8

Correlation

Overview

Correlation quantifies to what extent two uncertain quantities are linearly related. Section 8.1 defines the correlation coefficient, a measure of correlation between random variables with zero mean and unit variance. Section 8.2 extends the definition to general random variables, and introduces the covariance, a related measure of linear dependence. In Section 8.3 we describe how to estimate the correlation coefficient and the covariance from data. In Section 8.4 we define the problem of linear regression, and study its connection with correlation. Section 8.5 establishes several important properties of the correlation coefficient. In Section 8.6 we discuss the differences between uncorrelation, defined as a lack of linear dependence, and independence. Section 8.7 provides geometric intuition about correlation, based on the insight that the covariance can be interpreted as an inner product between random variables. Finally, Section 8.8 shows that correlation between two quantities does not necessarily imply a causal relationship.

8.1 Correlation Between Standardized Quantities

The correlation between two random variables $\tilde{a}$ and $\tilde{b}$ quantifies the average *linear* dependence between them. In this section, we assume that the random variables are *standardized*, which means that their mean is zero and their variance is equal to one (we extend our definitions to non-standardized variables in Section 8.2). To determine the linear relationship between the random variables, we ask: *To what extent can we approximate $\tilde{b}$ by scaling $\tilde{a}$?* Answering this question requires determining the optimal scaling factor. If we use mean squared error (MSE) as a metric for estimation accuracy (see Definition 7.29), the optimal factor is the mean of the product between the two variables. We call this quantity the correlation coefficient of $\tilde{a}$ and $\tilde{b}$.

Theorem 8.1 (Linear dependence between standardized random variables)**.** *Let $\tilde{a}$ and $\tilde{b}$ be two random variables with zero mean and unit variance, belonging to the same probability space. The correlation coefficient between $\tilde{a}$ and $\tilde{b}$, defined as*

$$\rho_{\tilde{a},\tilde{b}} := \mathrm{E}[\tilde{a}\tilde{b}] \tag{8.1}$$

satisfies

$$\rho_{\tilde{a},\tilde{b}} = \arg\min_{\beta} \mathrm{E}\left[(\tilde{b} - \beta\tilde{a})^2\right]. \tag{8.2}$$

Consequently, the linear minimum MSE estimator of $\tilde{b}$ *given* $\tilde{a}$ *is equal to* $\rho_{\tilde{a},\tilde{b}}\,\tilde{a}$. *The corresponding MSE equals* $1 - \rho^2_{\tilde{a},\tilde{b}}$.

Proof The MSE is quadratic in β. By linearity of expectation,

$$\begin{aligned} \mathrm{MSE}(\beta) := \mathrm{E}\left[(\tilde{b} - \beta\tilde{a})^2\right] &= \mathrm{E}\left[\tilde{b}^2 - 2\beta\tilde{a}\tilde{b} + \beta^2\tilde{a}^2\right] && (8.3)\\ &= \mathrm{E}[\tilde{b}^2] + \beta^2\mathrm{E}[\tilde{a}^2] - 2\beta\mathrm{E}[\tilde{a}\tilde{b}] && (8.4)\\ &= 1 + \beta^2 - 2\beta\mathrm{E}[\tilde{a}\tilde{b}]. && (8.5) \end{aligned}$$

Taking derivatives reveals that the quadratic function is convex:

$$\begin{aligned} \frac{\mathrm{d}\,\mathrm{MSE}(\beta)}{\mathrm{d}\beta} &= 2\beta - 2\mathrm{E}[\tilde{a}\tilde{b}], && (8.6)\\ \frac{\mathrm{d}^2\,\mathrm{MSE}(\beta)}{\mathrm{d}\beta^2} &= 2, && (8.7) \end{aligned}$$

so we can set the first derivative to zero to find the minimum.

The MSE corresponding to the estimator $\rho_{\tilde{a},\tilde{b}}\,\tilde{a}$ equals

$$\begin{aligned} \mathrm{E}\left[(\tilde{b} - \rho_{\tilde{a},\tilde{b}}\,\tilde{a})^2\right] &= \mathrm{E}[\tilde{b}^2] + \rho^2_{\tilde{a},\tilde{b}}\mathrm{E}[\tilde{a}^2] - 2\rho_{\tilde{a},\tilde{b}}\mathrm{E}[\tilde{a}\tilde{b}] && (8.8)\\ &= 1 - \rho^2_{\tilde{a},\tilde{b}} && (8.9) \end{aligned}$$

by linearity of expectation, the definition of the correlation coefficient, and the assumption that $\tilde{a}$ and $\tilde{b}$ are standardized. ■

Consider the following decomposition of $\tilde{b}$:

$$\tilde{b} = \underbrace{\rho_{\tilde{a},\tilde{b}}\tilde{a}}_{\text{Best linear estimate given } \tilde{a}} + \underbrace{\tilde{b} - \rho_{\tilde{a},\tilde{b}}\tilde{a}}_{\text{Residual}}. \tag{8.10}$$

The first term is the linear estimate obtained by scaling $\tilde{a}$, whereas the second term is the error or residual of the linear estimate. The sign of the correlation coefficient $\rho_{\tilde{a},\tilde{b}}$ indicates whether the component is proportional or inversely proportional to $\tilde{a}$. When it is positive, we say that $\tilde{a}$ and $\tilde{b}$ are positively correlated. When it is negative, they are negatively correlated. When it is zero, they are uncorrelated, which means that the linear approximation is zero, and therefore there is no linear dependence between $\tilde{a}$ and $\tilde{b}$.

As we prove in Section 8.5.1, the correlation coefficient $\rho_{\tilde{a},\tilde{b}}$ is bounded between -1 and 1. By Theorem 8.1, the error of the best linear estimate of $\tilde{b}$ given $\tilde{a}$ equals $1 - {\rho_{\tilde{a},\tilde{b}}}^2$. The linear approximation is worst when the variables are uncorrelated ($\rho_{\tilde{a},\tilde{b}} = 0$), and improves as the magnitude of correlation coefficient increases. In fact, the approximation is perfect when $\rho_{\tilde{a},\tilde{b}}$ is 1 or -1! The correlation coefficient therefore quantifies to what extent $\tilde{a}$ and $\tilde{b}$ are linearly related. Figure 8.1 shows samples from random variables with different correlation coefficients, along with the corresponding linear estimates and residuals.

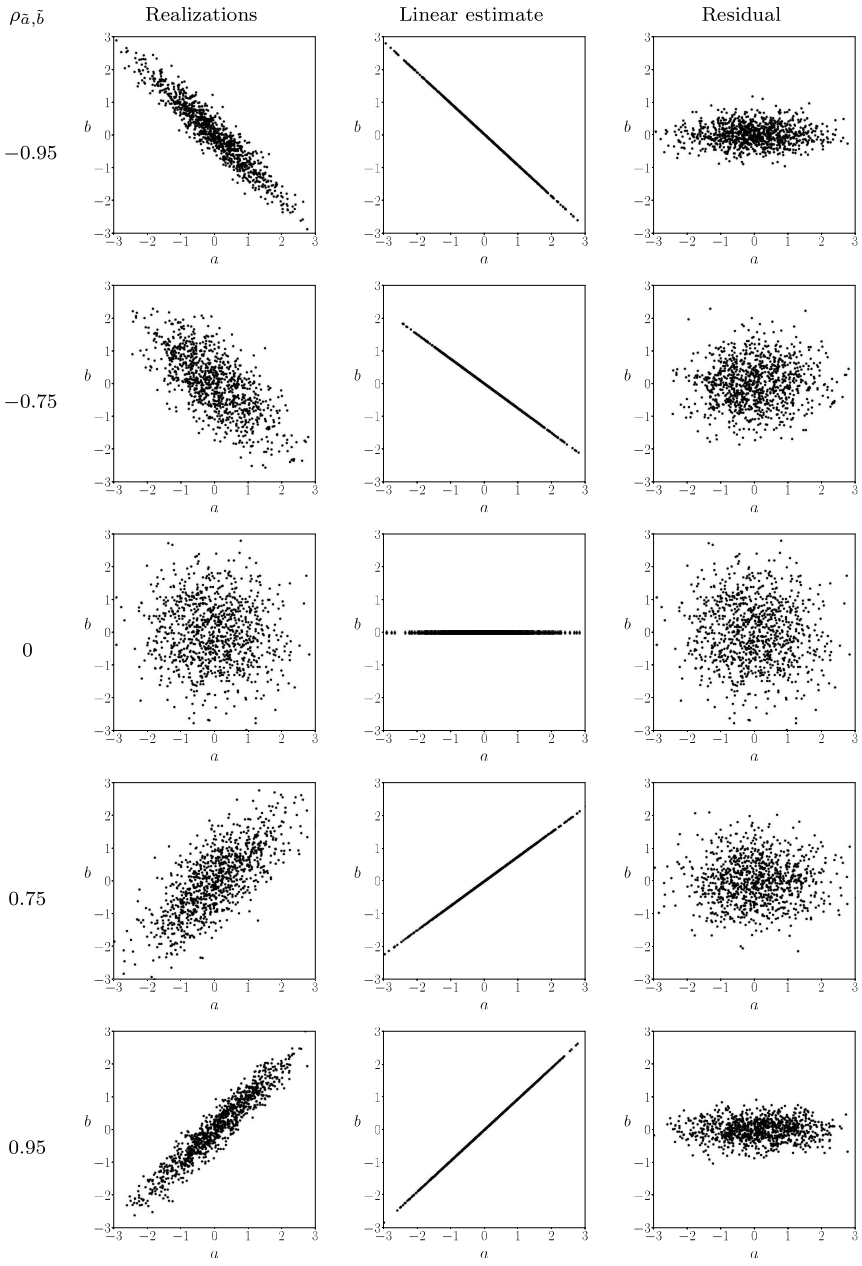

Figure 8.1 The correlation coefficient quantifies linear dependence between random variables. The left column shows scatterplots of 1,000 i.i.d. samples from two Gaussian random variables $\tilde{a}$ and $\tilde{b}$ with zero mean, unit variance and different correlation coefficients. The central column shows the best linear approximation $\rho_{\tilde{a},\tilde{b}}a$ of $\tilde{b}$ given $\tilde{a} = a$ for these samples. The right column shows the corresponding residual $\tilde{b} - \rho_{\tilde{a},\tilde{b}}\tilde{a}$. The sign of $\rho_{\tilde{a},\tilde{b}}$ indicates whether $\tilde{b}$ is proportional or inversely proportional to $\tilde{a}$ on average. The magnitude of $\rho_{\tilde{a},\tilde{b}}$ determines the accuracy of the linear approximation.

8.2 Correlation And Covariance

The previous section defines the correlation coefficient between standardized random variables with zero mean and unit variance. We now consider two random variables $\tilde{a}$ and $\tilde{b}$ with arbitrary means and variances. To ease notation, we denote the means of $\tilde{a}$ and $\tilde{b}$ by $\mu_{\tilde{a}}$ and $\mu_{\tilde{b}}$, and their standard deviations by $\sigma_{\tilde{a}}$ and $\sigma_{\tilde{b}}$. A useful first step is to *standardize* the random variables by centering and normalizing them. The resulting variables have zero mean and unit variance.

Definition 8.2 (Standardized variable)**.** *Given a random variable $\tilde{a}$ with mean $\mu_{\tilde{a}}$ and standard deviation $\sigma_{\tilde{a}}$, the standardized variable* $\mathrm{s}(\tilde{a})$ *of $\tilde{a}$ is obtained by subtracting its mean and dividing by its standard deviation,*

$$\mathrm{s}(\tilde{a}) := \frac{\tilde{a} - \mu_{\tilde{a}}}{\sigma_{\tilde{a}}}. \tag{8.11}$$

Lemma 8.3 (Mean and variance of a standardized variable)**.** *The standardized variable* $\mathrm{s}(\tilde{a})$ *corresponding to a random variable $\tilde{a}$ with mean $\mu_{\tilde{a}}$ and standard deviation $\sigma_{\tilde{a}}$ has zero mean and unit variance.*

Proof By linearity of expectation,

$$\begin{aligned}
\mathrm{E}\left[\mathrm{s}(\tilde{a})\right] &= \mathrm{E}\left[\frac{\tilde{a} - \mu_{\tilde{a}}}{\sigma_{\tilde{a}}}\right] && (8.12)\\
&= \frac{\mathrm{E}\left[\tilde{a}\right] - \mu_{\tilde{a}}}{\sigma_{\tilde{a}}} = 0, && (8.13)\\
\mathrm{Var}\left[\mathrm{s}(\tilde{a})\right] &= \mathrm{E}\left[\mathrm{s}(\tilde{a})^2\right] && (8.14)\\
&= \mathrm{E}\left[\frac{(\tilde{a} - \mu_{\tilde{a}})^2}{\sigma_{\tilde{a}}^2}\right] && (8.15)\\
&= \frac{\mathrm{E}\left[(\tilde{a} - \mu_{\tilde{a}})^2\right]}{\sigma_{\tilde{a}}^2} = 1. && (8.16)
\end{aligned}$$

■

Since the standardized random variables $s(\tilde{a})$ and $s(\tilde{b})$ have zero mean and unit variance, as explained in Section 8.1, the correlation coefficient $\rho_{\mathrm{s}(\tilde{a}),\mathrm{s}(\tilde{b})}$ quantifies to what extent $\mathrm{s}(\tilde{a})$ and $\mathrm{s}(\tilde{b})$ are linearly related. By Theorem 8.1, the best linear approximation of $s(\tilde{b})$ given $s(\tilde{a})$ is $\rho_{\mathrm{s}(\tilde{a}),\mathrm{s}(\tilde{b})}s(\tilde{a})$. Since $\tilde{b} = \sigma_{\tilde{b}}s(\tilde{b}) + \mu_{\tilde{b}}$, this immediately provides an affine approximation of $\tilde{b}$ given $\tilde{a}$, :

$$\begin{aligned}
\tilde{b} &\approx \sigma_{\tilde{b}}\rho_{\mathrm{s}(\tilde{a}),\mathrm{s}(\tilde{b})}s(\tilde{a}) + \mu_{\tilde{b}} && (8.17)\\
&= \frac{\sigma_{\tilde{b}}\rho_{\mathrm{s}(\tilde{a}),\mathrm{s}(\tilde{b})}(\tilde{a} - \mu_{\tilde{a}})}{\sigma_{\tilde{a}}} + \mu_{\tilde{b}}. && (8.18)
\end{aligned}$$

In Section 8.4 we show that this approximation is optimal in terms of mean squared error. The correlation coefficient $\rho_{\mathrm{s}(\tilde{a}),\mathrm{s}(\tilde{b})}$ therefore determines to what extent $\tilde{b}$ can be approximated using an affine function of $\tilde{a}$ (or vice versa). Motivated by this, we define the correlation coefficient between $\tilde{a}$ and $\tilde{b}$ as the correlation coefficient between their standardized counterparts $\mathrm{s}(\tilde{a})$ and $\mathrm{s}(\tilde{b})$. This

yields a measure of linear dependence that is invariant to positive scaling and shifting. For any positive scaling factors β_1 and β_2, and any additive constants α_1 and α_2, the standardized counterparts of $\beta_1 \tilde{a} + \alpha_1$ and $\beta_2 \tilde{b} + \alpha_2$ are $s(\tilde{a})$ and $s(\tilde{b})$, respectively, so the correlation coefficient remains the same.

