_1 & \cdots & \mathbf{v}_m\end{bmatrix}\mathbf{c}$, that is, every vector in the span corresponds to a *unique* vector of scalar coefficients.

Exercise 1.12

Let $\{\mathbf{v}_1, \ldots, \mathbf{v}_m\}$ be a basis for $\mathcal{V}$. How do you know that $\{\mathbf{v}_2, \ldots, \mathbf{v}_m\}$ is not also a basis for $\mathcal{V}$?

Trivially $\mathbf{c} = \mathbf{b} + (\mathbf{c} - \mathbf{b})$; we show that the second term is in the null space.

$$\begin{aligned}\begin{bmatrix}\mathbf{v}_1 & \cdots & \mathbf{v}_m\end{bmatrix}(\mathbf{c} - \mathbf{b}) &= \begin{bmatrix}\mathbf{v}_1 & \cdots & \mathbf{v}_m\end{bmatrix}\mathbf{c} - \begin{bmatrix}\mathbf{v}_1 & \cdots & \mathbf{v}_m\end{bmatrix}\mathbf{b} \\ &= \mathbf{0}\end{aligned}$$

because the two linear combinations are equal.

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

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

By linear independence, no linear combination of $\mathbf{v}_2, \ldots, \mathbf{v}_m$ equals $\mathbf{v}_1 \in \mathcal{V}$. Therefore, $\mathbf{v}_2, \ldots, \mathbf{v}_m$ do not span $\mathcal{V}$ and thus do not satisfy the definition of a basis.

With $\mathcal{N}$ representing the null space of $\begin{bmatrix}\mathbf{v}_1 & \cdots & \mathbf{v}_m\end{bmatrix}$, the set of coefficient vectors producing $\begin{bmatrix}\mathbf{v}_1 & \cdots & \mathbf{v}_m\end{bmatrix}\mathbf{b}$ is exactly $\{\mathbf{b} + \mathbf{z} : \mathbf{z} \in \mathcal{N}\}$; this set has a single element ($\mathbf{b}$) if and only if the null space has only the zero vector. We've already established that this condition is equivalent to linear independence.

Exercise 1.13

Let $\{\mathbf{v}_1, \ldots, \mathbf{v}_m\}$ be a basis for $\mathcal{V}$, and let $\mathcal{S}$ be a *proper* subspace of $\mathcal{V}$. Explain why at least one of $\mathbf{v}_1, \ldots, \mathbf{v}_m$ is not in $\mathcal{S}$.

Exercise 1.14

Suppose $\{\mathbf{v}_1, \ldots, \mathbf{v}_m\}$ is a basis for $\mathcal{V}$. Prove that *every* basis for $\mathcal{V}$ has exactly m vectors.

Exercise 1.15

Let $\mathcal{V}$ be an m-dimensional vector space. Prove that any set of m linearly independent vectors in $\mathcal{V}$ must be basis for $\mathcal{V}$.

Exercise 1.16

Let $\mathcal{S}$ be a subspace of an n-dimensional vector space $\mathcal{V}$. Prove that a basis for $\mathcal{S}$ exists.

Suppose that $\{\mathbf{w}_1, \ldots, \mathbf{w}_k\}$ is also a basis for $\mathcal{V}$ and that $k < m$. The set $\{\mathbf{v}_2, \ldots, \mathbf{v}_m\}$ doesn't span $\mathcal{V}$, so there must be some vector in $\{\mathbf{w}_1, \ldots, \mathbf{w}_k\}$ that isn't in the span of $\{\mathbf{v}_2, \ldots, \mathbf{v}_m\}$. Assume without loss of generality that $\mathbf{w}_1$ isn't in the span of $\{\mathbf{v}_2, \ldots, \mathbf{v}_m\}$. Then $\{\mathbf{w}_1, \mathbf{v}_2, \ldots, \mathbf{v}_m\}$ is linearly independent. By repeating this logic with $\mathbf{v}_2$ and continuing through $\mathbf{v}_k$, we end up claiming that $\{\mathbf{w}_1, \ldots, \mathbf{w}_k, \mathbf{v}_{k+1}, \ldots, \mathbf{v}_m\}$ is a linearly independent set. However, since $\{\mathbf{w}_1, \ldots, \mathbf{w}_k\}$ is a basis for $\mathcal{V}$, the remaining vectors $\{\mathbf{v}_{k+1}, \ldots, \mathbf{v}_m\} \subset \mathcal{V}$ have to be in their span which is a contradiction. This proves that one basis for $\mathcal{V}$ can't be larger than another.

Any $\mathbf{w}$ in $\mathcal{V}$ has a unique representation as $b_1\mathbf{v}_1 + \ldots b_m\mathbf{v}_m$ for some scalar coefficients. Assume every one of the basis vectors were in $\mathcal{S}$. Because $\mathcal{S}$ is a subspace it contains all linear combinations of its vectors which would include $\mathbf{w}$. Therefore, if $\mathcal{S}$ is a subspace of $\mathcal{V}$ that includes an entire basis for $\mathcal{V}$, then it would have to be equal to $\mathcal{V}$.

We'll describe a constructive proof. Take any vector of $\mathcal{S}$ and call it $\mathbf{v}_1$. If $\mathcal{S}$ equals the span of $\mathbf{v}_1$, then $\mathbf{v}_1$ is a basis. Otherwise, take a vector from $\mathcal{S}$ that is outside of the span of $\mathbf{v}_1$ and call it $\mathbf{v}_2$. If $\mathcal{S}$ equals the span of $\{\mathbf{v}_1, \mathbf{v}_2\}$, then they form a basis. Otherwise, continue to repeat this process of adding one more linearly independent vector at a time until the vectors span $\mathcal{S}$. The algorithm is guaranteed to terminate with no more than n vectors, because any set of n linearly independent vectors is a basis for $\mathcal{V}$ according to Exercise 1.15.

Let $V := \{\mathbf{v}_1, \ldots, \mathbf{v}_m\}$ be a set of linearly independent vectors in $\mathcal{V}$, and let $W := \{\mathbf{w}_1, \ldots, \mathbf{w}_m\}$ be a basis for $\mathcal{V}$. Assume that the span of V is not $\mathcal{V}$. From Exercise 1.13 there exists at least one vector in W that isn't in the span of V; without loss of generality, let it be $\mathbf{w}_1$. By continuing to add one vector at a time from V to W in this way, you would maintain linear independence and you would eventually produce a set that does have $\mathcal{V}$ as its span. But that set will have more than m vectors, which contradicts the fact from Exercise 1.14 that every basis for $\mathcal{V}$ has m vectors.

Exercise 1.17

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

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 $\mathbb{T}\mathbf{v}$.

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.

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.

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 $\mathbb{T}$.

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.

Exercise 1.21

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

Exercise 1.22

Let $\mathbb{T}^{-1}$ be the inverse of a linear operator $\mathbb{T}$, that is, $\mathbb{T}^{-1}\mathbb{T}\mathbf{v} = \mathbf{v}$ for every $\mathbf{v}$ in the domain of $\mathbb{T}$. Show that $\mathbb{T}$ is also the inverse of $\mathbb{T}^{-1}$, that is, $\mathbb{T}\mathbb{T}^{-1}\mathbf{w} = \mathbf{w}$ for every $\mathbf{w}$ in the domain of $\mathbb{T}^{-1}$ (which we've defined to be the range of $\mathbb{T}$).

Exercise 1.23

Suppose a linear operator $\mathbb{T}$ has an inverse $\mathbb{T}^{-1}$. Show that $\mathbb{T}^{-1}$ is also linear:

$$\mathbb{T}^{-1}(a_1\mathbf{w}_1 + a_2\mathbf{w}_2) = a_1\mathbb{T}^{-1}\mathbf{w}_1 + a_2\mathbb{T}^{-1}\mathbf{w}_2$$

for all vectors $\mathbf{w}_1, \mathbf{w}_2$ and scalars a_1, a_2.

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.

We know that $\mathbf{w} = \mathbb{T}\mathbf{v}$ for some $\mathbf{v}$; by making this substitution,

$$\begin{aligned}\mathbb{T}\mathbb{T}^{-1}\mathbf{w} &= \mathbb{T}\mathbb{T}^{-1}\mathbb{T}\mathbf{v} \\ &= \mathbb{T}\mathbf{v} \\ &= \mathbf{w}.\end{aligned}$$

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 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 $\mathbf{q}_1, \ldots, \mathbf{q}_n$.

As an inverse function, $\mathbb{T}^{-1}$ maps from $\mathbb{T}$'s range to its domain. $\mathbf{w}_1$ and $\mathbf{w}_2$ are in the range of $\mathbb{T}$, so we know that they can be represented by $\mathbf{w}_1 = \mathbb{T}\mathbf{v}_1$ and $\mathbf{w}_2 = \mathbb{T}\mathbf{v}_2$ for some vectors $\mathbf{v}_1$ and $\mathbf{v}_2$.

$$\begin{aligned}\mathbb{T}^{-1}(a_1\mathbf{w}_1 + a_2\mathbf{w}_2) &= \mathbb{T}^{-1}(a_1\mathbb{T}\mathbf{v}_1 + a_2\mathbb{T}\mathbf{v}_2) \\ &= \mathbb{T}^{-1}\mathbb{T}(a_1\mathbf{v}_1 + a_2\mathbf{v}_2) \\ &= a_1\mathbf{v}_1 + a_2\mathbf{v}_2 \\ &= a_1\mathbb{T}^{-1}\mathbf{w}_1 + a_2\mathbb{T}^{-1}\mathbf{w}_2\end{aligned}$$

Exercise 1.25

Show that inner products are also additive in their second argument:

$$\langle \mathbf{v}, \mathbf{w} + \mathbf{x} \rangle = \langle \mathbf{v}, \mathbf{w} \rangle + \langle \mathbf{v}, \mathbf{x} \rangle.$$

Exercise 1.26

Show that inner products are also homogeneous in their second argument when the scalar is real: for every $a \in \mathbb{R}$,

$$\langle \mathbf{v}, a\mathbf{w} \rangle = a\langle \mathbf{v}, \mathbf{w} \rangle.$$

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 1.28

Show that $\|a\mathbf{v}\| = |a|\|\mathbf{v}\|$ for any scalar a.

The complex conjugate of a product is equal to the product of the complex conjugates, and the complex conjugate of a real number is itself, so

$$\begin{aligned}\langle \mathbf{v}, a\mathbf{w}\rangle &= \overline{\langle a\mathbf{w}, \mathbf{v}\rangle} \\ &= \overline{a\langle \mathbf{w}, \mathbf{v}\rangle} \\ &= \overline{a}\overline{\langle \mathbf{w}, \mathbf{v}\rangle} \\ &= a\langle \mathbf{v}, \mathbf{w}, \rangle.\end{aligned}$$

The complex conjugate of a sum is equal to the sum of the complex conjugates, so

$$\begin{aligned}\langle \mathbf{v}, \mathbf{w}+\mathbf{x}\rangle &= \overline{\langle \mathbf{w}+\mathbf{x}, \mathbf{v}\rangle} \\ &= \overline{\langle \mathbf{w}, \mathbf{v}\rangle} + \overline{\langle \mathbf{x}, \mathbf{v}\rangle} \\ &= \langle \mathbf{v}, \mathbf{w}\rangle + \langle \mathbf{v}, \mathbf{x}\rangle.\end{aligned}$$

From steps in our solution to Exercise 1.26, we can realize that $\langle \mathbf{x}, a\mathbf{y}\rangle = \overline{a}\langle \mathbf{x}, \mathbf{y}\rangle$. Using the definition of norm,

$$\begin{aligned}\|a\mathbf{v}\| &= \sqrt{\langle a\mathbf{v}, a\mathbf{v}\rangle} \\ &= \sqrt{a\overline{a}}\sqrt{\langle \mathbf{v}, \mathbf{v}\rangle} \\ &= |a|\|\mathbf{v}\|.\end{aligned}$$

Note that with $a = b + ic$, the squared absolute value $|a|^2$ is defined to be $b^2 + c^2$.

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.

Exercise 1.29

Suppose $\mathbf{v}_1, \ldots, \mathbf{v}_m$ are orthogonal to each other and none of them is the zero vector. Show that they must be linearly independent.

Exercise 1.30

Suppose $\langle \mathbf{x}, \mathbf{v} \rangle = \langle \mathbf{y}, \mathbf{v} \rangle$ for all $\mathbf{v}$. Show that $\mathbf{x}$ and $\mathbf{y}$ must be the same vector.

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 $\frac{1}{\|\mathbf{v}\|}\mathbf{v}$.

Subtracting, $\langle \mathbf{y}, \mathbf{v} \rangle$ from both sides of the assumption,

$$\begin{aligned} 0 &= \langle \mathbf{x}, \mathbf{v} \rangle - \langle \mathbf{y}, \mathbf{v} \rangle \\ &= \langle \mathbf{x} - \mathbf{y}, \mathbf{v} \rangle \end{aligned}$$

for all $\mathbf{v}$. In particular, substitute $\mathbf{x} - \mathbf{y}$ for $\mathbf{v}$ to see that $\|\mathbf{x} - \mathbf{y}\|^2 = 0$ which implies that $\mathbf{x} - \mathbf{y} = \mathbf{0}$, i.e. $\mathbf{x} = \mathbf{y}$.

Without loss of generality, we will consider whether or not $\mathbf{v}_m$ can be equal to some linear combination $b_1\mathbf{v}_1 + \ldots + b_{m-1}\mathbf{v}_{m-1}$. The inner product of this linear combination with $\mathbf{v}_m$ equals

$$\begin{aligned} \langle b_1\mathbf{v}_1 + \ldots + b_{m-1}\mathbf{v}_{m-1}, \mathbf{v}_m \rangle &= b_1 \underbrace{\langle \mathbf{v}_1, \mathbf{v}_m \rangle}_{0} + \ldots + b_{m-1} \underbrace{\langle \mathbf{v}_{m-1}, \mathbf{v}_m \rangle}_{0} \\ &= 0. \end{aligned}$$

But the inner product of $\mathbf{v}_m$ with itself is equal to its squared length, which must be greater than zero by the assumption that $\mathbf{v}_m$ isn't the zero vector. Therefore, no such linear combination can be equal to $\mathbf{v}_m$.

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 $\mathbf{y}$ 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 1.45

Let $\mathcal{S}$ be an m-dimensional subspace of a d-dimensional vector space $\mathcal{V}$. Verify that the dimension of $\mathcal{S}^{\perp}$ is $d - m$.

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.

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.

Let B be a basis for $\mathcal{S}$; it contains m vectors. Suppose there exist more than $d-m$ linearly independent vectors in $\mathcal{S}^\perp$. All of them are also linearly independent of B, so the two bases taken together would contain a total of *more than* d linearly independent vectors in $\mathcal{V}$ which is impossible. On the other hand, suppose $\mathcal{S}^\perp$ has a basis of fewer than $d-m$ vectors. Then that basis, taken together with B would have fewer than d vectors, so it wouldn't span $\mathcal{V}$. However, this can't be true because we've seen than *any* vector in $\mathcal{V}$ can be represented as the sum of a vector in $\mathcal{S}$ and a vector in $\mathcal{S}^\perp$.

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.

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.

Exercise 1.49

Let $\mathbb{H}_1$ be the orthogonal projection operator onto $\mathcal{S}_1$, and let $\mathbb{H}_0$ be the orthogonal projection operator onto $\mathcal{S}_0 \subseteq \mathcal{S}_1$. Explain why $\mathbb{H}_1 - \mathbb{H}_0$ is the orthogonal projection operator onto the orthogonal complement of $\mathcal{S}_0$ within $\mathcal{S}_1$.

Exercise 1.50

Let $\mathcal{S}_0 \subseteq \mathcal{S}_1$ be subspaces, and let $\mathbb{H}_0$ and $\mathbb{H}_1$ be orthogonal projection operators onto $\mathcal{S}_0$ and $\mathcal{S}_1$ respectively. Explain why $\mathbb{H}_0 \circ \mathbb{H}_1 = \mathbb{H}_1 \circ \mathbb{H}_0 = \mathbb{H}_0$.

Exercise 1.51

Let $\mathbb{M} \in \mathcal{F}^{n \times m}$ and $\mathbb{L} \in \mathcal{F}^{m \times n}$. Show that the trace of $\mathbb{ML}$ is equal to the trace of $\mathbb{LM}$.

Exercise 1.52

Let $\mathbb{M}$ and $\mathbb{L}$ be matrices such that the product $\mathbb{ML}$ is well-defined. Show that $(\mathbb{ML})^T = \mathbb{L}^T\mathbb{M}^T$.

In our discussion regarding Equation 1.4, we realized that the orthogonal projection of $\mathbf{y}$ onto $\mathcal{S}_0$ is the same as the orthogonal projection of $\mathbb{H}_1\mathbf{y}$ onto $\mathcal{S}_0$. In other words, it doesn't matter which order you compose the operators, you end up at $\mathbb{H}_0\mathbf{y}$ either way.

The operator $\mathbb{H}_1 - \mathbb{H}_0$ evaluated at $\mathbf{y}$ has the value $\mathbb{H}_1\mathbf{y} - \mathbb{H}_0\mathbf{y}$. From our discussion regarding Equation 1.4, we know that this is precisely the orthogonal projection of $\mathbf{y}$ onto the orthogonal complement of $\mathcal{S}_0$ within $\mathcal{S}_1$.

The (i, j) entry of $(\mathbb{ML})^T$ is the (j, i) entry of $\mathbb{ML}$, which is $\sum_k \mathbb{M}_{j,k}\mathbb{L}_{k,i}$. The (i, j) entry of $\mathbb{L}^T\mathbb{M}^T$ works out to be the same:

$$\sum_k (\mathbb{L}^T)_{i,k}(\mathbb{M}^T)_{k,j} = \sum_k \mathbb{L}_{k,i}\mathbb{M}_{j,k}.$$

The ith diagonal entry of $\mathbb{ML}$ is $\sum_{j=1}^m \mathbb{M}_{i,j}\mathbb{L}_{j,i}$. We express the trace as the sum of these diagonals then reverse the order of the summations.

$$\begin{aligned}
\operatorname{tr}\mathbb{ML} &= \sum_{i=1}^{n}\sum_{j=1}^{m} \mathbb{M}_{i,j}\mathbb{L}_{j,i} \\
&= \sum_{j=1}^{m}\sum_{i=1}^{n} \mathbb{L}_{j,i}\mathbb{M}_{i,j} \\
&= \operatorname{tr}\mathbb{LM}
\end{aligned}$$

Exercise 1.53

If the columns of $\mathbb{U}$ are orthonormal, show that $\mathbb{U}^T\mathbb{U}$ equals the identity matrix $\mathbb{I}$.

Exercise 1.54

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

Exercise 1.55

Show that a square matrix is invertible if and only if its columns are linearly independent.

Exercise 1.56

Suppose $\mathbb{M} \in \mathbb{R}^{n \times n}$ has orthonormal eigenvectors $\mathbf{q}_1, \ldots, \mathbf{q}_n$ with eigenvalues $\lambda_1, \ldots, \lambda_n$ Show that $\mathbb{M}$ can't have any other eigenvalues.