In terms of $\tilde{a}$ and $\tilde{b}$ and their means and variances, the correlation coefficient equals

$$\begin{aligned} \rho_{\tilde{a},\tilde{b}} &:= \rho_{s(\tilde{a}),s(\tilde{b})} && (8.19)\\ &= \mathrm{E}\left[s(\tilde{a})s(\tilde{b})\right] && (8.20)\\ &= \mathrm{E}\left[\frac{\tilde{a}-\mu_{\tilde{a}}}{\sigma_{\tilde{a}}} \cdot \frac{\tilde{b}-\mu_{\tilde{b}}}{\sigma_{\tilde{b}}}\right] && (8.21)\\ &= \frac{\mathrm{E}\left[(\tilde{a}-\mu_{\tilde{a}})(\tilde{b}-\mu_{\tilde{b}})\right]}{\sigma_{\tilde{a}}\,\sigma_{\tilde{b}}}. && (8.22) \end{aligned}$$

The numerator is the mean of the product between $\tilde{a}$ and $\tilde{b}$ after centering. This quantity is known as the covariance between $\tilde{a}$ and $\tilde{b}$.

Definition 8.4 (Covariance)**.** *Let $\tilde{a}$ and $\tilde{b}$ be two random variables with finite mean belonging to the same probability space. The covariance of $\tilde{a}$ and $\tilde{b}$ is*

$$\mathrm{Cov}[\tilde{a},\tilde{b}] := \mathrm{E}\left[(\tilde{a}-\mu_{\tilde{a}})\,(\tilde{b}-\mu_{\tilde{b}})\right], \tag{8.23}$$

where $\mu_{\tilde{a}}$ and $\mu_{\tilde{b}}$ denote the means of $\tilde{a}$ and $\tilde{b}$.

The following lemma shows that the covariance can be obtained by subtracting the product of the means from the mean of the product.

Lemma 8.5. *The covariance of two random variables $\tilde{a}$ and $\tilde{b}$ with means $\mu_{\tilde{a}}$ and $\mu_{\tilde{b}}$ equals*

$$\mathrm{Cov}[\tilde{a},\tilde{b}] := \mathrm{E}[\tilde{a}\tilde{b}] - \mu_{\tilde{a}}\mu_{\tilde{b}}. \tag{8.24}$$

Proof By linearity of expectation,

$$\begin{aligned} \mathrm{E}\left[(\tilde{a}-\mu_{\tilde{a}})\,(\tilde{b}-\mu_{\tilde{b}})\right] &= \mathrm{E}[\tilde{a}\tilde{b}] - \mathrm{E}[\tilde{a}]\mu_{\tilde{b}} - \mu_{\tilde{a}}\mathrm{E}[\tilde{b}] + \mu_{\tilde{a}}\,\mu_{\tilde{b}} && (8.25)\\ &= \mathrm{E}[\tilde{a}\tilde{b}] - \mu_{\tilde{a}}\mu_{\tilde{b}}. && (8.26) \end{aligned}$$

■

The correlation coefficient is usually defined as the covariance normalized by the standard deviations. This coincides with our initial definition in (8.22).

Definition 8.6 (Correlation coefficient)**.** *Let $\tilde{a}$ and $\tilde{b}$ be two random variables with finite mean and variance belonging to the same probability space. The correlation coefficient of $\tilde{a}$ and $\tilde{b}$ is*

$$\rho_{\tilde{a},\tilde{b}} := \frac{\mathrm{Cov}[\tilde{a},\tilde{b}]}{\sigma_{\tilde{a}}\sigma_{\tilde{b}}}, \tag{8.27}$$

where $\sigma_{\tilde{a}}$ and $\sigma_{\tilde{b}}$ are the standard deviations of $\tilde{a}$ and $\tilde{b}$, respectively.

The covariance and the correlation coefficient determine whether the two variables are directly or inversely proportional on average. If they are positive, we say that the variables are positively correlated. If they are negative, they are negatively correlated. If they are zero, they are uncorrelated, which means that there is no linear dependence.

Example 8.7 (Correlation between cats and dogs)**.** In Example 7.6 we study the distribution of the number of cats and dogs per household in a new market entered by a fictitious pet-food producer. In order to summarize the dependence between the two quantities, we can compute the correlation between the random variables $\tilde{c}$ and $\tilde{d}$, which represent the number of cats and dogs respectively. We have

$$\begin{aligned}
\mathrm{E}[\tilde{c}\,\tilde{d}] &:= \sum_{c=0}^{3}\sum_{d=0}^{2} c\,d\,p_{\tilde{c},\tilde{d}}(c,d) && (8.28)\\
&= 1\cdot 0.05 + 2(0.03+0.02) && (8.29)\\
&= 0.15, && (8.30)\\
\mathrm{E}[\tilde{c}] &:= \sum_{c=0}^{3}\sum_{d=0}^{2} c\,p_{\tilde{c},\tilde{d}}(c,d) = 0.63, && (8.31)\\
\mathrm{E}[\tilde{d}] &:= \sum_{c=0}^{3}\sum_{d=0}^{2} d\,p_{\tilde{c},\tilde{d}}(c,d) = 0.42, && (8.32)
\end{aligned}$$

so the covariance equals

$$\begin{aligned}
\mathrm{Cov}[\tilde{c},\tilde{d}] &= \mathrm{E}[\tilde{c}\,\tilde{d}] - \mathrm{E}[\tilde{c}]\mathrm{E}[\tilde{d}] && (8.33)\\
&= -0.115. && (8.34)
\end{aligned}$$

The number of cats and the number of dogs are negatively correlated. This indicates that, on average, people with more dogs have less cats (and vice versa). The variances of $\tilde{c}$ and $\tilde{d}$ equal

$$\begin{aligned}
\mathrm{Var}[\tilde{c}] &= \mathrm{E}[\tilde{c}^2] - \mathrm{E}[\tilde{c}]^2 = 0.793, && (8.35)\\
\mathrm{Var}[\tilde{d}] &= \mathrm{E}[\tilde{d}^2] - \mathrm{E}[\tilde{d}]^2 = 0.384. && (8.36)
\end{aligned}$$

To obtain the correlation coefficient we divide the covariance by the product of the standard deviations,

$$\begin{aligned}
\rho_{\tilde{c},\tilde{d}} &:= \frac{\mathrm{Cov}[\tilde{c},\tilde{d}]}{\sqrt{\mathrm{Var}[\tilde{c}]\mathrm{Var}[\tilde{d}]}} && (8.37)\\
&= -0.208. && (8.38)
\end{aligned}$$

This provides a normalized quantification of the linear dependence between the number of cats and the number of dogs.

Example 8.8 (Dependence between Gaussian random variables)**.** In Example 5.23 we study a bivariate Gaussian random vector $(\tilde{a}, \tilde{b})$ with zero mean and covariance matrix

$$\Sigma := \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}. \tag{8.39}$$

In particular, we establish that the conditional distribution of $\tilde{b}$ given $\tilde{a} = a$ is Gaussian with mean ρa and variance $1 - \rho^2$. We now derive the correlation coefficient of $\tilde{a}$ and $\tilde{b}$. We begin by computing the conditional mean function of $\tilde{a}\tilde{b}$ given $\tilde{a}$. Conditioned on $\tilde{a} = a$, $\tilde{a}$ is just a constant, so

$$\mu_{\tilde{a}\tilde{b} \,|\, \tilde{a}}(a) = \int_{b=-\infty}^{\infty} ab \, f_{\tilde{b} \,|\, \tilde{a}}(b \,|\, a) \, \mathrm{d}b \tag{8.40}$$

$$= a \, \mu_{\tilde{b} \,|\, \tilde{a}}(a) \tag{8.41}$$

$$= \rho a^2. \tag{8.42}$$

Consequently, by iterated expectation

$$\rho_{\tilde{a},\tilde{b}} = \mathrm{E}\left[\tilde{a}\tilde{b}\right] \tag{8.43}$$

$$= \mathrm{E}\left[\mu_{\tilde{a}\tilde{b} \,|\, \tilde{a}}(\tilde{a})\right] \tag{8.44}$$

$$= \mathrm{E}\left[\rho \tilde{a}^2\right] \tag{8.45}$$

$$= \rho \mathrm{E}\left[\tilde{a}^2\right] \tag{8.46}$$

$$= \rho. \tag{8.47}$$

The correlation coefficient is equal to the parameter ρ. In Example 5.23 we show that ρ completely determines the dependence between $\tilde{a}$ and $\tilde{b}$. When it is zero, they are independent, there is no dependence at all. When ρ is close to 1 or -1, $\tilde{b}$ is closely concentrated around $\tilde{a}$ or $-\tilde{a}$, respectively (see Figure 5.17).

8.3 Estimating Correlation From Data

The covariance of two random variables is the mean of their product after centering. It is therefore natural to use averaging in order to estimate the covariance from data. As in the definition of the sample variance, we average dividing by $n - 1$, where n is the number of data points, in order to ensure that the estimate is unbiased (see Exercise 9.4).

Definition 8.9 (Sample covariance)**.** *Given a dataset formed by pairs of real-valued examples* (x_1, y_1), (x_2, y_2), $\ldots$, (x_n, y_n), $1 \leq i \leq n$, *let* $X := \{x_1, x_2, \ldots, x_n\}$ *and* $Y := \{y_1, y_2, \ldots, y_n\}$. *The sample covariance of the data equals*

$$c(X, Y) := \frac{1}{n-1} \sum_{i=1}^{n} (x_i - m(X))(y_i - m(Y)), \tag{8.48}$$

where $m(X)$ *and* $m(Y)$ *denote the sample means of* X *and* Y*, respectively.*

In order to estimate the correlation coefficient from data, we apply Definition 8.6 replacing the covariance and variances by the sample covariance and sample variances, respectively.

Definition 8.10 (Sample correlation coefficient)**.** *Given a dataset formed by pairs of real-valued examples* (x_1, y_1), (x_2, y_2), $\ldots$, (x_n, y_n), $1 \leq i \leq n$, *let* $X := \{x_1, x_2, \ldots, x_n\}$ *and* $Y := \{y_1, y_2, \ldots, y_n\}$. *The sample correlation coefficient of the data equals*

$$\rho_{X,Y} := \frac{c(X,Y)}{\sqrt{v(X)v(Y)}}, \tag{8.49}$$

where $c(X, Y)$ *is the sample covariance of* X *and* Y, *and* $v(X)$ *and* $v(Y)$ *are the sample variances of* X *and* Y, *respectively.*

As explained in Section 8.2, the correlation coefficient of two random variables $\tilde{a}$ and $\tilde{b}$ can be interpreted as the optimal scaling factor when approximating the standardized variable $s(\tilde{b})$ as a linear function of the standardized variable $s(\tilde{a})$. A similar interpretation holds for the sample correlation coefficient. To show this, we first need to define a standardization operation that can be applied to data.

Definition 8.11 (Standardized data)**.** *Let* $X := \{x_1, x_2, \ldots, x_n\}$ *denote a real-valued dataset. The corresponding standardized data are obtained by subtracting the sample mean* $m(X)$ *and dividing by the sample standard deviation* $\sqrt{v(X)}$,

$$s(x_i) := \frac{x_i - m(X)}{\sqrt{v(X)}}, \quad 1 \leq i \leq n. \tag{8.50}$$

The following theorem is the finite-data counterpart to Theorem 8.1. It shows that the sample correlation coefficient is an optimal scaling for standardized data, in the sense that it produces the best linear approximation, in terms of the sum of squared errors.