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

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

Letting $\mathbf{u}_1, \ldots, \mathbf{u}_m$ denote the columns,

$$\begin{aligned}\mathbb{U}^T\mathbb{U} &= \begin{bmatrix} - & \mathbf{u}_1 & - \\ & \vdots & \\ - & \mathbf{u}_m & - \end{bmatrix} \begin{bmatrix} | & & | \\ \mathbf{u}_1 & \cdots & \mathbf{u}_m \\ | & & | \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{u}_1^T\mathbf{u}_1 & \cdots & \mathbf{u}_1^T\mathbf{u}_n \\ \vdots & \ddots & \vdots \\ \mathbf{u}_n^T\mathbf{u}_1 & \cdots & \mathbf{u}_n^T\mathbf{u}_n \end{bmatrix} \\ &= \begin{bmatrix} 1 & & \\ & \ddots & \\ & & 1 \end{bmatrix}\end{aligned}$$

with every off-diagonal entry equal to zero.

Let us check what would be required for an arbitrary vector $\mathbf{w}$ to be an eigenvector for $\mathbb{M}$. We can express $\mathbf{w}$ with respect to the eigenvector basis:

$$\begin{aligned}\mathbb{M}\mathbf{w} &= \mathbb{M}(\langle \mathbf{w}, \mathbf{q}_1\rangle \mathbf{q}_1 + \ldots + \langle \mathbf{w}, \mathbf{q}_n\rangle \mathbf{q}_n) \\ &= \langle \mathbf{w}, \mathbf{q}_1\rangle \mathbb{M}\mathbf{q}_1 + \ldots + \langle \mathbf{w}, \mathbf{q}_n\rangle \mathbb{M}\mathbf{q}_n \\ &= \langle \mathbf{w}, \mathbf{q}_1\rangle \lambda_1\mathbf{q}_1 + \ldots + \langle \mathbf{w}, \mathbf{q}_n\rangle \lambda_n\mathbf{q}_n.\end{aligned}$$

This is only proportional to $\mathbf{w} = \langle \mathbf{w}, \mathbf{q}_1\rangle \mathbf{q}_1 + \ldots + \langle \mathbf{w}, \mathbf{q}_n\rangle \mathbf{q}_n$ if all of the non-zero terms share the same eigenvalue coefficient. That coefficient, which is one of $\lambda_1, \ldots, \lambda_n$, would be the eigenvalue of $\mathbf{w}$.

Let $\mathbb{M}$ be a square matrix. First, assume it's invertible. Then what linear combinations satisfy $\mathbb{M}\mathbf{b} = \mathbf{0}$? Multiplying both sides by the inverse, we see that the coefficients $\mathbf{b} = \mathbb{M}^{-1}\mathbf{0}$ must be the zero vector.

Next, suppose the n columns of $\mathbb{M}$ are linearly independent. Then for each canonical basis vector $\mathbf{e}_j$, there is some coefficient vector $\mathbf{b}_j$ such that $\mathbb{M}\mathbf{b}_j = \mathbf{e}_j$. The matrix with these coefficient vectors as its columns is the inverse of $\mathbb{M}$.

$$\begin{aligned}\mathbb{M}\begin{bmatrix} | & & | \\ \mathbf{b}_1 & \cdots & \mathbf{b}_n \\ | & & | \end{bmatrix} &= \begin{bmatrix} | & & | \\ \mathbb{M}\mathbf{b}_1 & \cdots & \mathbb{M}\mathbf{b}_n \\ | & & | \end{bmatrix} \\ &= \underbrace{\begin{bmatrix} | & & | \\ \mathbf{e}_1 & \cdots & \mathbf{e}_n \\ | & & | \end{bmatrix}}_{\mathbb{I}_n}\end{aligned}$$

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 1.63

Explain why $\mathbb{M}^T\mathbb{M}$ is invertible if and only if the columns of $\mathbb{M} \in \mathbb{R}^{n \times m}$ are linearly independent.

Exercise 1.64

Let $\mathbb{M} \in \mathbb{R}^{n \times n}$. Prove that $\mathbb{M}$ is symmetric if and only if $\langle \mathbf{v}, \mathbb{M}\mathbf{w} \rangle = \langle \mathbb{M}\mathbf{v}, \mathbf{w} \rangle$ for every $\mathbf{v}, \mathbf{w} \in \mathbb{R}^n$.

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.

First, assume that $\mathbb{M}$ is symmetric. For an arbitrary $\mathbf{v}, \mathbf{w} \in \mathbb{R}^n$, we can use a spectral decomposition of $\mathbb{M}$ to see that

$$\begin{aligned}\langle \mathbf{v}, \mathbb{M}\mathbf{w}\rangle &= \mathbf{v}^T\mathbb{M}\mathbf{w} \\ &= \mathbf{v}^T(\lambda_1\mathbf{q}_1\mathbf{q}_1^T + \ldots + \lambda_n\mathbf{q}_n\mathbf{q}_n^T)\mathbf{w} \\ &= \lambda_1(\mathbf{v}^T\mathbf{q}_1)(\mathbf{q}_1^T\mathbf{w}) + \ldots + \lambda_n(\mathbf{v}^T\mathbf{q}_n)(\mathbf{q}_n^T\mathbf{w}) \\ &= \lambda_1(\mathbf{w}^T\mathbf{q}_1)(\mathbf{q}_1^T\mathbf{v}) + \ldots + \lambda_n(\mathbf{w}^T\mathbf{q}_n)(\mathbf{q}_n^T\mathbf{v}) \\ &= \mathbf{w}^T(\lambda_1\mathbf{q}_1\mathbf{q}_1^T + \ldots + \lambda_n\mathbf{q}_n\mathbf{q}_n^T)\mathbf{v} \\ &= \mathbf{w}^T\mathbb{M}\mathbf{v} \\ &= \langle \mathbf{w}, \mathbb{M}\mathbf{v}\rangle.\end{aligned}$$

Next, suppose $\langle \mathbf{v}, \mathbb{M}\mathbf{w}\rangle = \langle \mathbb{M}\mathbf{v}, \mathbf{w}\rangle$ for every $\mathbf{v}, \mathbf{w} \in \mathbb{R}^n$. In particular, apply two canonical basis vectors $\mathbf{e}_i$ and $\mathbf{e}_j$. The vector $\mathbb{M}\mathbf{e}_j$ is the jth column of $\mathbb{M}$, so $\langle \mathbf{e}_i, \mathbb{M}\mathbf{e}_j\rangle$ is the (i,j)-entry of $\mathbb{M}$. By our assumption, it is equal to $\langle \mathbf{e}_j, \mathbb{M}\mathbf{e}_i\rangle$ which is the (j,i)-entry of $\mathbb{M}$ which shows that $\mathbb{M}$ must be symmetric.

The columns of $\mathbb{M}$ are linearly independent if and only if $C(\mathbb{M})$ is m-dimensional; that's what we'll check.

First, suppose $m \leq n$, and let $\mathbb{V}$ have the singular value decomposition

$$\mathbb{V} = \sigma_1\mathbf{u}_1\mathbf{v}_1^T + \ldots + \sigma_m\mathbf{u}_m\mathbf{v}_m^T.$$

Then $\mathbb{V}^T\mathbb{V}$ has the spectral decomposition

$$\mathbb{V}^T\mathbb{V} = \sigma_1^2\mathbf{v}_1\mathbf{v}_1^T + \ldots + \sigma_m^2\mathbf{v}_m\mathbf{v}_m^T.$$

We know that $\mathbb{V}^T\mathbb{V}$ is invertible if and only if its eigenvalues $\sigma_1^2, \ldots, \sigma_m^2$ are positive. The column space of $\mathbb{V}$ is a linear combination of all of $\mathbf{u}_1, \ldots, \mathbf{u}_m$ if and only if none of the singular values $\sigma_1, \ldots, \sigma_m$ are zero. These conditions are the same.

Otherwise, if $n < m$, then the singular decomposition can't possibly represent a linear combination of m column vectors, so $\mathbb{M}$ can't have linearly independent. Similarly, the squared singular values can't account for the m positive eigenvalues that $\mathbb{V}^T\mathbb{V}$ would need to be invertible.

Exercise 1.65

Let $\mathbb{H}$ be a real matrix. Use Exercise 1.64 to show that if $\mathbb{H}$ is an orthogonal projection matrix then it must be symmetric.

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.

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 $\mathbb{M}$.

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 $\mathbf{u}$.

$$\begin{aligned}\langle \mathbb{H}\mathbf{v}, \mathbf{w} \rangle &= \langle \mathbb{H}\mathbf{v}, \mathbb{H}\mathbf{w} + (\mathbf{w} - \mathbb{H}\mathbf{w}) \rangle \\ &= \langle \mathbb{H}\mathbf{v}, \mathbb{H}\mathbf{w} \rangle + \langle \mathbb{H}\mathbf{v}, \mathbf{w} - \mathbb{H}\mathbf{w} \rangle\end{aligned}$$

Because $\mathbb{H}\mathbf{w}$ is in the space that $\mathbb{H}$ projects onto while $\mathbf{w} - \mathbb{H}\mathbf{w}$ is orthogonal to it, the second term is zero. Similarly,

$$\begin{aligned}\langle \mathbf{v}, \mathbb{H}\mathbf{w} \rangle &= \langle \mathbb{H}\mathbf{v} + (\mathbf{v} - \mathbb{H}\mathbf{v}), \mathbb{H}\mathbf{w} \rangle \\ &= \langle \mathbb{H}\mathbf{v}, \mathbb{H}\mathbf{w} \rangle + \underbrace{\langle \mathbf{v} - \mathbb{H}\mathbf{v}, \mathbb{H}\mathbf{w} \rangle}_{0}.\end{aligned}$$

Both $\langle \mathbf{v}, \mathbb{H}\mathbf{w} \rangle$ and $\langle \mathbb{H}\mathbf{v}, \mathbf{w} \rangle$ simplify to $\langle \mathbb{H}\mathbf{v}, \mathbb{H}\mathbf{w} \rangle$, so they are equal to each other.

It's also clear from the formula derived in Exercise 1.71 that orthogonal projection matrices are symmetric. However, I prefer the argument used here because it's more readily extended to orthogonal projection *operators*.

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 $C(\mathbb{M})$.

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.

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 $C(\mathbb{M})$.

Exercise 1.72

Show that $\mathbb{M}^{-}\mathbb{M}$ equals the orthogonal projection matrix onto the row space of $\mathbb{M}$.

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

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.

Let $\mathbb{M} = \mathbb{U}\mathbb{S}\mathbb{V}^T$ be a singular value decomposition for which $\mathbb{S}$ is square and has only positive values along its diagonal.

$$\begin{aligned}\mathbb{M}^-\mathbb{M} &= \mathbb{V}\mathbb{S}^{-1}\mathbb{U}^T\mathbb{U}\mathbb{S}\mathbb{V}^T \\ &= \mathbb{V}\mathbb{V}^T\end{aligned}$$

Exercise 1.71 indicates that the orthogonal projection matrix onto $C(\mathbb{V})$ is $\mathbb{V}(\mathbb{V}^T\mathbb{V})^{-1}\mathbb{V}^T$ which simplifies to $\mathbb{V}\mathbb{V}^T$ because $\mathbb{V}^T\mathbb{V}$ is the identity.

It only remains to establish that $C(\mathbb{V})$ is the row space of $\mathbb{M}$. Letting the entries of $\mathbf{w}$ be the coefficients of the linear combination,

$$\begin{aligned}\mathbf{w}^T\mathbb{M} &= \mathbf{w}^T(\sigma_1\mathbf{u}_1\mathbf{v}_1^T + \ldots \sigma_d\mathbf{u}_d\mathbf{v}_d^T) \\ &= \sigma_1\langle\mathbf{w}, \mathbf{u}_1\rangle\mathbf{v}_1^T + \ldots + \sigma_d\langle\mathbf{w}, \mathbf{u}_d\rangle\mathbf{v}_d^T.\end{aligned}$$

Any linear combination of $\mathbf{v}_1^T, \ldots, \mathbf{v}_d^T$ can be produced by the appropriate choice of $\mathbf{w}$, but no vector outside of their span can be produced.

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 $\mathbb{M}(\mathbb{M}^T\mathbb{M})^{-1}\mathbb{M}^T$.

Exercise 1.73

Explain why the Moore-Penrose inverse of an invertible matrix must be its inverse.

Exercise 1.74

Of all coefficient vectors that satisfy the Normal equation, show that $\hat{\mathbf{b}} = (\mathbb{M}^T\mathbb{M})^{-}\mathbb{M}^T\hat{\mathbf{y}}$ has the smallest norm.

Exercise 1.75

Show that $\mathbb{M}\mathbb{M}^{-}\mathbb{M} = \mathbb{M}$.

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 $\mathbb{M} \in \mathbb{R}^{n\times n}$.

From Section 1.5, we know that every solution can be represented as $\hat{\mathbf{b}} + \mathbf{w}$ for some $\mathbf{w}$ in the null space of $\mathbb{M}^T\mathbb{M}$. Let $\mathbb{M}^T\mathbb{M}$ have spectral decomposition

$$\mathbb{M}^T\mathbb{M} = \sigma_1^2\mathbf{v}_1\mathbf{v}_1^T + \ldots + \sigma_d^2\mathbf{v}_d\mathbf{v}_d^T + (0)\mathbf{v}_{d+1}\mathbf{v}_{d+1}^T + \ldots + (0)\mathbf{v}_m\mathbf{v}_m^T.$$

with positive $\sigma_1^2, \ldots, \sigma_d^2$. It's clear that the null space is exactly the span of $\{\mathbf{v}_{d+1}, \ldots, \mathbf{v}_m\}$. On the other hand, by definition of generalized inverse, $\hat{\mathbf{b}}$ is in the span of $\{\mathbf{v}_1, \ldots, \mathbf{v}_d\}$. The squared norm of any solution $\hat{\mathbf{b}} + \mathbf{w}$ is $\|\hat{\mathbf{b}}\|^2 + \|\mathbf{w}\|^2$, so the solution of smallest norm is $\hat{\mathbf{b}}$.

If $\mathbb{M} \in \mathbb{R}^{n\times n}$ is invertible, then its row space is $\mathbb{R}^n$. Exercise 1.72 implies that $\mathbb{M}^-\mathbb{M}\mathbf{w} = \mathbf{w}$ for every $\mathbf{w} \in \mathbb{R}^n$. $\mathbb{M}^-$ must be the inverse according to Exercise 1.22.

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.

Let $\mathbb{M}$ have singular value decomposition $\mathbb{U}\mathbb{S}\mathbb{V}^T$ where $\mathbb{S}$ is a square matrix with strictly positive diagonals. Then

$$\begin{aligned}
\mathbb{M}\mathbb{M}^-\mathbb{M} &= \mathbb{U}\mathbb{S}\underbrace{\mathbb{V}^T\mathbb{V}}_{\mathbb{I}}\mathbb{S}^{-1}\underbrace{\mathbb{U}^T\mathbb{U}}_{\mathbb{I}}\mathbb{S}\mathbb{V}^T \\
&= \mathbb{U}\mathbb{S}\mathbb{V}^T \\
&= \mathbb{M}.
\end{aligned}$$

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.

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 1.54 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.

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.

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

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

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 1.81

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 $\frac{1}{n}\mathbb{X}^T\mathbb{X}$.

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 $\mathbf{1} = (1, \ldots, 1)$.

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 $\frac{1}{n}\mathbf{x}_i\mathbf{x}_i^T$.

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.

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.

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.

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.

Exercise 2.5

Show that the correlation between two vectors equals the empirical covariance of their standardized versions.

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.

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,

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

$$\begin{aligned}
\rho_{\mathbf{x},\mathbf{y}} &:= \frac{\sigma_{\mathbf{x},\mathbf{y}}}{\sigma_{\mathbf{x}}\sigma_{\mathbf{y}}} \\
&= \frac{(1/n)\langle \mathbf{x} - \bar{x}\mathbf{1}, \mathbf{y} - \bar{y}\mathbf{1}\rangle}{(\sqrt{1/n}\|\mathbf{x} - \bar{x}\mathbf{1}\|)(\sqrt{1/n}\|\mathbf{y} - \bar{y}\mathbf{1}\|)} \\
&= \frac{\langle \mathbf{x} - \bar{x}\mathbf{1}, \mathbf{y} - \bar{y}\mathbf{1}\rangle}{\|\mathbf{x} - \bar{x}\mathbf{1}\|\|\mathbf{y} - \bar{y}\mathbf{1}\|}.
\end{aligned}$$

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

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 3.1

Explain why the sum of the probabilities of E_1 and E_2 is no greater than the probability of their union.

Exercise 3.2

Let $\mathbf{X}$ be a random vector mapping to the complex plane with the representation $\mathbf{X} = Y + iZ$ where Y and Z are random variables. Verify that $\mathbb{E}Y + i\mathbb{E}Z$ is the expectation of $\mathbf{X}$ by checking property (ii), assuming that property (iii) holds for *random variables*.

Exercise 3.3

Suppose $\mathbf{x}$ satisfies $\mathbb{E}\langle \mathbf{X}, \mathbf{v} \rangle = \langle \mathbf{x}, \mathbf{v} \rangle$ and $\mathbf{y}$ satisfies $\mathbb{E}\langle \mathbf{Y}, \mathbf{v} \rangle = \langle \mathbf{y}, \mathbf{v} \rangle$ for all $\mathbf{v}$. In order to justify an implicit claim in our definition of expectation for random vectors, verify that

$$\mathbb{E}\langle \mathbf{X} + \mathbf{Y}, \mathbf{v} \rangle = \langle \mathbf{x} + \mathbf{y}, \mathbf{v} \rangle$$

for all $\mathbf{v}$. In other words, verify that the expectation of a sum is indeed the sum of the expectations when all expectations are defined by property (ii), assuming that property (iii) holds for *random variables*.

Exercise 3.4

For a random vector $\mathbf{X}$ and scalar a, show that $\mathbb{E}a\mathbf{X} = a\mathbb{E}\mathbf{X}$.

Let $\mathbf{X} = Y + iZ$ with random variables Y and Z, and let $a = b + ic$ be a complex number.

$$\begin{aligned}
\mathbb{E}\langle \mathbf{X}, a\rangle &= \mathbb{E}\langle Y + iZ, b + ic\rangle \\
&= \mathbb{E}[\langle Y, ic\rangle + \langle Y, b\rangle + \langle iZ, b\rangle + \langle iZ, ic\rangle] \\
&= \mathbb{E}[(-i)\langle Y, c\rangle + \langle Y, b\rangle + i\langle Z, b\rangle + \underbrace{(i)(-i)}_{-i^2=1}\langle Z, c\rangle] \\
&= \mathbb{E}(\langle Y, b\rangle + \langle Z, c\rangle) + i\mathbb{E}(\langle Z, b\rangle - \langle Y, c\rangle) \\
&= \langle \mathbb{E}Y, b\rangle + \underbrace{\langle \mathbb{E}Z, c\rangle}_{\langle i\mathbb{E}Z, ic\rangle} + i\langle \mathbb{E}Z, b\rangle - \underbrace{i\langle \mathbb{E}Y, c\rangle}_{-\langle \mathbb{E}Y, ic\rangle} \\
&= \langle \mathbb{E}Y + i\mathbb{E}Z, b\rangle + \langle \mathbb{E}Y + i\mathbb{E}Z, ic\rangle \\
&= \langle \mathbb{E}Y + i\mathbb{E}Z, a\rangle
\end{aligned}$$

The union of E_1 and E_2 is the same as the union of the disjoint sets E_1 and E_2/E_1 (the part of E_2 that isn't in E_1). With $\mathbb{P}$ mapping each event to its probability,

$$\begin{aligned}
\mathbb{P}(E_1 \cup E_2) &= \mathbb{P}[E_1 \cup (E_2/E_1)] \\
&= \mathbb{P}E_1 + \mathbb{P}(E_2/E_1) \\
&\leq \mathbb{P}E_1 + \mathbb{P}E_2
\end{aligned}$$

To understand the last step, realize that E_2 can be represented as the disjoint union $E_2 = (E_2/E_1) \cup (E_2 \cap E_1)$.