Theorem 8.12 (Linear dependence in standardized data)**.** *Given a dataset formed by* n *pairs* (x_1, y_1), (x_2, y_2), $\ldots$, (x_n, y_n), $1 \leq i \leq n$, *let* $s(x_i)$ *and* $s(y_i)$ *denote the corresponding standardized data for* $1 \leq i \leq n$. *The sample correlation coefficient* $\rho_{X,Y}$ *provides an optimal linear approximation of one standardized variable given the other in terms of least-squares error:*

$$\rho_{X,Y} = \arg\min_{\beta} \sum_{i=1}^{n} (s(y_i) - \beta s(x_i))^2. \tag{8.51}$$

Proof It is straightforward to show that the sample mean of a standardized dataset is zero and its sample variance is one (see Exercise 8.7). Consequently,

$$\sum_{i=1}^{n} s(x_i)^2 = \sum_{i=1}^{n} s(y_i)^2 = n - 1, \tag{8.52}$$

and the correlation coefficient of the standardized variables is equal to their sample covariance,

$$\rho_{X,Y} = \frac{1}{n-1} \sum_{i=1}^{n} s(x_i) s(y_i). \tag{8.53}$$

This allows us to express the residual sum of squares of the linear approximation as a simple function of the scaling factor β:

$$\begin{aligned} \text{RSS}(\beta) &:= \sum_{i=1}^{n} (s(y_i) - \beta s(x_i))^2 && (8.54) \\ &= \sum_{i=1}^{n} s(y_i)^2 + \beta^2 \sum_{i=1}^{n} s(x_i)^2 - 2\beta \sum_{i=1}^{n} s(x_i) s(y_i) && (8.55) \\ &= (n-1) \left(1 + \beta^2 - 2\beta \rho_{X,Y}\right). && (8.56) \end{aligned}$$

The derivatives of the function are

$$\frac{\mathrm{d}\,\text{RSS}(\beta)}{\mathrm{d}\beta} = 2(n-1)(\beta - \rho_{X,Y}), \tag{8.57}$$

$$\frac{\mathrm{d}^2\,\text{RSS}(\beta)}{\mathrm{d}\beta^2} = 2(n-1), \tag{8.58}$$

so it is strictly convex and we can minimize it by setting the first derivative to zero. We conclude that the optimal scaling is the sample correlation coefficient. ■

Example 8.13 (Height of NBA players)**.** Height is obviously very important in basketball. Here we study how to it relates to the offensive productivity of players in the NBA between 1996 and 2019, extracted from Dataset 13. Our data consist of the height of each player and their points, rebounds and assists per game. The sample correlation coefficient is -0.06 for height and points, 0.42 for height and rebounds, and -0.46 for height and assists. On average, taller players capture more rebounds, give less assists, and score slightly less than shorter players. Figure 8.2 shows scatterplots of the data, together with the corresponding densities before and after standardizing. The bottom two rows show the linear estimate associated to the correlation coefficient and the corresponding residual, obtained by subtracting the linear estimate from each data point.

8.4 Simple Linear Regression

As we explain in Section 7.8.3, regression is the problem of estimating a quantity of interest, called the response, as a function of the observed features. In this section, we consider simple linear regression, where there is a single feature and the regression estimate is constrained to be affine. The general regression problem with multiple input variables is discussed in detail in Chapter 12.

We model the response and the feature as random variables. Our goal is to

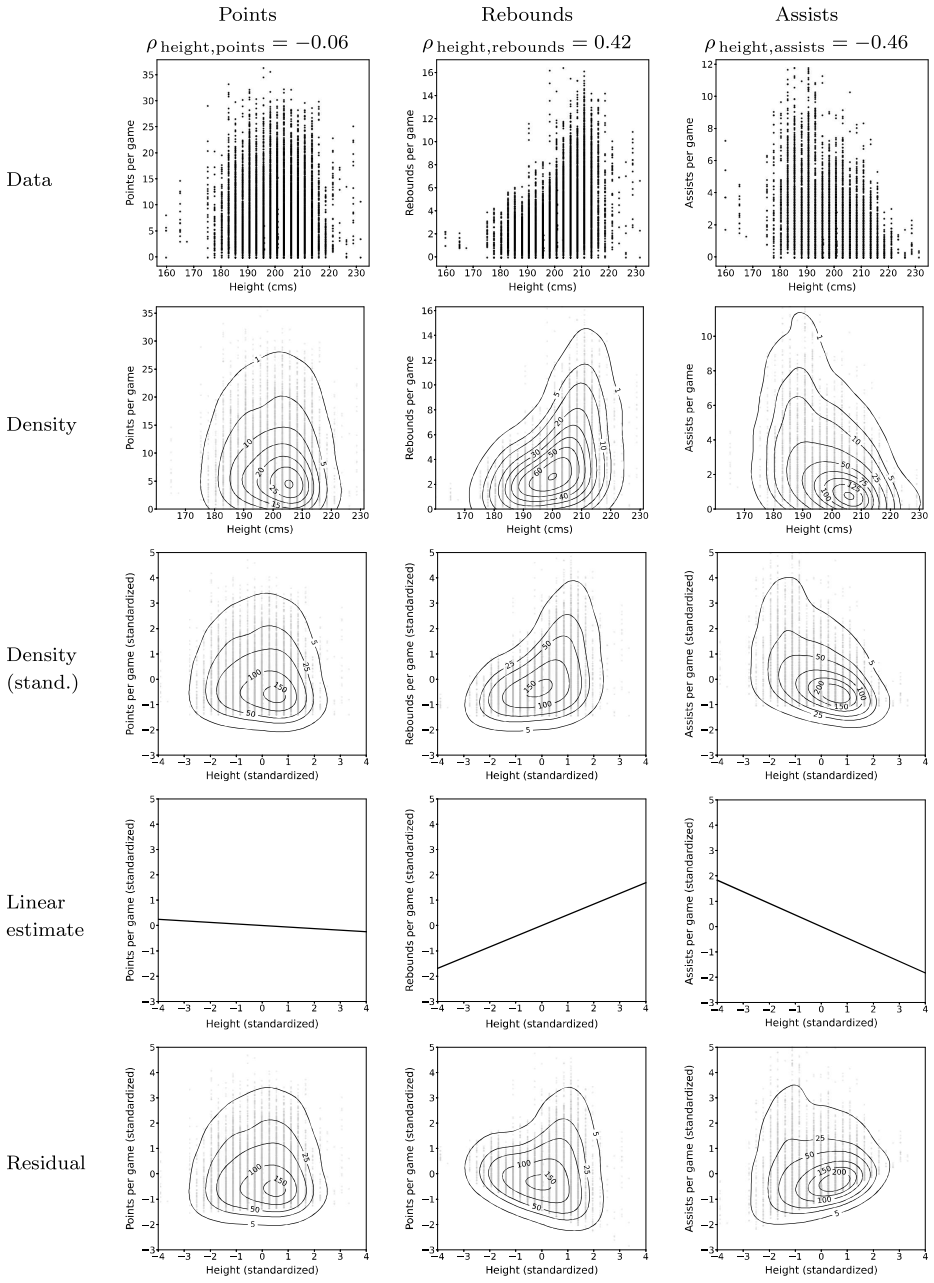

Figure 8.2 Linear dependence with height. The top row shows scatterplots of points (left), rebounds (center) and assists (right) per game against height for NBA players. The second row shows the corresponding bivariate probability densities estimated via kernel density estimation. The third row shows the density of the data standardized using the sample means and variances. The fourth row shows the best linear approximation given height. The fifth row shows the corresponding residuals, obtained by subtracting the linear estimate from each data point. The correlation coefficients indicate that rebounds are positively correlated with height, and assists are negatively correlated with height. Points are also negatively correlated with height, but the linear relationship is much weaker.

approximate the response $\tilde{b}$ using a linear transformation of the feature $\tilde{a}$. Theorem 8.1 establishes that $\rho_{\tilde{a},\tilde{b}} s(\tilde{a})$, where $s(\tilde{a})$ is the result of standardizing $\tilde{a}$, is the linear minimum MSE estimator of the standardized variable $s(\tilde{b})$. Since

$$\tilde{b} = \sigma_{\tilde{b}} s(\tilde{b}) + \mu_{\tilde{b}}, \tag{8.59}$$

as we explain at the beginning of Section 8.2, a reasonable estimator for $\tilde{b}$ is

$$\sigma_{\tilde{b}} \rho_{\tilde{a},\tilde{b}} s(\tilde{a}) + \mu_{\tilde{b}}. \tag{8.60}$$

The transformation is illustrated in Figure 8.3. Strictly speaking, the estimator is affine, because it consists of a linear term and an additive constant. However, it is commonly referred to as a linear estimator of $\tilde{b}$ given $\tilde{a}$. It turns out that this is the best possible affine estimator in terms of MSE, so we call it the linear MMSE (minimum MSE) estimator.

Theorem 8.14 (Linear MMSE estimator)**.** *Let $\tilde{a}$ and $\tilde{b}$ be two random variables with means $\mu_{\tilde{a}}$ and $\mu_{\tilde{b}}$, variances $\sigma^2_{\tilde{a}}$ and $\sigma^2_{\tilde{b}}$, and correlation coefficient $\rho_{\tilde{a},\tilde{b}}$. The linear minimum MSE estimator of $\tilde{b}$ given $\tilde{a} = a$ is*

$$\ell_{\text{MMSE}}(a) := \beta_{\text{MMSE}} a + \alpha_{\text{MMSE}} \tag{8.61}$$

$$= \sigma_{\tilde{b}} \, \rho_{\tilde{a},\tilde{b}} \left(\frac{a - \mu_{\tilde{a}}}{\sigma_{\tilde{a}}} \right) + \mu_{\tilde{b}}. \tag{8.62}$$

This estimator yields the best affine estimate $\tilde{b}$ given $\tilde{a}$ in terms of MSE:

$$\beta_{\text{MMSE}}, \alpha_{\text{MMSE}} = \arg\min_{\beta,\alpha} \mathrm{E}\left[(\tilde{b} - \beta\tilde{a} - \alpha)^2 \right]. \tag{8.63}$$

Proof We denote the MSE as a function of β and α by

$$\text{MSE}(\beta, \alpha) := \mathrm{E}\left[(\tilde{b} - \beta\tilde{a} - \alpha)^2 \right]. \tag{8.64}$$

If we fix $\beta \in \mathbb{R}$, the optimal value of α, denoted by $\alpha^*(\beta)$, for that particular β is the best constant estimate of the random variable $\tilde{b} - \beta\tilde{a}$, which equals

$$\alpha^*(\beta) := \arg\min_{\alpha} \mathrm{E}\left[(\tilde{b} - \beta\tilde{a} - \alpha)^2 \right] \tag{8.65}$$

$$= \mathrm{E}\left[\tilde{b} - \beta\tilde{a} \right] \tag{8.66}$$

$$= \mu_{\tilde{b}} - \beta\mu_{\tilde{a}}, \tag{8.67}$$

by Theorem 7.30 and linearity of expectation. Consequently, for any β and any α $\text{MSE}(\beta, \alpha) \geq \text{MSE}(\beta, \alpha^*(\beta))$. To obtain β_{MMSE}, we fix α to $\alpha^*(\beta)$ and minimize the resulting MSE,

$$\beta_{\text{MMSE}} = \arg\min_{\beta} \text{MSE}(\beta, \alpha^*(\beta)). \tag{8.68}$$

By linearity of expectation,

$$\text{MSE}(\beta, \alpha^*(\beta)) = \mathrm{E}\left[(\tilde{b} - \beta\tilde{a} - \mu_{\tilde{b}} + \beta\mu_{\tilde{a}})^2 \right] \tag{8.69}$$

$$= \mathrm{E}\left[(\tilde{b} - \mu_{\tilde{b}})^2 \right] + \beta^2 \mathrm{E}\left[(\tilde{a} - \mu_{\tilde{a}})^2 \right] - 2\beta \mathrm{E}\left[(\tilde{a} - \mu_{\tilde{a}}) (\tilde{b} - \mu_{\tilde{b}}) \right]$$

$$= \sigma^2_{\tilde{b}} + \sigma^2_{\tilde{a}} \beta^2 - 2\text{Cov}[\tilde{a}, \tilde{b}]\beta. \tag{8.70}$$

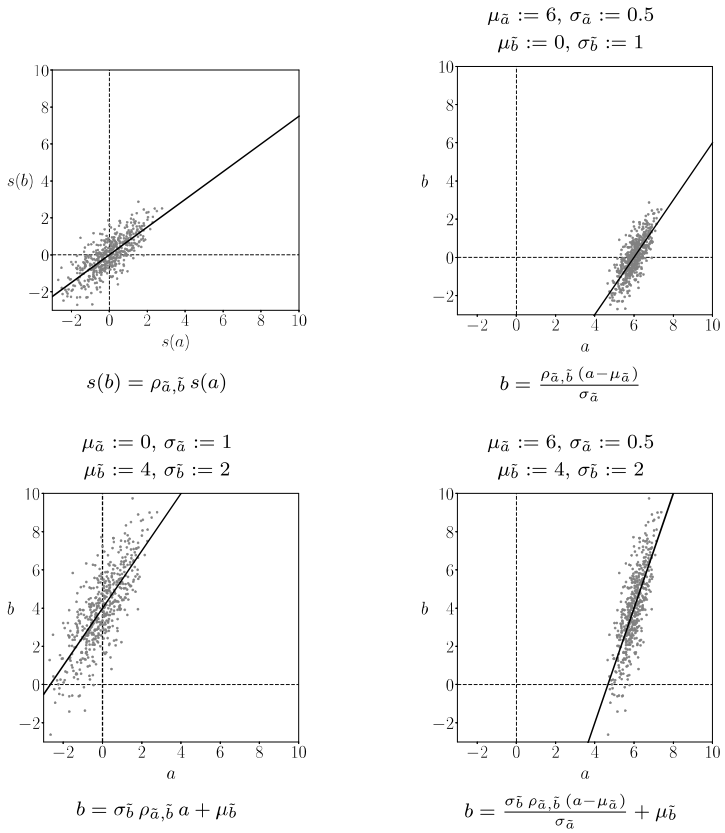

Figure 8.3 Linear MMSE estimator of non-standardized random variables. The upper left plot shows 500 samples (gray dots) from two standardized random variables $s(\tilde{a})$ and $s(\tilde{b})$, and the corresponding linear MMSE estimate of $s(\tilde{b})$ given $s(\tilde{a})$ (black line). The three other plots show different possible original data before standardization for different values of the means and standard deviations of $\tilde{a}$ and $\tilde{b}$ (gray dots). The linear estimate for the standardized variables can be shifted and scaled according to the means and standard deviations to obtain an estimate for the original data. The resulting affine estimate for each case is plotted as a black line, and its equation is provided below each graph.