If $\mathbf{X}$ is a random vector taking values in the complex plane, then inner product is the ordinary product, so property (ii) of our definition of expectation says that $\mathbb{E}a\mathbf{X} = a\mathbb{E}\mathbf{X}$.

Now we'll use that fact to establish the more general result for random vectors.

$$\begin{aligned}
\mathbb{E}\langle a\mathbf{X}, \mathbf{v}\rangle &= \mathbb{E}a\langle \mathbf{X}, \mathbf{v}\rangle \\
&= a\mathbb{E}\langle \mathbf{X}, \mathbf{v}\rangle \\
&= a\langle \mathbb{E}\mathbf{X}, \mathbf{v}\rangle \\
&= \langle a\mathbb{E}\mathbf{X}, \mathbf{v}\rangle
\end{aligned}$$

First, let's verify the claim in question when $\mathbf{X}$ and $\mathbf{Y}$ map to the complex plane; we'll represent them by $X_1 + iX_2$ and $Y_1 + iY_2$ respectively. Based on Exercise 3.2,

$$\begin{aligned}
\mathbb{E}(\mathbf{X} + \mathbf{Y}) &= \mathbb{E}(X_1 + iX_2 + Y_1 + iY_2) \\
&= \mathbb{E}[(X_1 + Y_1) + i(X_2 + Y_2)] \\
&= \mathbb{E}(X_1 + Y_1) + i\mathbb{E}(X_2 + Y_2) \\
&= \mathbb{E}X_1 + \mathbb{E}Y_1 + i\mathbb{E}X_2 + i\mathbb{E}Y_2 \\
&= \underbrace{(\mathbb{E}X_1 + i\mathbb{E}X_2)}_{\mathbb{E}\mathbf{X}} + \underbrace{(\mathbb{E}Y_1 + i\mathbb{E}Y_2)}_{\mathbb{E}\mathbf{Y}}
\end{aligned}$$

Using this result, we can prove the general case.

$$\begin{aligned}
\mathbb{E}\langle \mathbf{X} + \mathbf{Y}, \mathbf{v}\rangle &= \mathbb{E}[\langle \mathbf{X}, \mathbf{v}\rangle + \langle \mathbf{Y}, \mathbf{v}\rangle] \\
&= \mathbb{E}\langle \mathbf{X}, \mathbf{v}\rangle + \mathbb{E}\langle \mathbf{Y}, \mathbf{v}\rangle \\
&= \langle \mathbf{x}, \mathbf{v}\rangle + \langle \mathbf{y}, \mathbf{v}\rangle \\
&= \langle \mathbf{x} + \mathbf{y}, \mathbf{v}\rangle
\end{aligned}$$

Exercise 3.5

Suppose the random vector $\mathbf{X}$ maps every point in the sample space to $\mathbf{w}$. Show that $\mathbb{E}\mathbf{X} = \mathbf{w}$.

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 $\mathbf{u}_1, \ldots, \mathbf{u}_m$.

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, $\mathbb{E}\mathbf{v} = \mathbf{v}$.

$$\begin{aligned}\mathbb{E}\langle \mathbf{X}, \mathbf{v}\rangle &= \mathbb{E}\langle \mathbf{w}, \mathbf{v}\rangle \\ &= \langle \mathbf{w}, \mathbf{v}\rangle \underbrace{\mathbb{E}\mathbb{1}_\Omega}_{1} \\ &= \langle \mathbf{w}, \mathbf{v}\rangle\end{aligned}$$

where Ω represents the whole sample space and therefore has probability 1.

This is a simple consequence of Parseval's identity.

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

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

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

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

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 3.12

Suppose X is uncorrelated with each of $Y_1, \ldots, Y_n$. Show that X is also uncorrelated with $a_1 Y_1 + \ldots + a_n Y_n$.

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.

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,

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

$$\begin{aligned}&\mathbb{E}(X - \mathbb{E}X)[a_1Y_1 + \ldots + a_nY_n - \mathbb{E}(a_1Y_1 + \ldots + a_nY_n)] \\ &\quad= \mathbb{E}(X - \mathbb{E}X)[a_1Y_1 + \ldots + a_nY_n - (a_1\mathbb{E}Y_1 + \ldots + a_n\mathbb{E}Y_n)] \\ &\quad= a_1 \underbrace{\mathbb{E}(X - \mathbb{E}X)(Y_1 - \mathbb{E}Y_1)}_{0} + \ldots + a_n \underbrace{\mathbb{E}(X - \mathbb{E}X)(Y_n - \mathbb{E}Y_n)}_{0} \\ &\quad= 0\end{aligned}$$

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

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

Exercise 3.13

If events $E_1, \ldots, E_m$ are *independent* (meaning that their indicator functions are independent random variables) and each has probability q, what's the probability that at least one of them occurs?

Exercise 3.14

Let $\mathbf{X}$ be a random vector and $\mathbf{v}$ be a non-random vector. Show that if $\mathbb{E}\langle \mathbf{X}, \mathbf{v} \rangle$ is real, then it's equal to $\mathbb{E}\langle \mathbf{v}, \mathbf{X} \rangle$.

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.

We can express the first inner product as $\langle \mathbf{X}, \mathbf{v} \rangle = Y + iZ$ for some random variables Y and Z. If its expectation $\mathbb{E}Y + i\mathbb{E}Z$ is real, then $\mathbb{E}Z$ must be 0. The other inner product is the complex conjugate $\langle \mathbf{v}, \mathbf{X} \rangle = Y - iZ$. Its expectation $\mathbb{E}Y - i\mathbb{E}Z$ simplifies to $\mathbb{E}Y$ as well.

Each event has probability $1 - q$ of not occurring. The probability of an intersection of independent events equals the product of their probabilities, so the probability that *none* of the events occur is $(1-q)^m$. The probability that *at least one* occurs is $1 - (1 - q)^m$ since it's the complement of the event that none of them occur.

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,

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

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 3.20

Let $\mathbf{X}$ be a random matrix whose entries have finite expectations, and let $\mathbb{M}$ be a non-random matrix. Assuming $\mathbb{M}\mathbf{X}$ is well-defined, show that $\mathbb{E}\mathbb{M}\mathbf{X} = \mathbb{M}\mathbb{E}\mathbf{X}$. Alternatively, assuming $\mathbf{X}\mathbb{M}$ is well-defined, show that $\mathbb{E}\mathbf{X}\mathbb{M} = (\mathbb{E}\mathbf{X})\mathbb{M}$.

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 $\lambda \in [0, 1]$.

The (i, j) entry of $\mathbb{E}\mathbb{M}\mathbf{X}$ is $\mathbb{E}\mathbf{m}_i^T\mathbf{X}_j = \mathbf{m}_i^T\mathbb{E}\mathbf{X}_j$ where $\mathbf{m}_i$ represents the ith row of $\mathbb{M}$ and $\mathbf{X}_j$ represents the jth column of $\mathbf{X}$. This is also the (i, j) entry of $\mathbb{M}\mathbb{E}\mathbf{X}$. Similarly, the (i, j) entry of $\mathbb{E}\mathbf{X}\mathbb{M}$ is $\mathbb{E}\mathbf{X}_i^T\mathbf{m}_j = (\mathbb{E}\mathbf{X}_i)^T\mathbf{m}_j$ where $\mathbf{X}_i$ represents the ith row of $\mathbf{X}$ and $\mathbf{m}_j$ represents the jth column of $\mathbb{M}$

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 3.21

Show that an alternative expression for the covariance matrix of $\mathbf{Y}$ is $\mathbb{E}[(\mathbf{Y} - \mathbb{E}\mathbf{Y})(\mathbf{Y} - \mathbb{E}\mathbf{Y})^T]$.

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 $\mathbf{Y}$.

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.

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

We'll work out the matrix resulting from the multiplication then move the expectation into the matrix entries.

$$\begin{aligned}\mathbb{E}[(\mathbf{Y}-\mathbb{E}\mathbf{Y})(\mathbf{Y}-\mathbb{E}\mathbf{Y})^T] &= \mathbb{E}\begin{bmatrix} Y_1-\mathbb{E}Y_1 \\ \vdots \\ Y_n-\mathbb{E}Y_n \end{bmatrix}\begin{bmatrix} Y_1-\mathbb{E}Y_1 & \cdots & Y_n-\mathbb{E}Y_n \end{bmatrix} \\ &= \mathbb{E}\begin{bmatrix} (Y_1-\mathbb{E}Y_1)(Y_1-\mathbb{E}Y_1) & \cdots & (Y_1-\mathbb{E}Y_1)(Y_n-\mathbb{E}Y_n) \\ \vdots & \ddots & \vdots \\ (Y_n-\mathbb{E}Y_n)(Y_1-\mathbb{E}Y_1) & \cdots & (Y_n-\mathbb{E}Y_n)(Y_n-\mathbb{E}Y_n) \end{bmatrix} \\ &= \begin{bmatrix} \mathbb{E}[(Y_1-\mathbb{E}Y_1)(Y_1-\mathbb{E}Y_1)] & \cdots & \mathbb{E}[(Y_1-\mathbb{E}Y_1)(Y_n-\mathbb{E}Y_n)] \\ \vdots & \ddots & \vdots \\ \mathbb{E}[(Y_n-\mathbb{E}Y_n)(Y_1-\mathbb{E}Y_1)] & \cdots & \mathbb{E}[(Y_n-\mathbb{E}Y_n)(Y_n-\mathbb{E}Y_n)] \end{bmatrix}\end{aligned}$$

The empirical covariance matrix expression in Equation 1.7 can be understood as an empirical version of this.