The derivatives with respect to β equal

$$\frac{\mathrm{d}\,\mathrm{MSE}(\beta, \alpha^*(\beta))}{\mathrm{d}\beta} = 2\left(\sigma_{\tilde{a}}^2\beta - \mathrm{Cov}[\tilde{a}, \tilde{b}]\right), \tag{8.71}$$

$$\frac{\mathrm{d}^2\,\mathrm{MSE}(\beta, \alpha^*(\beta))}{\mathrm{d}\beta^2} = 2\sigma_{\tilde{a}}^2 \geq 0. \tag{8.72}$$

The function is strictly convex, as long as the variance of $\tilde{a}$ is nonzero, so we can

obtain β_{MMSE} by setting the derivative to zero,

$$\beta_{\text{MMSE}} = \frac{\text{Cov}[\tilde{a}, \tilde{b}]}{\sigma_{\tilde{a}}^2} \tag{8.73}$$

$$= \frac{\rho_{\tilde{a},\tilde{b}}\, \sigma_{\tilde{b}}}{\sigma_{\tilde{a}}}. \tag{8.74}$$

The corresponding value of $\alpha^*(\beta)$ is

$$\alpha_{\text{MMSE}} = \alpha^*(\beta_{\text{MMSE}}) \tag{8.75}$$

$$= \mu_{\tilde{b}} - \beta_{\text{MMSE}} \mu_{\tilde{a}} \tag{8.76}$$

$$= \mu_{\tilde{b}} - \frac{\rho_{\tilde{a},\tilde{b}}\, \sigma_{\tilde{b}}\, \mu_{\tilde{a}}}{\sigma_{\tilde{a}}}. \tag{8.77}$$

Finally, we verify that

$$\ell_{\text{MMSE}}(a) := \beta_{\text{MMSE}} a + \alpha_{\text{MMSE}} \tag{8.78}$$

$$= \sigma_{\tilde{b}}\, \rho_{\tilde{a},\tilde{b}} \left(\frac{a - \mu_{\tilde{a}}}{\sigma_{\tilde{a}}} \right) + \mu_{\tilde{b}}. \tag{8.79}$$

■

Example 8.15 (Cats and dogs: Linear MMSE estimation)**.** In Example 8.7 we compute the means, variances and covariances of the number of cats and dogs from the joint pmf in Example 7.6. Here we use these statistics to obtain a linear estimate of the number of cats given the number of dogs. By Theorem 8.14, the linear MMSE estimator is

$$\ell_{\text{MMSE}}(d) := \sigma_{\tilde{c}}\, \rho_{\tilde{d},\tilde{c}} \left(\frac{d - \mu_{\tilde{d}}}{\sigma_{\tilde{d}}} \right) + \mu_{\tilde{c}} \tag{8.80}$$

$$= -0.3d + 0.755, \tag{8.81}$$

which yields the estimates

$$\ell_{\text{MMSE}}(0) = 0.755, \tag{8.82}$$

$$\ell_{\text{MMSE}}(1) = 0.455, \tag{8.83}$$

$$\ell_{\text{MMSE}}(2) = 0.155. \tag{8.84}$$

The corresponding MSE equals

$$\text{E}\left[\left(\tilde{c} - \ell_{\text{MMSE}}(\tilde{d}) \right)^2 \right] = \sum_{c=0}^{3} \sum_{d=0}^{2} p_{\tilde{c},\tilde{d}}(c, d)\, (c + 0.3d - 0.756)^2 \tag{8.85}$$

$$= 0.759. \tag{8.86}$$

The MSE is very close to, but larger than, the optimal MSE (0.756) achieved by the conditional-mean estimator derived in Example 7.60.

As illustrated by Example 8.15, the MSE of the linear minimum MSE estimator is lower bounded by the optimal MSE achieved by the nonlinear conditional-mean

estimator. For Gaussian random variables, linear estimation is optimal in terms of MSE, because the conditional mean function is linear.

Theorem 8.16 (Linear estimation is optimal for Gaussian random variables)**.** *Let $\tilde{a}$ and $\tilde{b}$ be random variables representing the response and the feature in a simple regression problem. We assume that $\tilde{a}$ and $\tilde{b}$ are jointly Gaussian, meaning that*

$$\begin{bmatrix} \tilde{a} \\ \tilde{b} \end{bmatrix} \tag{8.87}$$

is a Gaussian random vector with mean parameter

$$\mu := \begin{bmatrix} \mu_{\tilde{a}} \\ \mu_{\tilde{b}} \end{bmatrix} \tag{8.88}$$

and covariance-matrix parameter

$$\Sigma := \begin{bmatrix} \sigma_{\tilde{a}}^2 & \rho\sigma_{\tilde{a}}\sigma_{\tilde{b}} \\ \rho\sigma_{\tilde{a}}\sigma_{\tilde{b}} & \sigma_{\tilde{b}}^2 \end{bmatrix}, \tag{8.89}$$

where $\sigma_{\tilde{a}} > 0$, $\sigma_{\tilde{b}} > 0$, and $-1 < \rho < 1$ to ensure that Σ is full rank. The parametrization of the covariance matrix is without loss of generality; any positive definite symmetric matrix can be written like this. The minimum MSE (MMSE) estimator of $\tilde{b}$ given $\tilde{a} = a$ is the linear estimator

$$\ell(a) = \frac{\rho\sigma_{\tilde{b}}(a - \mu_{\tilde{a}})}{\sigma_{\tilde{a}}} + \mu_{\tilde{b}}. \tag{8.90}$$

Proof By Theorem 5.24, the conditional mean function of $\tilde{b}$ given $\tilde{a}$ is

$$\mu_{\tilde{b}\,|\,\tilde{a}}(a) = \frac{\rho\sigma_{\tilde{b}}(a - \mu_{\tilde{a}})}{\sigma_{\tilde{a}}} + \mu_{\tilde{b}}. \tag{8.91}$$

This is an affine function of $\tilde{a}$, so it must equal the linear MMSE estimate of $\tilde{b}$ given $\tilde{a}$, because the conditional mean is the optimal estimator in terms of MSE by Theorem 7.59. As a bonus, matching the estimator with the optimal linear estimator in Theorem 8.14 establishes that the parameter ρ is equal to the correlation coefficient between $\tilde{a}$ and $\tilde{b}$. ■

In practice, we perform regression using data. Assume that we have available n pairs of a feature and a corresponding response: (x_1, y_1), (x_2, y_2), ..., (x_n, y_n). We can approximate the linear minimum MSE estimator in Theorem 8.14 by replacing the means with sample means, the variances with sample variances and the correlation coefficient with the sample correlation coefficient. This yields

$$\beta_{\text{MMSE}} \approx \rho_{X,Y}\sqrt{\frac{v(Y)}{v(X)}}, \tag{8.92}$$

$$\alpha_{\text{MMSE}} \approx m(Y) - \rho_{X,Y}\sqrt{\frac{v(Y)}{v(X)}}m(X), \tag{8.93}$$

where $m(X)$ and $m(Y)$ are the sample means, $v(X)$ and $v(Y)$ the sample variances and $\rho_{X,Y}$ the sample correlation coefficient of the feature $X := \{x_1, x_2, \ldots, x_n\}$ and the response $Y := \{y_1, y_2, \ldots, y_n\}$.

Alternatively, we can fit an affine model to the data directly, approximating the response of each data point y_i by an affine function of the feature $\ell(x_i) := \beta x_i + \alpha$. If we select the affine function to minimize the residual sum of squares,

$$\text{RSS}\,(\beta, \alpha) := \sum_{i=1}^{n} (y_i - \beta x_i - \alpha)^2, \tag{8.94}$$

the resulting estimator is known as the *ordinary least-squares* estimator. It turns out that these two approaches are equivalent, as established in the following theorem.

Theorem 8.17 (Simple linear regression via ordinary least squares)**.** *Let* (x_1, y_1), (x_2, y_2), $\ldots$, (x_n, y_n) *be a dataset formed by* n *pairs of a single feature and the corresponding response. The ordinary least-squares estimator of the response when the feature equals* x *is*

$$\ell_{\text{OLS}}(x_i) := \beta_{\text{OLS}}\, x_i + \alpha_{\text{OLS}} \tag{8.95}$$

$$= \sqrt{v(Y)}\rho_{X,Y} \left(\frac{x - m(X)}{\sqrt{v(X)}} \right) + m(Y), \tag{8.96}$$

$$\tag{8.97}$$

where $\rho_{X,Y}$ *is the sample correlation coefficient of* X *and* Y*, and* $v(X)$ *and* $v(Y)$ *are the sample variances of* $X := \{x_1, x_2, \ldots, x_n\}$ *and* $Y := \{y_1, y_2, \ldots, y_n\}$*, respectively.*

The OLS estimator produces an optimal affine estimate, in the sense that it minimizes the sum of squared errors,

$$\beta_{\text{OLS}}, \alpha_{\text{OLS}} = \arg\min_{\beta,\alpha} \sum_{i=1}^{n} (y_i - \beta x_i - \alpha)^2. \tag{8.98}$$

Proof We follow a similar argument to the proof of Theorem 8.14. We denote the residual sum of squares, as a function of the linear coefficient β and the additive constant α, by

$$\text{RSS}(\beta, \alpha) := \sum_{i=1}^{n} (y_i - \beta x_i - \alpha)^2. \tag{8.99}$$

We denote by $\alpha^*(\beta)$ the optimal value of α for a fixed value of β. Equivalently, $\alpha^*(\beta)$ is the best constant estimate of $y_i - \beta x_i$, $1 \leq i \leq n$, in terms of the sum of

squared errors. By Theorem 7.31,

$$\alpha^*(\beta) := \arg\min_{\alpha} \sum_{i=1}^{n} (y_i - \beta x_i - \alpha)^2 \tag{8.100}$$

$$= \frac{1}{n} \sum_{i=1}^{n} (y_i - \beta x_i) \tag{8.101}$$

$$= m(Y) - \beta m(X). \tag{8.102}$$

Consequently, for any β and any α $\mathrm{RSS}(\beta, \alpha) \geq \mathrm{RSS}(\beta, \alpha^*(\beta))$. To obtain β_{OLS}, we fix α to $\alpha^*(\beta)$ and minimize the sum of squared errors,

$$\beta_{\mathrm{OLS}} = \arg\min_{\beta} \mathrm{RSS}(\beta, \alpha^*(\beta)). \tag{8.103}$$

We have

$$\mathrm{RSS}(\beta, \alpha^*(\beta)) \tag{8.104}$$

$$= \sum_{i=1}^{n} (y_i - \beta x_i - m(Y) + \beta m(X))^2 \tag{8.105}$$

$$= \sum_{i=1}^{n} (y_i - m(Y))^2 + \beta^2 \sum_{i=1}^{n} (x_i - m(X))^2 - 2\beta \sum_{i=1}^{n} (y_i - m(Y))(x_i - m(X))$$

$$= (n-1)\left(v(Y) + \beta^2 v(X) - 2\beta c(X,Y)\right). \tag{8.106}$$

The derivatives with respect to β equal

$$\frac{\mathrm{d}\,\mathrm{RSS}(\beta, \alpha^*(\beta))}{\mathrm{d}\beta} = 2(n-1)\left(v(X)\beta - c(X,Y)\right), \tag{8.107}$$

$$\frac{\mathrm{d}^2\,\mathrm{RSS}(\beta, \alpha^*(\beta))}{\mathrm{d}\beta^2} = 2(n-1)v(X) \geq 0. \tag{8.108}$$

The function is strictly convex, as long as the sample variance of the feature is nonzero, so we can obtain β_{OLS} by setting the derivative to zero,

$$\beta_{\mathrm{OLS}} = \frac{c(X,Y)}{v(X)} \tag{8.109}$$

$$= \rho_{X,Y} \sqrt{\frac{v(Y)}{v(X)}}. \tag{8.110}$$

The corresponding value of $\alpha^*(\beta)$ is

$$\alpha_{\mathrm{OLS}} = \alpha^*(\beta_{\mathrm{OLS}}) \tag{8.111}$$

$$= m(Y) - \rho_{X,Y} \sqrt{\frac{v(Y)}{v(X)}} m(X). \tag{8.112}$$

■

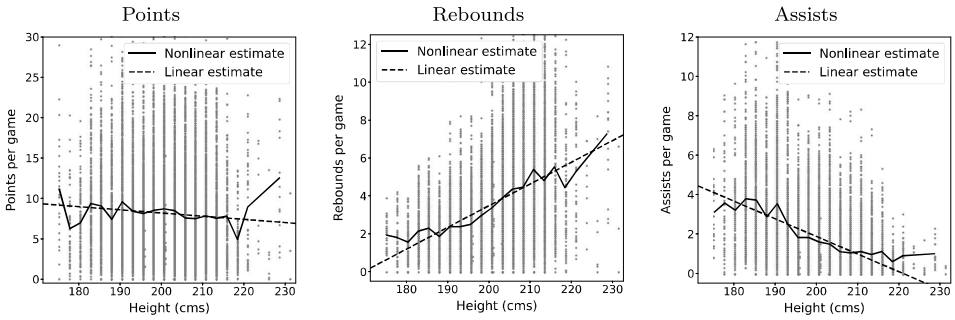

Figure 8.4 Comparison of linear and nonlinear regression estimates. Estimate of the points (left), rebounds (center), and assists (right) per game of NBA players between 1996 and 2019 based on their height. The graphs show the ordinary least-squares linear estimate (dashed line) and the sample conditional mean (solid line), superposed on a scatterplot of the data. The linear estimate reflects the overall relationship between each response (points, rebounds or assists) and the height, but cannot capture nonlinear structure that is reflected in the conditional mean. For example, the nonlinear estimate of assists is approximately constant for players shorter than 190 cm.

Example 8.18 (Comparing linear and nonlinear regression)**.** We consider the problem of estimating points, rebounds and assists of NBA players as a function of their height, using the same data as in Example 8.13. We compare the linear ordinary least-squares estimator defined in Theorem 8.17 with a nonlinear estimator based on the sample conditional mean (see Definition 7.47).

The results are shown in Figure 8.4. The conditional mean is more flexible than the linear estimate. As a result, it is much noisier in regions where we don't have a lot of data. However, this flexibility also means that it can capture nonlinear structure. For example, the general trend is for assists to be inversely correlated to height, because shorter players tend to be responsible for creating scoring opportunities for teammates. This is captured by the linear model. However, the conditional mean of assists per game is roughly constant for players under 190 cm. This makes sense: if we only consider short players, then height is no longer indicative of their role (all of them are point guards). Consequently, height is not inversely correlated with assists for such players. This pattern is captured by the nonlinear conditional-mean estimator, but not by the linear model, because it cannot be expressed in terms of linear dependence.