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

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

Exercise 3.25

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

Exercise 3.26

Suppose $X_1, \ldots, X_n$ are uncorrelated random variables. Show that the variance of their sum equals the sum of their variances.

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 $\mathbb{H}(\mathbf{v} + \boldsymbol{\epsilon})$.

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.

Let $\mathbf{X} := (X_1, \ldots, X_n)$, and let $\sigma_1^2, \ldots, \sigma_n^2$ represent the variances. Using Exercise 3.23,

$$\begin{aligned}
\operatorname{var} \sum_i X_i &= \operatorname{var} \mathbf{1}^T \mathbf{X} \\
&= \mathbf{1}^T \begin{bmatrix} \sigma_1^2 & & \\ & \ddots & \\ & & \sigma_n^2 \end{bmatrix} \mathbf{1} \\
&= \mathbf{1}^T \begin{bmatrix} \sigma_1^2 \\ \vdots \\ \sigma_n^2 \end{bmatrix} \\
&= \sum_i \sigma_i^2.
\end{aligned}$$

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

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

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

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

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.

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 $\mathbf{X} - \boldsymbol{\mu}$.

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 $\mathbb{M}$.

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.

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

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

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 $\lambda_1, \ldots, \lambda_n$.

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

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

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

Exercise 3.36

Let $X_1, \ldots, X_n$ all have expectation $\boldsymbol{\mu}_X$, and let $Y_1, \ldots, Y_n$ all have expectation $\boldsymbol{\mu}_Y$. Suppose $\mathrm{cov}(X_i, Y_j)$ equals $\sigma_{X,Y}$ if $i = j$ and zero otherwise. Find the expectation of

$$\sum_i (X_i - \bar{X})(Y_i - \bar{Y}).$$

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.

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

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

Let $\mathbf{Y} := (Y_1, \ldots, Y_n)$ and $\mathbf{X} := (X_1, \ldots, X_n)$. The matrix

$$(\mathbf{X} - \boldsymbol{\mu}_X\mathbf{1})(\mathbf{Y} - \boldsymbol{\mu}_Y\mathbf{1})^T = \begin{bmatrix} (X_1 - \boldsymbol{\mu}_X)(Y_1 - \boldsymbol{\mu}_Y) & \cdots & (X_1 - \boldsymbol{\mu}_X)(Y_n - \boldsymbol{\mu}_Y) \\ \vdots & \ddots & \vdots \\ (X_n - \boldsymbol{\mu}_X)(Y_1 - \boldsymbol{\mu}_Y) & \cdots & (X_n - \boldsymbol{\mu}_X)(Y_n - \boldsymbol{\mu}_Y) \end{bmatrix}$$

has expectation $\sigma_{X,Y}\mathbb{I}$.

Let $\mathbb{J}$ be the orthogonal projection matrix onto the span of $\{\mathbf{1}\}$. We'll use the same trace cyclic permutation trick that was advantageous for evaluating expected quadratic forms.

$$\begin{aligned}
\mathbb{E}\sum_i (X_i - \bar{X})(Y_i - \overline{Y}) &= \mathbb{E}(\mathbf{Y} - \overline{Y}\mathbf{1})^T(\mathbf{X} - \bar{X}\mathbf{1}) \\
&= \mathbb{E}[(\mathbb{I} - \mathbb{J})\mathbf{Y}]^T[(\mathbb{I} - \mathbb{J})\mathbf{X}] \\
&= \mathbb{E}[(\mathbb{I} - \mathbb{J})(\mathbf{Y} - \boldsymbol{\mu}_Y\mathbf{1})]^T[(\mathbb{I} - \mathbb{J})(\mathbf{X} - \boldsymbol{\mu}_X\mathbf{1})] \\
&= \mathbb{E}(\mathbf{Y} - \boldsymbol{\mu}_Y\mathbf{1})^T(\mathbb{I} - \mathbb{J})(\mathbf{X} - \boldsymbol{\mu}_X\mathbf{1}) \\
&= \mathbb{E}\operatorname{tr}\left[(\mathbf{Y} - \boldsymbol{\mu}_Y\mathbf{1})^T(\mathbb{I} - \mathbb{J})(\mathbf{X} - \boldsymbol{\mu}_X\mathbf{1})\right] \\
&= \operatorname{tr}\left[(\mathbb{I} - \mathbb{J})\underbrace{\mathbb{E}(\mathbf{X} - \boldsymbol{\mu}_X\mathbf{1})(\mathbf{Y} - \boldsymbol{\mu}_Y\mathbf{1})^T}_{\sigma_{X,Y}\mathbb{I}}\right] \\
&= (n-1)\sigma_{X,Y}
\end{aligned}$$

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

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

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

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 4.3

Let $x_1, \ldots, x_n \in \mathbb{R}$ be values of an explanatory variable, and suppose that the response variable $Y_1, \ldots, Y_n$ satisfies a simple linear model

$$Y_i = \alpha + \beta(x_i - \bar{x}) + \epsilon_i.$$

Assuming $x_1, \ldots, x_n$ are all the same number, show that the coefficients $\hat{\alpha}$ and $\hat{\beta}$ in the least-squares line $y = \hat{\alpha} + \hat{\beta}(x - \bar{x})$ are unbiased estimators for α and β.

Exercise 4.4

Suppose $Y_1, \ldots, Y_n$ are uncorrelated and all have the same variance σ^2. If $x_1, \ldots, x_n$ are values of an explanatory variable, what's the covariance matrix of the coefficients $\hat{\alpha}$ and $\hat{\beta}$ in the least-squares line $y = \hat{\alpha} + \hat{\beta}(x - \bar{x})$?

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 variance of $\hat{a}$ works out to be $\frac{\sigma^2}{n}$, exactly as in Exercise 4.2. The variance of $\hat{b}$ is

$$\begin{aligned}\operatorname{var}\hat{b} &= \operatorname{var}\frac{\frac{1}{n}\langle \mathbf{x}-\bar{x}\mathbf{1}, \mathbf{Y}\rangle}{\sigma^2_{\mathbf{x}}} = \tfrac{1}{n^2\sigma^4_{\mathbf{x}}}\operatorname{cov}(\mathbf{x}-\bar{x}\mathbf{1})^T\mathbf{Y}\\ &= \tfrac{1}{n^2\sigma^4_{\mathbf{x}}}(\mathbf{x}-\bar{x}\mathbf{1})^T\sigma^2\mathbb{I}(\mathbf{x}-\bar{x}\mathbf{1}) = \frac{\sigma^2}{n^2\sigma^4_{\mathbf{x}}}\underbrace{\|\mathbf{x}-\bar{x}\mathbf{1}\|^2}_{n\sigma^2_{\mathbf{x}}} = \frac{\sigma^2}{n\sigma^2_{\mathbf{x}}}\end{aligned}$$

unless $\sigma^2_{\mathbf{x}} = 0$ in which case $\hat{\beta} \equiv 0$ has variance 0. These two variances are the diagonals of the covariance matrix. The off-diagonals are equal to the covariance between $\hat{a}$ and $\hat{b}$. It will be important to realize that the average of the entries of $\mathbb{E}\mathbf{Y}$ is

$$\tfrac{1}{n}\mathbf{1}^T\mathbb{E}\mathbf{Y} = \tfrac{1}{n}\mathbf{1}^T[\alpha\mathbf{1} - \beta(\mathbf{x}-\bar{x}\mathbf{1})] = \alpha - \beta(\bar{x}-\bar{x}) = \alpha.$$

Thus $\overline{Y}-\alpha$ can be rewritten as $\frac{1}{n}\mathbf{1}^T(\mathbf{Y}-\mathbb{E}\mathbf{Y})$.

$$\begin{aligned}\mathbb{E}(\overline{Y}-\alpha)\left(\frac{\frac{1}{n}\langle \mathbf{x}-\bar{x}\mathbf{1}, \mathbf{Y}\rangle}{\sigma^2_{\mathbf{x}}}\right) &= \mathbb{E}\tfrac{1}{n^2\sigma^2_{\mathbf{x}}}\mathbf{1}^T(\mathbf{Y}-\mathbb{E}\mathbf{Y})\mathbf{Y}^T(\mathbf{x}-\bar{x}\mathbf{1})\\ &= \tfrac{1}{n^2\sigma^2_{\mathbf{x}}}\mathbf{1}^T\underbrace{[\mathbb{E}(\mathbf{Y}-\mathbb{E}\mathbf{Y})\mathbf{Y}^T]}_{\sigma^2\mathbb{I}}(\mathbf{x}-\bar{x}\mathbf{1})\\ &= \tfrac{\sigma^2}{n\sigma^2_{\mathbf{x}}}\underbrace{\tfrac{1}{n}\mathbf{1}^T(\mathbf{x}-\bar{x}\mathbf{1})}_{\bar{x}-\bar{x}} = 0\end{aligned}$$