8.5 Properties Of The Correlation Coefficient

In this section we establish three important properties of the correlation coefficient. Similar arguments can be applied to show that analogous properties hold for the sample correlation coefficient (see Exercises 8.8 and 8.9). Section 8.5.1

proves that the correlation coefficient $\rho_{\tilde{a},\tilde{b}}$ between two random variables $\tilde{a}$ and $\tilde{b}$ is always bounded between -1 and 1. Section 8.5.2 establishes that when $\rho_{\tilde{a},\tilde{b}}$ equals -1 or 1, this implies a linear relationship between $\tilde{a}$ and $\tilde{b}$. Section 8.5.3 shows that the squared correlation coefficient is equal to the fraction of variance in $\tilde{b}$ that can be explained linearly in terms of $\tilde{a}$ (and vice versa). These three properties are all consequences of the following theorem, which provides a formula for the MSE of the linear MMSE estimator as a function of the correlation coefficient.

Theorem 8.19 (Mean squared error of linear MMSE estimate)**.** *Let $\tilde{a}$ and $\tilde{b}$ be two random variables with finite variance, belonging to the same probability space. Let $\ell_{\text{MMSE}}(\tilde{a})$ be the linear MMSE estimator of $\tilde{b}$ given $\tilde{a}$, as defined in Theorem 8.14. The MSE of this estimator equals*

$$\mathrm{E}\left[(\tilde{b} - \ell_{\text{MMSE}}(\tilde{a}))^2\right] = \left(1 - \rho^2_{\tilde{a},\tilde{b}}\right) \sigma^2_{\tilde{b}}, \tag{8.113}$$

where $\sigma^2_{\tilde{b}}$ is the variance of $\tilde{b}$ and $\rho_{\tilde{a},\tilde{b}}$ the correlation coefficient of $\tilde{a}$ and $\tilde{b}$.

Proof Let $s(\tilde{a})$ and $s(\tilde{b})$ denote the standardized counterparts of $\tilde{a}$ and $\tilde{b}$. By Theorem 8.14 and Lemma 8.3

$$\begin{aligned}
\mathrm{E}\left[(\ell_{\text{MMSE}}(\tilde{a}) - \tilde{b})^2\right] &= \mathrm{E}\left[(\rho_{\tilde{a},\tilde{b}}\, \sigma_{\tilde{b}} s(\tilde{a}) + \mu_{\tilde{b}} - \tilde{b})^2\right] && (8.114)\\
&= \sigma^2_{\tilde{b}}\ \mathrm{E}\left[(\rho_{\tilde{a},\tilde{b}} s(\tilde{a}) - s(\tilde{b}))^2\right] && (8.115)\\
&= \sigma^2_{\tilde{b}} \left(\rho^2_{\tilde{a},\tilde{b}} \mathrm{E}[s(\tilde{a})^2] + \mathrm{E}[s(\tilde{b})^2] - 2\rho_{\tilde{a},\tilde{b}} \mathrm{E}\left[s(\tilde{a}) s(\tilde{b})\right]\right) && (8.116)\\
&= \sigma^2_{\tilde{b}} \left(1 - \rho^2_{\tilde{a},\tilde{b}}\right). && (8.117)
\end{aligned}$$

■

8.5.1 The Correlation Coefficient Is Bounded

The following lemma establishes that the magnitude of the correlation coefficient cannot be greater than one.

Theorem 8.20. *The coefficient $\rho_{\tilde{a},\tilde{b}}$ of any random variables $\tilde{a}$ and $\tilde{b}$ with bounded variance satisfies*

$$-1 \leq \rho_{\tilde{a},\tilde{b}} \leq 1. \tag{8.118}$$

Proof By Theorem 8.19

$$\sigma^2_{\tilde{b}} \left(1 - \rho^2_{\tilde{a},\tilde{b}}\right) = \mathrm{E}\left[(\tilde{b} - \ell_{\text{MMSE}}(\tilde{a}))^2\right] \geq 0, \tag{8.119}$$

where $\ell_{\text{MMSE}}(\tilde{a})$ is the linear MMSE estimate of $\tilde{b}$ given $\tilde{a}$. The inequality follows from the fact that the expected value of a nonnegative quantity is nonnegative, as it is just a sum or integral of nonnegative values. Since $\sigma^2_{\tilde{b}} \geq 0$, if $\sigma_{\tilde{b}} \neq 0$, this implies $0 \leq \rho^2_{\tilde{a},\tilde{b}} \leq 1$. If $\sigma^2_{\tilde{b}} = 0$, then $\tilde{b}$ is equal to its mean with probability one by Corollary 9.20, and the correlation coefficient is therefore zero, because $s(\tilde{b}) = 0$. ■

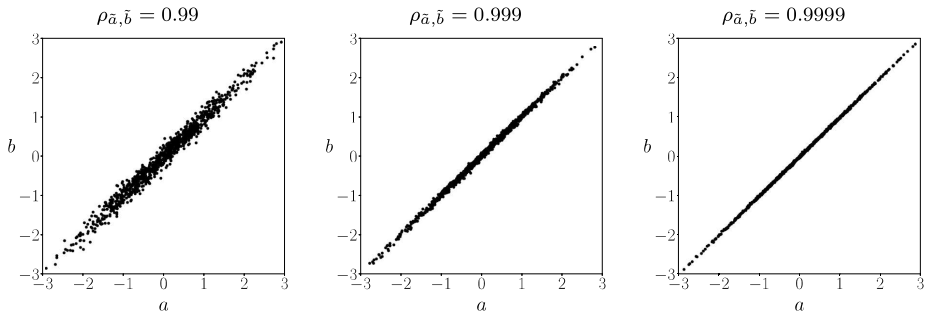

Figure 8.5 Complete linear dependence. The plots show 1,000 i.i.d. samples from two Gaussian random variables $\tilde{a}$ and $\tilde{b}$ with zero mean, unit variance and different correlation coefficients. As the correlation coefficient approaches one, the samples lie increasingly closer to a diagonal line, which implies that $\tilde{b}$ essentially equals $\tilde{a}$, as predicted by Theorem 8.21.

8.5.2 Complete Linear Dependence

If the correlation coefficient between two random variables equals -1 or 1, then the random variables can be expressed *exactly* as an affine function of each other with probability one. The reason is that the residual of the linear MMSE estimate has zero variance.

Theorem 8.21 (Complete linear dependence)**.** *Let $\tilde{a}$ and $\tilde{b}$ be two random variables with bounded variance belonging to the same probability space. If $\rho_{\tilde{a},\tilde{b}} = 1$ or $\rho_{\tilde{a},\tilde{b}} = -1$, then $\tilde{b}$ is an affine function of $\tilde{a}$ with probability one. The affine function is the linear MMSE estimate of $\tilde{b}$ given $\tilde{a}$, denoted by $\ell_{\text{MMSE}}(\tilde{a})$:*

$$\mathrm{P}(\tilde{b} = \ell_{\text{MMSE}}(\tilde{a})) = 1. \tag{8.120}$$

Proof By Theorem 8.19, if $|\rho_{\tilde{a},\tilde{b}}| = 1$, the mean squared error of the linear MMSE estimate is zero,

$$\mathrm{E}\left[(\ell_{\text{MMSE}}(\tilde{a}) - \tilde{b})^2\right] = \left(1 - \rho^2_{\tilde{a},\tilde{b}}\right)\sigma^2_{\tilde{b}} = 0, \tag{8.121}$$

which implies (8.120) by Corollary 9.20. ■

Figure 8.5 illustrates the result by showing the scatterplots of samples from Gaussian random variables with different correlation coefficients. As the correlation coefficient approaches one, the random variables become completely linearly dependent.

8.5.3 Explained Variance

In this section we provide an interpretation of the correlation coefficient in terms of explained variance. We will need the following result, which expresses the variance of a sum of random variables in terms of their variances and covariance.

Theorem 8.22 (Variance of the sum of two random variables)**.** *The variance of the sum of two random variables $\tilde{a}$ and $\tilde{b}$ with finite variance equals*

$$\operatorname{Var}[\tilde{a}+\tilde{b}] = \operatorname{Var}[\tilde{a}] + \operatorname{Var}[\tilde{b}] + 2\operatorname{Cov}[\tilde{a},\tilde{b}]. \tag{8.122}$$

Proof The result follows from linearity of expectation:

$$\begin{aligned}\operatorname{Var}[\tilde{a}+\tilde{b}] &= \mathrm{E}\left[(\tilde{a}+\tilde{b}-\mathrm{E}[\tilde{a}+\tilde{b}])^2\right] &(8.123)\\ &= \mathrm{E}\left[(\tilde{a}-\mathrm{E}[\tilde{a}])^2\right] + \mathrm{E}\left[(\tilde{b}-\mathrm{E}[\tilde{b}])^2\right] + 2\mathrm{E}\left[(\tilde{a}-\mathrm{E}[\tilde{a}])\,(\tilde{b}-\mathrm{E}[\tilde{b}])\right]\\ &= \operatorname{Var}[\tilde{a}] + \operatorname{Var}[\tilde{b}] + 2\operatorname{Cov}[\tilde{a},\tilde{b}]. &(8.124)\end{aligned}$$

■

If two random variables are positively correlated, then their fluctuations reinforce each other. Conversely, if they are negatively correlated, the fluctuations cancel out. If they are uncorrelated, then the variance of their sum equals the sum of their variances.

Corollary 8.23. *If $\tilde{a}$ and $\tilde{b}$ are uncorrelated, then*

$$\operatorname{Var}[\tilde{a}+\tilde{b}] = \operatorname{Var}[\tilde{a}] + \operatorname{Var}[\tilde{b}]. \tag{8.125}$$

Consider the decomposition of $\tilde{b}$ into the sum of the linear MMSE estimator $\ell_{\mathrm{MMSE}}(\tilde{a})$ and the corresponding residual

$$\tilde{b} = \underbrace{\ell_{\mathrm{MMSE}}(\tilde{a})}_{\text{Linear MMSE estimator}} + \underbrace{\tilde{b}-\ell_{\mathrm{MMSE}}(\tilde{a})}_{\text{Residual}}, \tag{8.126}$$

The residual is uncorrelated with $\tilde{a}$, which makes sense: if they were correlated, then we could use an affine function of $\tilde{a}$ to approximate the residual and improve the linear MMSE estimator. However, this is impossible, because the linear MMSE estimator is the optimal affine estimator.

Lemma 8.24 (Uncorrelated residual)**.** *Let $\tilde{a}$ and $\tilde{b}$ be two random variables with bounded variance belonging to the same probability space. The residual of the linear MMSE estimator $\ell_{\mathrm{MMSE}}(\tilde{a})$ of $\tilde{b}$ given $\tilde{a}$, defined in Theorem 8.14, has zero mean and is uncorrelated with $\tilde{a}$.*

Proof By linearity of expectation

$$\begin{aligned}\mathrm{E}\left[\tilde{b}-\ell_{\mathrm{MMSE}}(\tilde{a})\right] &= \mathrm{E}\left[\tilde{b}-\sigma_{\tilde{b}}\,\rho_{\tilde{a},\tilde{b}}\left(\frac{\tilde{a}-\mu_{\tilde{a}}}{\sigma_{\tilde{a}}}\right)-\mu_{\tilde{b}}\right] &(8.127)\\ &= \mu_{\tilde{b}}-\mu_{\tilde{b}}-\sigma_{\tilde{b}}\,\rho_{\tilde{a},\tilde{b}}\left(\frac{\mu_{\tilde{a}}-\mu_{\tilde{a}}}{\sigma_{\tilde{a}}}\right)=0. &(8.128)\end{aligned}$$

Consequently, by Theorem 8.14,

$$\begin{aligned}
\mathrm{Cov}\left[\tilde{a}, \tilde{b} - \ell_{\mathrm{MMSE}}(\tilde{a})\right] &= \mathrm{E}\left[(\tilde{a} - \mu_{\tilde{a}})\left(\tilde{b} - \sigma_{\tilde{b}}\,\rho_{\tilde{a},\tilde{b}}\left(\frac{\tilde{a} - \mu_{\tilde{a}}}{\sigma_{\tilde{a}}}\right) - \mu_{\tilde{b}}\right)\right] && (8.129)\\
&= \sigma_{\tilde{a}}\,\sigma_{\tilde{b}}\mathrm{E}\left[s(\tilde{a})(s(\tilde{b}) - \rho_{\tilde{a},\tilde{b}}s(\tilde{a}))\right] && (8.130)\\
&= \sigma_{\tilde{a}}\,\sigma_{\tilde{b}}\left(\mathrm{E}[s(\tilde{a})s(\tilde{b})] - \rho_{\tilde{a},\tilde{b}}\mathrm{E}[s(\tilde{a})^2]\right) && (8.131)\\
&= \sigma_{\tilde{a}}\,\sigma_{\tilde{b}}(\rho_{\tilde{a},\tilde{b}} - \rho_{\tilde{a},\tilde{b}}) && (8.132)\\
&= 0, && (8.133)
\end{aligned}$$

since $\mathrm{E}[s(\tilde{a})^2] = \mathrm{Var}\left[s(\tilde{a})\right] = 1$ by Lemma 8.3 and $\mathrm{E}[s(\tilde{a})s(\tilde{b})] = \rho_{\tilde{a},\tilde{b}}$, where $s(\tilde{a})$ and $s(\tilde{b})$ denote the standardized counterparts of $\tilde{a}$ and $\tilde{b}$. ■

Lemma 8.24 establishes that the two components in (8.126) are uncorrelated. Therefore, the variance of $\tilde{b}$ is equal to the sum of their variances by Corollary 8.23, which yields a variance decomposition that is directly tied to the correlation coefficient. Figures 8.6 and 8.7 illustrate the decomposition.