The *least-squares line* has $\hat{\alpha} = \overline{Y}$, and its expectation is

$$\begin{aligned}\mathbb{E}\overline{Y} &= \mathbb{E}(\tfrac{1}{n}\sum_i Y_i)\\ &= \tfrac{1}{n}\sum_i \underbrace{\mathbb{E}Y_i}_{\alpha+\beta(x_i-\bar{x})}\\ &= \alpha + \beta\underbrace{\tfrac{1}{n}\sum_i(x_i-\bar{x})}_{0}\\ &= \alpha.\end{aligned}$$

The other coefficient's expectation is

$$\begin{aligned}\mathbb{E}\hat{\beta} &= \mathbb{E}\frac{\langle \mathbf{x}-\bar{x}\mathbf{1}, \mathbf{Y}-\overline{Y}\mathbf{1}\rangle}{\|\mathbf{x}-\bar{x}\mathbf{1}\|^2}\\ &= \frac{\langle \mathbf{x}-\bar{x}\mathbf{1}, \mathbb{E}\mathbf{Y}-\mathbb{E}\overline{Y}\mathbf{1}\rangle}{\|\mathbf{x}-\bar{x}\mathbf{1}\|^2}\\ &= \frac{\langle \mathbf{x}-\bar{x}\mathbf{1}, [\alpha\mathbf{1}+\beta(\mathbf{x}-\bar{x}\mathbf{1})]-\alpha\mathbf{1}\rangle}{\|\mathbf{x}-\bar{x}\mathbf{1}\|^2}\\ &= \beta\frac{\langle \mathbf{x}-\bar{x}\mathbf{1}, \mathbf{x}-\bar{x}\mathbf{1}\rangle}{\|\mathbf{x}-\bar{x}\mathbf{1}\|^2}\\ &= \beta.\end{aligned}$$

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.

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 $y = \hat{\alpha} + \hat{\beta}_1(x^{(1)} - \bar{x}^{(1)}) + \ldots + \hat{\beta}_m(x^{(m)} - \bar{x}^{(m)})$?

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.

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

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

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

$$\mathbb{E}\overline{Y} = \alpha + \tfrac{1}{n}\sum_i \underbrace{\mathbb{E}\epsilon_i}_{0} = \alpha.$$

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,

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

Exercise 4.9

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, show that $\hat{\alpha}$ is uncorrelated with every $\hat{\beta}_1, \ldots, \hat{\beta}_m$ in the least-squares hyperplane.

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 $y = \hat{\gamma}_1 g_1(\mathbf{x}) + \ldots + \hat{\gamma}_d g_d(\mathbf{x})$?

Exercise 4.12

Assume the columns of $\mathbb{M}$ are linearly independent. Suppose $\mathbf{Y} = \mathbb{M}\boldsymbol{\gamma} + \boldsymbol{\epsilon}$ where $\boldsymbol{\epsilon}$ is a random vector with mean $\mathbf{0}$ and covariance $\sigma^2 \mathbb{I}$. Let $\hat{\boldsymbol{\gamma}} := (\mathbb{M}^T \mathbb{M})^{-1} \mathbb{M}^T \mathbf{Y}$ denote the least-squares estimator for the coefficients. Suppose an alternative estimator $\check{\boldsymbol{\gamma}} := \mathbb{L}\mathbf{Y}$ is also unbiased for $\boldsymbol{\gamma}$. Use the Gauss-Markov theorem to show that $\mathbb{E}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})^T \mathbb{L} (\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}) \leq \mathbb{E}(\check{\boldsymbol{\gamma}} - \boldsymbol{\gamma})^T \mathbb{L} (\check{\boldsymbol{\gamma}} - \boldsymbol{\gamma})$ for every positive semi-definite matrix $\mathbb{L}$.

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.)

Let $\widetilde{\mathbb{X}}$ be the centered version of the explanatory data matrix, and let Σ_j^- be the jth row of the generalized inverse of its empirical covariance matrix (as a column vector). Borrowing tricks from Exercise 4.4, the covariance between $\hat{\alpha}$ and $\hat{\beta}_j$ is

$$\begin{aligned} \mathbb{E}(\overline{Y} - \alpha)((\Sigma_j^-)^T \tfrac{1}{n}\widetilde{\mathbb{X}}^T\mathbf{Y}) &= \mathbb{E}\tfrac{1}{n^2}\mathbf{1}^T(\mathbf{Y} - \mathbb{E}\mathbf{Y})\mathbf{Y}^T\widetilde{\mathbb{X}}\Sigma_j^- \\ &= \tfrac{1}{n^2}\mathbf{1}^T[\underbrace{\mathbb{E}(\mathbf{Y} - \mathbb{E}\mathbf{Y})\mathbf{Y}^T}_{\operatorname{cov}\mathbf{Y} = \sigma^2\mathbb{I}}]\widetilde{\mathbb{X}}\Sigma_j^- \\ &= \tfrac{\sigma^2}{n}\underbrace{\tfrac{1}{n}\mathbf{1}^T\widetilde{\mathbb{X}}}_{\mathbf{0}^T}\Sigma_j^- \\ &= 0. \end{aligned}$$

First, consider the claim of the Gauss-Markov theorem: the variance of $\hat{\boldsymbol{\gamma}}^T\mathbf{v}$ is no greater than the variance of $\tilde{\boldsymbol{\gamma}}^T\mathbf{v}$. An alternative expression for the squared deviation of $\hat{\boldsymbol{\gamma}}^T\mathbf{v}$ from its mean is

$$\begin{aligned} (\hat{\boldsymbol{\gamma}}^T\mathbf{v} - \mathbb{E}\hat{\boldsymbol{\gamma}}^T\mathbf{v})^2 &= (\hat{\boldsymbol{\gamma}}^T\mathbf{v} - \boldsymbol{\gamma}^T\mathbf{v})^2 \\ &= [(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})^T\mathbf{v}][(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})^T\mathbf{v}] \\ &= (\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})^T\mathbf{v}\mathbf{v}^T(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}), \end{aligned}$$

and likewise for $\tilde{\boldsymbol{\gamma}}$. The *expected* squared deviation is the variance, so Gauss-Markov tells us that $\mathbb{E}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})^T\mathbf{v}\mathbf{v}^T(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})$ is no greater than $(\tilde{\boldsymbol{\gamma}} - \boldsymbol{\gamma})^T\mathbf{v}\mathbf{v}^T(\tilde{\boldsymbol{\gamma}} - \boldsymbol{\gamma})$ for every $\mathbf{v}$.

With a spectral decomposition for $\mathbb{L}$,

$$\begin{aligned} (\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})^T\mathbb{L}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}) &= (\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})^T(\lambda_1\mathbf{q}_1\mathbf{q}_1^T + \ldots + \lambda_d\mathbf{q}_d\mathbf{q}_d^T)(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}) \\ &= \lambda_1(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})^T\mathbf{q}_1\mathbf{q}_1^T(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}) + \ldots + \lambda_d(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})^T\mathbf{q}_d\mathbf{q}_d^T)(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}). \end{aligned}$$

Each eigenvalue is non-negative, so each term is no greater than the corresponding expression with $\tilde{\boldsymbol{\gamma}}$ in place of $\hat{\boldsymbol{\gamma}}$.

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.

Exercise 4.13

Suppose $\mathbf{Y} = \mathbb{M}\boldsymbol{\gamma} + \boldsymbol{\epsilon}$ with $\mathbb{E}\boldsymbol{\epsilon} = 0$, and let $\hat{\boldsymbol{\gamma}}$ be coefficients of least-squares linear regression estimators for the correctly specified model. If a new explanatory observation $\mathbf{v}_{n+1}$ is in the row space of $\mathbb{M}$, and $Y_{n+1} = \mathbf{v}_{n+1}^T\boldsymbol{\gamma} + \epsilon_{n+1}$ with $\mathbb{E}\epsilon_{n+1} = 0$, show that the expectation of the predictor $\widehat{Y}_{n+1} = \mathbf{v}_{n+1}^T\hat{\boldsymbol{\gamma}}$ equals the expectation of Y_{n+1} regardless of whether or not $\mathbb{M}$ has full rank.

Exercise 4.14

Let $\mathbf{x} = (x_1, \ldots, x_n)$ and $\mathbf{z} = (z_1, \ldots, z_n)$ be explanatory variables such that $\mathbf{x} = a\mathbf{z}$ for some $c \in \mathbb{R}$. Assume $\mathbf{Y} = (Y_1, \ldots, Y_n)$ satisfy $\mathbf{Y} = b_1\mathbf{x} + b_2\mathbf{z} + \boldsymbol{\epsilon}$ for some $b_1, b_2 \in \mathbb{R}$ and $\mathbb{E}\boldsymbol{\epsilon} = \mathbf{0}$. Argue that the derived parameter $c := ab_1 + b_2$ can be estimated.

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 4.16

Suppose the predictor $\widehat{Y}_{n+1}$ is a function of $Y_1, \ldots, Y_n$ which are *independent* of Y_{n+1}. Show that

$$\mathbb{E}(Y_{n+1} - \widehat{Y}_{n+1})^2 = \operatorname{var} Y_{n+1} + \mathbb{E}(\widehat{Y}_{n+1} - \mathbb{E}Y_{n+1})^2.$$

The expectation of the response variable is

$$\begin{aligned}\mathbb{E}\mathbf{Y} &= b_1\mathbf{x} + b_2\mathbf{z}\\ &= b_1 a\mathbf{z} + b_2\mathbf{z}\\ &= (ab_1 + b_2)\mathbf{z}\\ &= c\mathbf{z}.\end{aligned}$$

Every possible value of a implies a different expectation for the response variable, so it can be estimated. In fact the orthogonal projection's coefficient $\frac{\langle \mathbf{Y},\mathbf{z}\rangle}{\|\mathbf{z}\|^2}$ is unbiased based on Exercise 4.10.