Theorem 8.25 (Variance decomposition)**.** *Let $\tilde{a}$ and $\tilde{b}$ be two random variables with bounded variance belonging to the same probability space. The variance of $\tilde{b}$ can be decomposed into the sum of the variance of the linear MMSE estimator $\ell_{\mathrm{MMSE}}(\tilde{a})$ of $\tilde{b}$ given $\tilde{a}$ and the corresponding residual,*

$$\mathrm{Var}\left[\tilde{b}\right] = \mathrm{Var}\left[\ell_{\mathrm{MMSE}}(\tilde{a})\right] + \mathrm{Var}\left[\tilde{b} - \ell_{\mathrm{MMSE}}(\tilde{a})\right]. \tag{8.134}$$

The fraction of variance corresponding to the linear estimator is equal to the squared correlation coefficient $\rho_{\tilde{a},\tilde{b}}$ of $\tilde{a}$ and $\tilde{b}$:

$$\begin{aligned}
\mathrm{Var}\left[\ell_{\mathrm{MMSE}}(\tilde{a})\right] &= \rho^2_{\tilde{a},\tilde{b}}\mathrm{Var}\left[\tilde{b}\right], && (8.135)\\
\mathrm{Var}[\tilde{b} - \ell_{\mathrm{MMSE}}(\tilde{a})] &= (1 - \rho^2_{\tilde{a},\tilde{b}})\mathrm{Var}\left[\tilde{b}\right]. && (8.136)
\end{aligned}$$

Proof By Lemma 8.24 the residual and $\tilde{a}$ are uncorrelated. This immediately implies that the residual is also uncorrelated with any affine function of $\tilde{a}$ (see Exercise 8.2), and in particular with $\ell_{\mathrm{MMSE}}(\tilde{a})$. The variance decomposition in (8.134) then follows from Corollary 8.23.

Again, by Lemma 8.24, the mean of the residual is zero, so its variance is equal to the MSE of the linear MMSE estimator derived in Theorem 8.19:

$$\begin{aligned}
\mathrm{Var}[\tilde{b} - \ell_{\mathrm{MMSE}}(\tilde{a})] &= \mathrm{E}\left[(\tilde{b} - \ell_{\mathrm{MMSE}}(\tilde{a}))^2\right] && (8.137)\\
&= (1 - \rho^2_{\tilde{a},\tilde{b}})\mathrm{Var}\left[\tilde{b}\right]. && (8.138)
\end{aligned}$$

Consequently, by (8.138)

$$\begin{aligned}
\mathrm{Var}\left[\ell_{\mathrm{MMSE}}(\tilde{a})\right] &= \mathrm{Var}[\tilde{b}] - \mathrm{Var}[\tilde{b} - \ell_{\mathrm{MMSE}}(\tilde{a})] && (8.139)\\
&= \rho^2_{\tilde{a},\tilde{b}}\mathrm{Var}\left[\tilde{b}\right]. && (8.140)
\end{aligned}$$

■

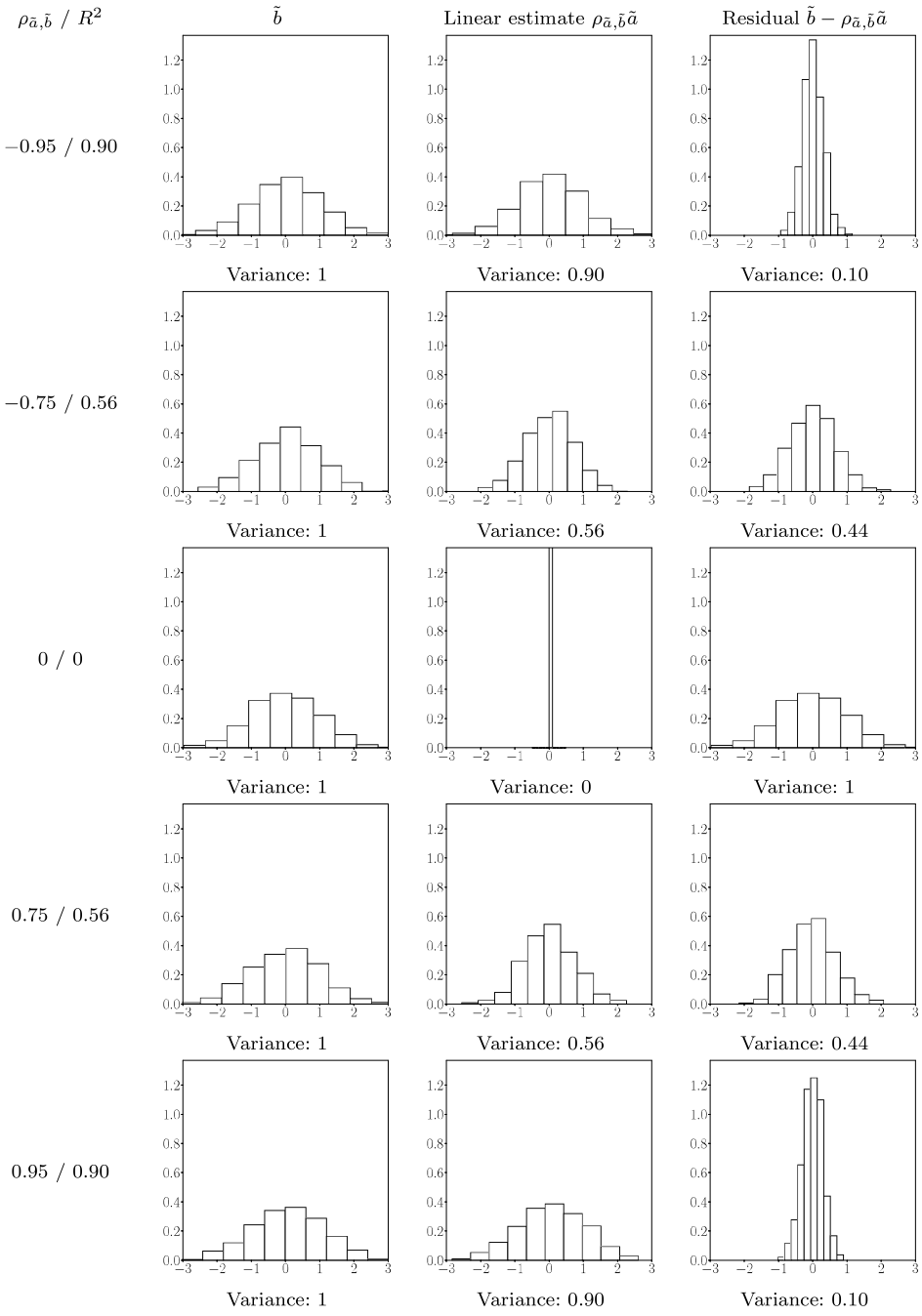

Figure 8.6 Decomposition of variance and coefficient of determination. The figure depicts the different terms in the variance decomposition of Theorem 8.25 for the two Gaussian random variables $\tilde{a}$ and $\tilde{b}$ in Figure 8.1. The left column shows the histogram of $\tilde{b}$. The central column shows the histogram of the linear MMSE estimate of $\tilde{b}$ given $\tilde{a}$, $\rho_{\tilde{a},\tilde{b}}\tilde{a}$. The right column shows the histogram of the residual $\tilde{b} - \rho_{\tilde{a},\tilde{b}}\tilde{a}$. The coefficient of determination, which captures the fraction of variance explained by the linear estimator, equals the square of the correlation coefficient $\rho^2_{\tilde{a},\tilde{b}}$, as established in Theorem 8.27.

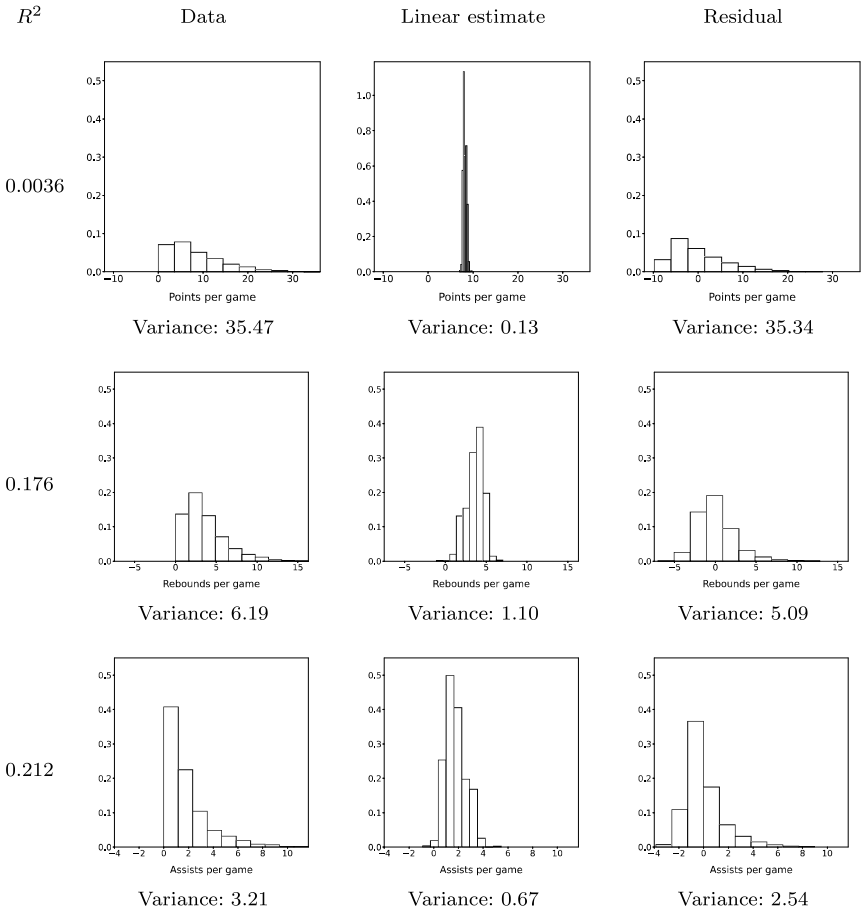

Figure 8.7 Decomposition of variance and coefficient of determination for NBA data. The figure depicts the different terms in the variance decomposition of Theorem 8.25 for the NBA data in Figure 8.2. The left column shows the histogram of the points (top), rebounds (center) and assists (bottom) per game. The central column shows the histogram of the linear MMSE estimate of each stat given height. The right column shows the histogram of the corresponding residual. The coefficient of determination, which captures the fraction of variance explained by the linear estimator, equals $\rho^2_{\tilde{a},\tilde{b}}$, as established in Theorem 8.27.

A popular metric to evaluate the linear MMSE estimation is the fraction of response variance *explained* by the estimator. This metric is known as the *coefficient of determination* R^2.

Definition 8.26 (Coefficient of determination in simple linear regression)**.** *Let $\tilde{a}$ and $\tilde{b}$ be random variables representing the feature and the response of a regression problem, and let ℓ_{MMSE} be the linear minimum MSE estimator of $\tilde{b}$ given $\tilde{a}$. The*

coefficient of determination is the ratio between the variance of the linear MMSE estimator and the variance of the response,

$$R^2 := \frac{\text{Var}\left[\ell_{\text{MMSE}}(\tilde{a})\right]}{\text{Var}[\tilde{b}]}. \tag{8.141}$$

By Theorem 8.25 the coefficient of determination equals the square of the correlation coefficient, and can also be expressed in terms of the MSE. By Theorem 8.20 it is bounded between zero and one.

Theorem 8.27 (Properties of the coefficient of determination)**.** *Let $\tilde{a}$ and $\tilde{b}$ be random variables representing the feature and the response of a regression problem, and let ℓ_{MMSE} be the linear minimum MSE estimator of $\tilde{b}$ given $\tilde{a}$. The coefficient of determination equals*

$$R^2 = \rho^2_{\tilde{a},\tilde{b}} \tag{8.142}$$

$$= 1 - \frac{\text{MSE}}{\text{Var}[\tilde{b}]}, \qquad \text{MSE} := \text{E}\left[(\tilde{b} - \ell_{\text{MMSE}}(\tilde{a}))^2\right], \tag{8.143}$$

where $\rho_{\tilde{a},\tilde{b}}$ is the correlation coefficient between $\tilde{a}$ and $\tilde{b}$. In addition, R^2 is bounded between zero and one:

$$0 \leq R^2 \leq 1. \tag{8.144}$$

Proof By (8.135) in Theorem 8.25

$$R^2 := \frac{\text{Var}[\ell_{\text{MMSE}}(\tilde{a})]}{\text{Var}[\tilde{b}]} = \rho^2_{\tilde{a},\tilde{b}}. \tag{8.145}$$

Consequently, by Theorem 8.20, $0 \leq R^2 \leq 1$.

By (8.134) in Theorem 8.25

$$R^2 = \frac{\text{Var}\left[\tilde{b}\right] - \text{Var}[\tilde{b} - \ell_{\text{MMSE}}(\tilde{a})]}{\text{Var}[\tilde{b}]} \tag{8.146}$$

$$= 1 - \frac{\text{Var}[\tilde{b} - \ell_{\text{MMSE}}(\tilde{a})]}{\text{Var}[\tilde{b}]} \tag{8.147}$$

$$= 1 - \frac{\text{E}\left[(\tilde{b} - \ell_{\text{MMSE}}(\tilde{a}))^2\right]}{\text{Var}[\tilde{b}]}, \tag{8.148}$$

where the last equality holds because the mean of the residual $\tilde{b} - \ell_{\text{MMSE}}(\tilde{a})$ is zero by Lemma 8.24. ■

As illustrated by Figures 8.6 and 8.7, R^2 provides a normalized evaluation of the linear MMSE estimator. When R^2 is close to one, the estimator explains most of the variance in the response, because the feature and the response are highly correlated, so the MSE is close to zero. Conversely, when R^2 is close to zero, the estimator explains almost no variance, because the feature and the response are almost uncorrelated.

8.6 Uncorrelation And Independence

If two random variables are independent, they are also uncorrelated. This makes sense, since lack of dependence necessarily implies lack of linear dependence.

Lemma 8.28 (Independence implies uncorrelation)**.** *If two random variables are independent, then they are uncorrelated.*

Proof By Theorem 7.19, if two random variables $\tilde{a}$ and $\tilde{b}$ are independent

$$\begin{aligned} \mathrm{Cov}[\tilde{a}, \tilde{b}] &= \mathrm{E}[\tilde{a}\tilde{b}] - \mathrm{E}\left[\tilde{a}\right] \mathrm{E}[\tilde{b}] && (8.149) \\ &= \mathrm{E}\left[\tilde{a}\right] \mathrm{E}[\tilde{b}] - \mathrm{E}\left[\tilde{a}\right] \mathrm{E}[\tilde{b}] && (8.150) \\ &= 0. && (8.151) \end{aligned}$$

■

In the case of Gaussian random variables, uncorrelation implies independence. This establishes that dependence between Gaussian random variables is purely linear.