Let $\mathbf{v}_{n+1}^T = \mathbf{w}^T\mathbb{M}$. The expectation of the new response value can be represented as the $\mathbf{w}$ linear combination of the expectations of previous response values:

$$\begin{aligned}\mathbb{E}Y_{n+1} &= \mathbf{v}_{n+1}^T\boldsymbol{\gamma}\\ &= \mathbf{w}^T\mathbb{M}\boldsymbol{\gamma}\\ &= \mathbf{w}^T\mathbb{E}\mathbf{Y}.\end{aligned}$$

The predictor can be written as the same linear combination of the previous predicted values:

$$\begin{aligned}\widehat{Y}_{n+1} &= \mathbf{v}_{n+1}^T\hat{\boldsymbol{\gamma}}\\ &= \mathbf{w}^T\mathbb{M}\hat{\boldsymbol{\gamma}}\\ &= \mathbf{w}^T\widehat{\mathbf{Y}}.\end{aligned}$$

Because $\widehat{\mathbf{Y}}$ is unbiased for $\mathbb{E}\mathbf{Y}$, we have

$$\begin{aligned}\mathbb{E}\widehat{Y}_{n+1} &= \mathbf{w}^T\mathbb{E}\widehat{\mathbf{Y}}\\ &= \mathbf{w}^T\mathbb{E}\mathbf{Y}\end{aligned}$$

which is the same expectation we found for $\mathbb{E}Y_{n+1}$.

$$\begin{aligned}\mathbb{E}(Y_{n+1} - \widehat{Y}_{n+1})^2 &= \mathbb{E}[(Y_{n+1} - \mathbb{E}Y_{n+1}) - (\widehat{Y}_{n+1} - \mathbb{E}Y_{n+1})]^2\\ &= \mathbb{E}(Y_{n+1} - \mathbb{E}Y_{n+1})^2 - 2\mathbb{E}(Y_{n+1} - \mathbb{E}Y_{n+1})(\widehat{Y}_{n+1} - \mathbb{E}Y_{n+1}) + \mathbb{E}(\widehat{Y}_{n+1} - \mathbb{E}Y_{n+1})^2\\ &= \text{var } Y_{n+1} - 2\underbrace{\mathbb{E}(Y_{n+1} - \mathbb{E}Y_{n+1})}_{0}\mathbb{E}(\widehat{Y}_{n+1} - \mathbb{E}Y_{n+1}) + \mathbb{E}(\widehat{Y}_{n+1} - \mathbb{E}Y_{n+1})^2\\ &= \text{var } Y_{n+1} + \mathbb{E}(\widehat{Y}_{n+1} - \mathbb{E}Y_{n+1})^2\end{aligned}$$

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.

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

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 $\mathbb{R}^n$.

Exercise 5.2

For standard Normal random vectors $\mathbf{Z}_1$ and $\mathbf{Z}_2$, suppose $\mathbb{M}_1\mathbf{Z}_1$ and $\mathbb{M}_2\mathbf{Z}_2$ have the same covariance. Show that they have the same distribution.

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.

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.

From Exercise 3.23, we calculate the covariances to be $\mathbb{M}_1\mathbb{M}_1^T$ and $\mathbb{M}_2\mathbb{M}_2^T$.

Suppose $\mathbb{M}_1$ has the singular value decomposition $\mathbb{U}\mathbb{S}\mathbb{V}_1^T$. Then the spectral decomposition of $\mathbb{M}_1\mathbb{M}_1^T$ is $\mathbb{U}\mathbb{S}\mathbb{U}^T$. By the assumption that the covariances are equal, we see that $\mathbb{M}_2\mathbb{M}_2^T$ must also be equal to $\mathbb{U}\mathbb{S}\mathbb{U}^T$. Thus, a singular value decomposition of $\mathbb{M}_2$ has the same matrix $\mathbb{U}$ on the left and the same matrix of singular values; we'll write $\mathbb{M}_2 = \mathbb{U}\mathbb{S}\mathbb{V}_2^T$.

We need to compare the distributions of $\mathbb{U}\mathbb{S}\mathbb{V}_1^T\mathbf{Z}_1$ and $\mathbb{U}\mathbb{S}\mathbb{V}_2^T\mathbf{Z}_2$. The entries of $\mathbb{V}_1^T\mathbf{Z}_1$ are the coordinates of $\mathbf{Z}_1$ with respect to the orthonormal columns of $\mathbb{V}$, so they're iid standard Normal according to our discussion in Section 5.1. Likewise, the entries of $\mathbb{V}_2^T\mathbf{Z}_2$ are standard Normal, so we can conclude that the two random vectors in question have the same distribution.

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 $(Z_1, \ldots, Z_n)$.

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

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 5.4

Show that if $\mathbf{X}_1$ and $\mathbf{X}_2$ are Normal random vectors, then so is $\mathbf{X}_1 + \mathbf{X}_2$.

Exercise 5.5

If two random variables are *multivariate* Normal and are uncorrelated with each other, then they are independent; one can verify that their joint density factors into a product of their marginal densities. However, without *multivariate* Normality, uncorrelated doesn't necessarily imply independent. Construct a pair of Normal random variables that are uncorrelated but not independent.

Exercise 5.6

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

Let $\mathbf{X}_1 \sim N(\boldsymbol{\mu}_1, \mathbb{C}_1)$ and $\mathbf{X}_2 \sim N(\boldsymbol{\mu}_2, \mathbb{C}_2)$. With standard Normal $\mathbf{Z}_1$ and $\mathbf{Z}_2$, we can represent the sum as

$$\begin{aligned}\mathbf{X}_1 + \mathbf{X}_2 &= (\mathbb{C}_1^{1/2}\mathbf{Z}_1 + \boldsymbol{\mu}_1) + (\mathbb{C}_2^{1/2}\mathbf{Z}_2 + \boldsymbol{\mu}_2) \\ &= \begin{bmatrix}\mathbb{C}_1^{1/2} & \mathbb{C}_2^{1/2}\end{bmatrix}\begin{bmatrix}\mathbf{Z}_1 \\ \mathbf{Z}_2\end{bmatrix} + [\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2].\end{aligned}$$

We're almost finished, but consider carefully the vector $\begin{bmatrix}\mathbf{Z}_1 \\ \mathbf{Z}_2\end{bmatrix}$ that has the entries of $\mathbf{Z}_1$ stacked on top of the entries of $\mathbf{Z}_2$. We can't assume that the entries of $\mathbf{Z}_1$ are independent of the entries of $\mathbf{Z}_2$, so the stacked vector isn't necessarily standard Normal. However, with $\mathbb{C}$ representing the covariance matrix of $\begin{bmatrix}\mathbf{Z}_1 \\ \mathbf{Z}_2\end{bmatrix}$ and with $\mathbf{Z}$ a standard Normal random vector of the same size as $\begin{bmatrix}\mathbf{Z}_1 \\ \mathbf{Z}_2\end{bmatrix}$, then we can rewrite the expression as

$$\begin{bmatrix}\mathbb{C}_1^{1/2} & \mathbb{C}_2^{1/2}\end{bmatrix}\mathbb{C}^{1/2}\mathbf{Z} + [\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2]$$

which fits the definition of a Normal random vector.

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.

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 $\mathbb{R}^k$:

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

Let $Z \sim N(0,1)$. Independently of Z, let B take values -1 and 1 each with probability. Finally, define $Y := BZ$. By inspecting the cdf of Y,

$$\begin{aligned}\mathbb{P}(Y \le t) &= \mathbb{P}(B = 1 \cap Z \le t) + \mathbb{P}(B = -1 \cap Z \ge t) \\ &= \mathbb{P}(B = 1)\mathbb{P}(Z \le t) + \mathbb{P}(B = -1)\mathbb{P}(Z \ge -t) \\ &= \mathbb{P}(B = 1)\mathbb{P}(Z \le t) + \mathbb{P}(B = -1)\underbrace{\mathbb{P}(Z \ge -t)}_{\mathbb{P}(Z \le t)} \\ &= \underbrace{[\mathbb{P}(B = 1) + \mathbb{P}(B = -1)]}_{1}\mathbb{P}(Z \le t)\end{aligned}$$

we see that it is also standard Normal as it has the same cdf as Z. If you learn that $Z = z$, you know that Y is either z or $-z$, so Z and Y clearly aren't independent. However, their correlation is

$$\begin{aligned}\mathbb{E}ZY &= \mathbb{E}Z(BZ) \\ &= (\underbrace{\mathbb{E}B}_{0})(\mathbb{E}Z^2) \\ &= 0.\end{aligned}$$

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 $\|\mathbb{H}\mathbf{Z}\|^2$.

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

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

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

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

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.

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 $\frac{\|\mathbb{H}_1 \boldsymbol{\epsilon}\|^2 / \operatorname{rank} \mathbb{H}_1}{\|\mathbb{H}_2 \boldsymbol{\epsilon}\|^2 / \operatorname{rank} \mathbb{H}_2}$.

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.

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.

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 $f_{\operatorname{rank}\mathbb{H}_1,\operatorname{rank}\mathbb{H}_2,\|\mathbf{v}\|^2/\sigma^2}$.

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?

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.

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 $\operatorname{cov}\hat{\boldsymbol{\beta}} = \frac{\sigma^2}{n}\mathbb{\Sigma}^{-1}$:

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

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 $C(\mathbb{M})$.

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

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

The event can be rewritten as

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

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

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

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?

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

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

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

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

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

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

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

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

And in the other test statistic,

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

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 $2\tau_{n-m-1}(-|T_j|)$.

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

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