Theorem 8.29 (Uncorrelation implies independence for Gaussian random variables)**.** *Let $(\tilde{a}, \tilde{b})$ be a Gaussian random vector. If $\tilde{a}$ and $\tilde{b}$ are uncorrelated, then they are also independent.*

Proof Without loss of generality, we parametrize the covariance matrix of the Gaussian random vector as

$$\Sigma := \begin{bmatrix} \sigma_{\tilde{a}}^2 & \rho\sigma_{\tilde{a}}\sigma_{\tilde{b}} \\ \rho\sigma_{\tilde{a}}\sigma_{\tilde{b}} & \sigma_{\tilde{b}}^2 \end{bmatrix}. \tag{8.152}$$

In the proof of Theorem 8.16, we show that the correlation coefficient of $\tilde{a}$ and $\tilde{b}$ is equal to ρ, so if ρ is zero, then the covariance-matrix parameter is diagonal and its inverse equals

$$\Sigma^{-1} = \begin{bmatrix} \frac{1}{\sigma_{\tilde{a}}^2} & 0 \\ 0 & \frac{1}{\sigma_{\tilde{b}}^2} \end{bmatrix}. \tag{8.153}$$

In terms of the standardized variables

$$s(a) := \frac{a - \mu_{\tilde{a}}}{\sigma_{\tilde{a}}}, \qquad s(b) := \frac{b - \mu_{\tilde{b}}}{\sigma_{\tilde{b}}}, \tag{8.154}$$

we have

$$\begin{aligned} f_{\tilde{a},\tilde{b}}(a, b) &= \frac{1}{2\pi\sqrt{|\Sigma|}} \exp\left(-\frac{1}{2} \begin{bmatrix} a - \mu_{\tilde{a}} \\ b - \mu_{\tilde{b}} \end{bmatrix}^T \Sigma^{-1} \begin{bmatrix} a - \mu_{\tilde{a}} \\ b - \mu_{\tilde{b}} \end{bmatrix} \right) && (8.155) \\ &= \frac{1}{2\pi\sigma_{\tilde{a}}\sigma_{\tilde{b}}} \exp\left(-\frac{s(a)^2 + s(b)^2}{2} \right) && (8.156) \\ &= \frac{1}{\sqrt{2\pi}\sigma_{\tilde{a}}} \exp\left(-\frac{s(a)^2}{2} \right) \frac{1}{\sqrt{2\pi}\sigma_{\tilde{b}}} \exp\left(-\frac{s(b)^2}{2} \right) && (8.157) \\ &= f_{\tilde{a}}(a) f_{\tilde{b}}(b). && (8.158) \end{aligned}$$

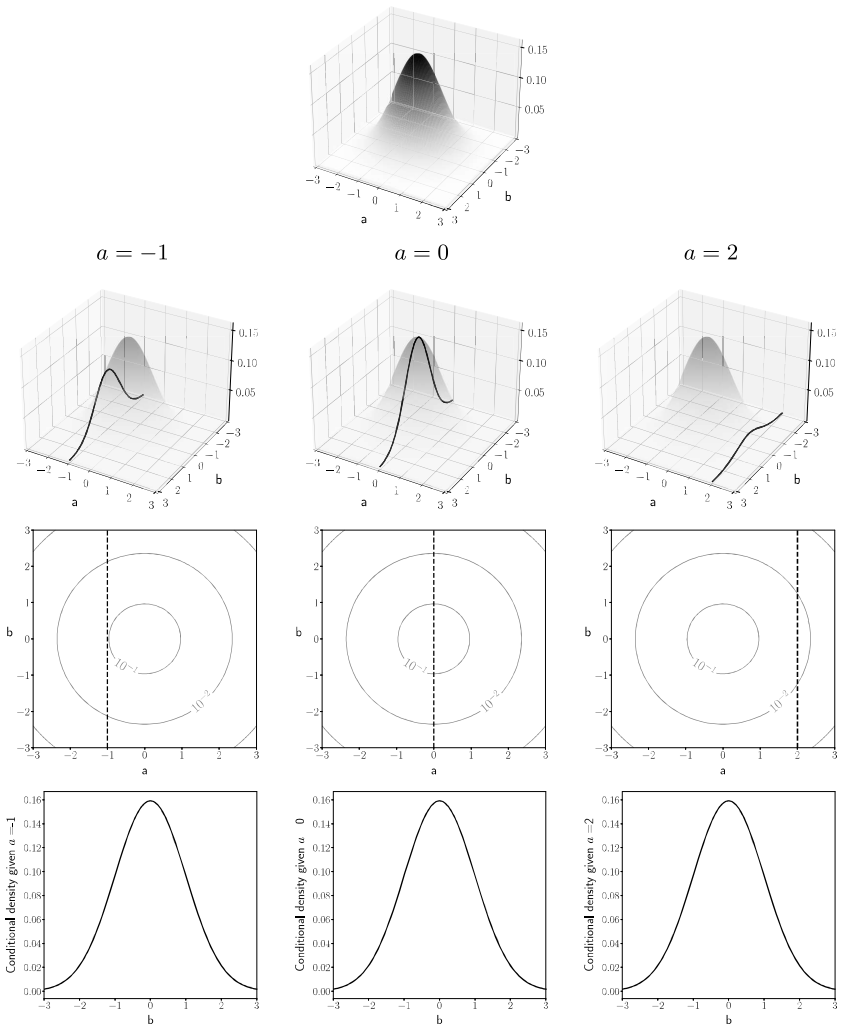

Figure 8.8 Uncorrelated Gaussian random variables are independent. The first row shows the joint pdf of two uncorrelated Gaussian random variables. In the second row we see the slices of the density corresponding to different values of a. The third row shows the position of the slices on the contour plot of the joint pdf. The fourth row shows the corresponding conditional pdf of $\tilde{b}$ given $\tilde{a} = a$, for the different values of a. They are all the same because a and b are independent, as established in Theorem 8.29.

The joint pdf factors into the product of the marginal pdfs, so the two random variables are independent. ■

Figure 8.8 shows the joint pdf of two uncorrelated Gaussian random variables.

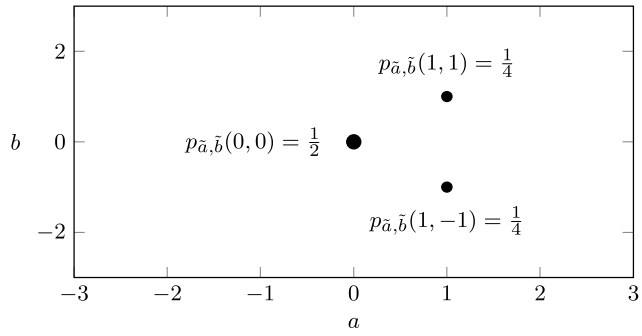

Figure 8.9 Uncorrelation does not imply independence. Joint probability mass function of the random variables in Example 8.30, which are uncorrelated but not independent.

The variables are independent, so all the conditional pdfs are the same. For general random variables, uncorrelation does not imply independence, as demonstrated by the following two examples.

Example 8.30 (Uncorrelation does not imply independence)**.** Let $\tilde{a}$ and $\tilde{b}$ be two random variables with the following joint probability mass function, depicted in Figure 8.9,

$$p_{\tilde{a},\tilde{b}}(0,0) = \frac{1}{2}, \qquad p_{\tilde{a},\tilde{b}}(1,-1) = \frac{1}{4}, \qquad p_{\tilde{a},\tilde{b}}(1,1) = \frac{1}{4}. \tag{8.159}$$

The two random variables are uncorrelated, since

$$\begin{aligned} \mathrm{E}[\tilde{b}] &= \sum_{a=0}^{1} \sum_{b=-1}^{1} b\, p_{\tilde{a},\tilde{b}}(a,b) && (8.160)\\ &= 0 \cdot \frac{1}{2} - 1 \cdot \frac{1}{4} + 1 \cdot \frac{1}{4} = 0, && (8.161)\\ \mathrm{Cov}[\tilde{a},\tilde{b}] &= \mathrm{E}[\tilde{a}\tilde{b}] - \mathrm{E}[\tilde{a}]\mathrm{E}[\tilde{b}] && (8.162)\\ &= \mathrm{E}[\tilde{a}\tilde{b}] && (8.163)\\ &= \sum_{a=0}^{1} \sum_{b=-1}^{1} ab\, p_{\tilde{a},\tilde{b}}(a,b) && (8.164)\\ &= 0 \cdot \frac{1}{2} - 1 \cdot \frac{1}{4} + 1 \cdot \frac{1}{4} = 0. && (8.165) \end{aligned}$$

However, they are not independent. The conditional pmf of $\tilde{b}$ given $\tilde{a} = 0$ is

$$p_{\tilde{b}\,|\,\tilde{a}}(b\,|\,0) = \begin{cases} 1 & \text{if } b = 0,\\ 0 & \text{otherwise.} \end{cases} \tag{8.166}$$

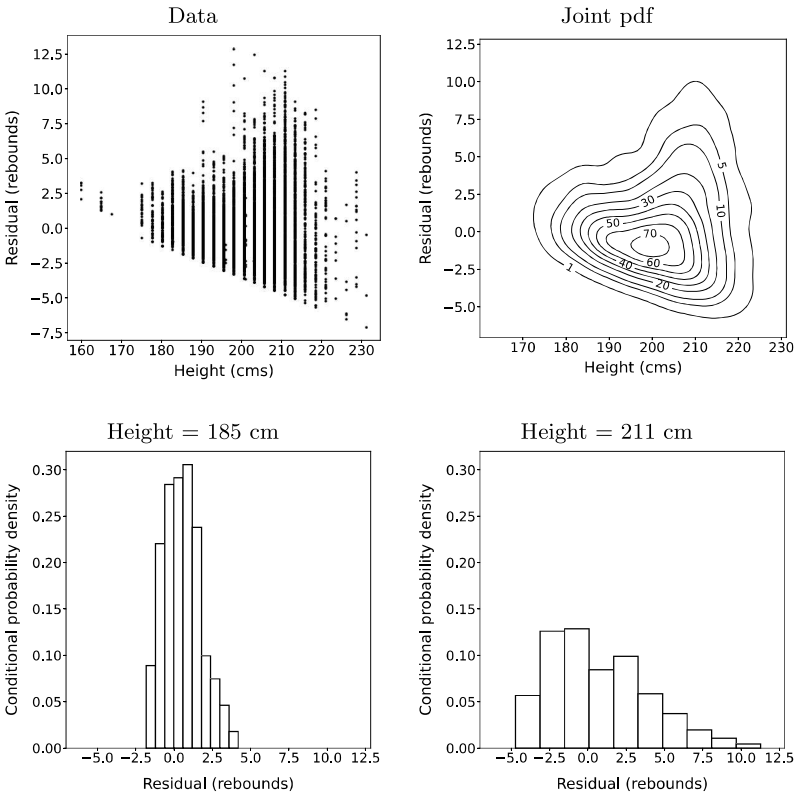

Figure 8.10 Uncorrelated residuals are not necessarily independent. The top row shows a scatterplot and a kernel density estimate of the residual of the OLS estimator of rebounds given height in Example 8.31. The bottom row shows histograms of the residual for two different heights. They are very different, demonstrating that the residual and height are not independent, even though their sample correlation coefficient is zero.

This is very different to the conditional pmf of $\tilde{b}$ given $\tilde{a} = 1$, which equals

$$p_{\tilde{b}\,|\,\tilde{a}}(b\,|\,1) = \begin{cases} 0.5 & \text{if } b = -1, \\ 0.5 & \text{if } b = 1, \\ 0 & \text{otherwise.} \end{cases} \tag{8.167}$$

Interestingly, the MMSE estimate of $\tilde{b}$ given $\tilde{a}$ actually coincides with the linear MMSE estimate, because the conditional mean of $\tilde{b}$ given $\tilde{a} = a$ is zero for both values of a. Nevertheless, the random variables are dependent because the conditional distribution of $\tilde{b}$ given $\tilde{a} = a$ is different for $a = 0$ and $a = 1$.

Example 8.31 (Uncorrelated OLS residual)**.** When we perform simple linear

regression via ordinary least squares (OLS), the residual is uncorrelated with the feature. Consider a dataset of real-valued pairs (x_1, y_1), (x_2, y_2), ..., (x_n, y_n), $1 \leq i \leq n$, and the corresponding OLS estimate of y_i given x_i by $\ell_{\text{OLS}}(x_i)$, defined in Theorem 8.17. The sample correlation coefficient between the features x_i and the corresponding residual

$$r_i := y_i - \ell_{\text{OLS}}(x_i), \qquad 1 \leq i \leq n, \tag{8.168}$$

is zero. We can prove this formally using a similar argument to the proof of Lemma 8.24 (see Exercise 8.9). It makes sense for the feature and the residual to be uncorrelated. Otherwise, we would be able to obtain a linear estimate of the residual which would allow us to improve the OLS estimate. However, this is impossible because Theorem 8.17 establishes that the OLS estimator is the optimal linear estimator.

Although they are always uncorrelated, the residual and the feature are not independent if there exists *nonlinear* dependence between the feature and the response. Figure 8.10 shows the residual of the OLS estimator of rebounds given height in Example 8.31. The residual is uncorrelated with height (the sample correlation coefficient equals zero), but the two quantities are definitely not independent. The bottom row of the figure shows histogram-based estimates of the conditional pdf of the residual given two different values of height. The two conditional distributions are very different. The nonlinear structure that induces this dependence is apparent in the comparison between the linear and nonlinear estimators shown in Figure 8.4.

8.7 A Geometric Analysis Of Correlation

In this section we study correlation from a geometric perspective. We interpret zero-mean random variables as vectors in a vector space where the covariance is a valid inner product, as explained in Section 8.7.1. Section 8.7.2 establishes that, in this vector space, the standard deviation can be interpreted as the *length* of a random variable, and the correlation coefficient as the cosine of the angle between random variables. Section 8.7.3 builds upon these insights to provide geometric intuition about simple linear regression, establishing an equivalence between mean-squared-error minimization and orthogonal projections.

8.7.1 The Inner-Product Space of Zero-Mean Random Variables

Random variables belonging to a common probability space can be interpreted as vectors. We often think of vectors as a list of numbers with fixed length, but the mathematical definition of vector is much more general.

Definition 8.32 (Vector space)**.** *A vector space is a set $\mathcal{V}$ of objects that admit two operations, a vector sum that is commutative and associative, and a multiplication between real scalars and vectors that is associative. The operations are*

distributive with respect to each other. In order for $\mathcal{V}$ to be a valid vector space, the following conditions must hold:

- *For any $v \in \mathcal{V}$ and any scalar $\beta \in \mathbb{R}$, the scalar multiple βv belongs to $\mathcal{V}$.*
- *For any $v_1, v_2 \in \mathcal{V}$, the vector sum $v_1 + v_2$ belongs to $\mathcal{V}$.*
- *There exists a zero vector 0, such that $v + 0 = v$ for any $v \in \mathcal{V}$.*
- *For any $v \in \mathcal{V}$, there exists an additive inverse $-v$ such that $v + (-v) = 0$.*

Theorem 8.33 (Random variables form a vector space)**.** *The random variables belonging to a probability space form a vector space.*

Proof The two first conditions hold because multiplying a random variable by a real scalar or summing two random variables yields another random variable in the same probability space. This is not completely obvious in the case of continuous random variables, but can be proved formally by showing that the resulting random variable is measurable.

In order to define the zero vector and the additive inverse in Definition 8.32, we first need to establish what it means for two random variables to be equal. Two random variables in the same probability space are considered equal if they are equal with probability one, i.e. $\tilde{a} = \tilde{b}$ means

$$\mathrm{P}\left(\tilde{a} = \tilde{b}\right) = 1. \tag{8.169}$$

We define the zero vector as a random variable $\tilde{0}$ that is equal to zero with probability one. Then, for any random variable $\tilde{a}$, $\tilde{a}+\tilde{0} = \tilde{a}$ because $\tilde{a}+0 = \tilde{a}$ with probability one. We define the additive inverse of an arbitrary random variable $\tilde{a}$ as $-\tilde{a}$. With probability one, $\tilde{a} - \tilde{a} = 0$, so $\tilde{a} + (-\tilde{a}) = \tilde{0}$. We conclude that the random variables form a vector space. ■

If we restrict our attention to zero-mean random variables, the covariance can be interpreted as an inner product between random variables. Let us recall the properties of inner products between vectors.

Definition 8.34 (Inner product)**.** *An inner product on a vector space $\mathcal{V}$ is an operation $\langle \cdot, \cdot \rangle$ that maps each pair of vectors to a real number and is:*

- *Symmetric: for any $v_1, v_2 \in \mathcal{V}$,*

$$\langle v_1, v_2 \rangle = \langle v_2, v_1 \rangle . \tag{8.170}$$

- *Linear: for any $\beta \in \mathbb{R}$ and any $v_1, v_2, v_3 \in \mathcal{V}$,*

$$\langle \beta\, v_1, v_2 \rangle = \beta \langle v_1, v_2 \rangle , \tag{8.171}$$

$$\langle v_1 + v_2, v_3 \rangle = \langle v_1, v_3 \rangle + \langle v_2, v_3 \rangle . \tag{8.172}$$

- *Positive semidefinite: $\langle v, v \rangle$ is nonnegative for all $v \in \mathcal{V}$ and if $\langle v, v \rangle = 0$, then $v = 0$.*

The following lemma establishes that the covariance is an inner product for zero-mean random variables. We require the mean of the random variables to be zero so that the operation is positive semidefinite. The covariance of zero-mean

random variables is equal to the mean of their product, so its interpretation as an inner product is very natural.

Theorem 8.35 (Covariance as an inner product)**.** *The covariance is a valid inner product between zero-mean random variables.*

Proof The covariance between two random variables $\tilde{a}$ and $\tilde{b}$ is symmetric

$$\begin{aligned}
\text{Cov}[\tilde{a}, \tilde{b}] &:= \text{E}\left[(\tilde{a} - \text{E}[\tilde{a}])(\tilde{b} - \text{E}[\tilde{b}])\right] && (8.173)\\
&= \text{E}\left[(\tilde{b} - \text{E}[\tilde{b}])(\tilde{a} - \text{E}[\tilde{a}])\right] && (8.174)\\
&= \text{Cov}[\tilde{b}, \tilde{a}]. && (8.175)
\end{aligned}$$

It is also linear by linearity of expectation. For any $\beta \in \mathbb{R}$ and any two other random variables $\tilde{a}_1$ and $\tilde{a}_2$,

$$\begin{aligned}
\text{Cov}[\beta\tilde{a}, \tilde{b}] &:= \text{E}\left[(\beta\tilde{a} - \text{E}[\beta\tilde{a}])(\tilde{b} - \text{E}[\tilde{b}])\right] && (8.176)\\
&= \beta\text{E}\left[(\tilde{a} - \text{E}[\tilde{a}])(\tilde{b} - \text{E}[\tilde{b}])\right] && (8.177)\\
&= \beta\text{Cov}[\tilde{a}, \tilde{b}], && (8.178)\\
\text{Cov}[\tilde{a}_1 + \tilde{a}_2, \tilde{b}] &:= \text{E}[(\tilde{a}_1 + \tilde{a}_2)\tilde{b}] - \text{E}[\tilde{a}_1 + \tilde{a}_2]\text{E}[\tilde{b}] && (8.179)\\
&= \text{E}[\tilde{a}_1\tilde{b}] - \text{E}[\tilde{a}_1]\text{E}[\tilde{b}] + \text{E}[\tilde{a}_2\tilde{b}] - \text{E}[\tilde{a}_2]\text{E}[\tilde{b}] && (8.180)\\
&= \text{Cov}[\tilde{a}_1, \tilde{b}] + \text{Cov}[\tilde{a}_2, \tilde{b}]. && (8.181)
\end{aligned}$$

To prove that the covariance is positive semidefinite for zero-mean random variables, we need to show that $\text{E}[\tilde{a}^2] = 0$ implies $\text{P}(\tilde{a} = 0) = 1$. This follows from Chebyshev's inequality, as we establish in Corollary 9.20. ■

8.7.2 The Geometry Of Zero-Mean Random Variables

We can use the inner product to define a norm for zero-mean random variables, which represents the *length* of the random variable interpreted as a vector. The norm associated to an inner product is

$$||v|| := \sqrt{\langle v, v \rangle}, \tag{8.182}$$

where v is an arbitrary vector. The norm of a zero-mean random variable $\tilde{a}$ associated to the covariance is therefore equal to the square root of its mean square or variance (they are the same for zero-mean random variables),

$$\begin{aligned}
||\tilde{a}|| &:= \sqrt{\text{Cov}[\tilde{a}, \tilde{a}]} && (8.183)\\
&= \sqrt{\text{Var}[\tilde{a}]}. && (8.184)
\end{aligned}$$

In our vector analogy, the length of a zero-mean random variable is its standard deviation, as depicted in the left image of Figure 8.11. When we divide a zero-mean random variable by its standard deviation, we normalize its vector representation so that it has unit length.

A natural measure of the similarity between vectors is the cosine of the angle between them, which equals their inner product normalized by their norms. Let us denote the angle by θ, as in the left image of Figure 8.11. Then,

$$\cos\theta = \frac{\langle \tilde{a}, \tilde{b} \rangle}{||\tilde{a}||\,||\tilde{b}||} \tag{8.185}$$

$$= \frac{\mathrm{Cov}[\tilde{a}, \tilde{b}]}{\sqrt{\mathrm{Var}[\tilde{a}]\mathrm{Var}[\tilde{b}]}} \tag{8.186}$$

$$= \rho_{\tilde{a},\tilde{b}}. \tag{8.187}$$

The cosine is exactly equal to the correlation coefficient! This analogy provides a geometric perspective on correlation. The vectors corresponding to two positively correlated random variables ($\rho_{\tilde{a},\tilde{b}} > 0$) point in the same direction. If the random variables are negatively correlated ($\rho_{\tilde{a},\tilde{b}} < 0$), then the vectors point in opposite directions. If they are uncorrelated ($\rho_{\tilde{a},\tilde{b}} = 0$), then the vectors are orthogonal.

The analogy between the cosine and the correlation coefficient implies some of the properties of the correlation coefficient derived in Section 8.5. Since it is the cosine of an angle, the correlation coefficient is between -1 and 1, as established in Theorem 8.20. When the correlation coefficient is exactly equal to 1 or -1, the two vectors are collinear, so each of the random variables can be expressed exactly as a linear scaling of the other, as established in Theorem 8.21.

8.7.3 Simple Linear Regression Via Orthogonal Projection

In this section, we show how the interpretation of zero-mean random variables as vectors can be leveraged to gain geometric intuition about the simple linear regression problem, and to easily rederive the linear MMSE estimator.

Our goal is to estimate a zero-mean random variable $\tilde{b}$ by scaling another zero-mean random variable $\tilde{a}$. Minimizing the mean squared error of the estimate is equivalent to minimizing the squared norm induced by the covariance inner product. Indeed, since $\tilde{a}$ and $\tilde{b}$ have mean zero, then $\tilde{b} - \beta\tilde{a}$ also has zero mean for any β, so

$$\mathrm{E}\left[(\tilde{b} - \beta\tilde{a})^2\right] = \mathrm{Var}\left[\tilde{b} - \beta\tilde{a}\right] \tag{8.188}$$

$$= ||\tilde{b} - \beta\tilde{a}||^2. \tag{8.189}$$

Interpreting $\tilde{a}$ and $\tilde{b}$ as vectors, we want the closest point to $\tilde{b}$ that is collinear with $\tilde{a}$. In linear algebra, this is called the *projection* of $\tilde{b}$ onto $\tilde{a}$ (or rather the subspace spanned by $\tilde{a}$). To find the projection, we realize that the distance between $\tilde{b}$ and $\beta\tilde{a}$ is minimized when the residual $\tilde{b} - \beta\tilde{a}$ is orthogonal to $\tilde{a}$ (see the right image in Figure 8.11),

$$\langle \tilde{a}, \tilde{b} - \beta\tilde{a} \rangle = 0, \tag{8.190}$$

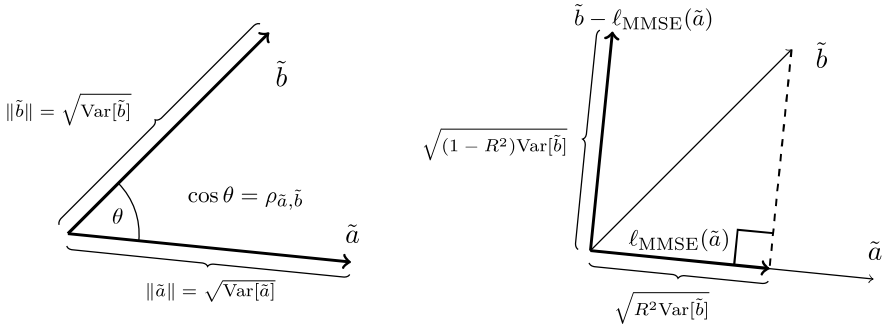

Figure 8.11 Geometric interpretation of correlation for zero-mean random variables. The left image illustrates the geometric interpretation of correlation for zero-mean random variables described in Section 8.7. The lengths of the vectors representing the random variables are equal to their standard deviations. The cosine of the angle between the vectors is equal to the correlation coefficient. The right image depicts the connection between linear regression and orthogonal projections described in Section 8.7.3. The linear MMSE estimator of the response $\tilde{b}$ given the feature $\tilde{a}$ is the projection of $\tilde{b}$ onto the line spanned by $\tilde{a}$. The estimate and the residual are orthogonal and hence uncorrelated. The squared length of $\tilde{b}$ (its variance) can be decomposed into the sum of the squared lengths (variances) of the estimate and of the residual. The ratio between the length of the estimate and of the response is equal to the coefficient of determination R^2.

which implies $\beta \left|\left|\tilde{a}\right|\right|^2 = \langle \tilde{a}, \tilde{b} \rangle$, so

$$\beta = \frac{\langle \tilde{a}, \tilde{b} \rangle}{\left|\left|\tilde{a}\right|\right|^2} \tag{8.191}$$

$$= \frac{\text{Cov}\left[\tilde{a}, \tilde{b}\right]}{\text{Var}\left[\tilde{a}\right]} \tag{8.192}$$

$$= \rho_{\tilde{a},\tilde{b}} \sqrt{\frac{\text{Var}[\tilde{b}]}{\text{Var}[\tilde{a}]}}. \tag{8.193}$$

Therefore, $\beta\tilde{a} = \ell_{\text{MMSE}}(\tilde{a})$, the linear MMSE estimator derived in Theorem 8.14 (for zero-mean random variables).

The orthogonality between $\tilde{a}$ and the residual implies that they are uncorrelated, which is consistent with Lemma 8.24. It also implies that $\ell_{\text{MMSE}}(\tilde{a})$ and the residual are orthogonal (since $\ell_{\text{MMSE}}(\tilde{a})$ lies in the same direction as $\tilde{a}$), so by Pythagoras's theorem,

$$\text{Var}[\tilde{b}] = \left|\left|\tilde{b}\right|\right|^2 = \left|\left|\ell_{\text{MMSE}}(\tilde{a})\right|\right|^2 + \left|\left|\tilde{b} - \ell_{\text{MMSE}}(\tilde{a})\right|\right|^2 \tag{8.194}$$

$$= \text{Var}\left[\ell_{\text{MMSE}}(\tilde{a})\right] + \text{Var}\left[\tilde{b} - \ell_{\text{MMSE}}(\tilde{a})\right]. \tag{8.195}$$

We obtain the same decomposition of variance as in Theorem 8.25. The squared

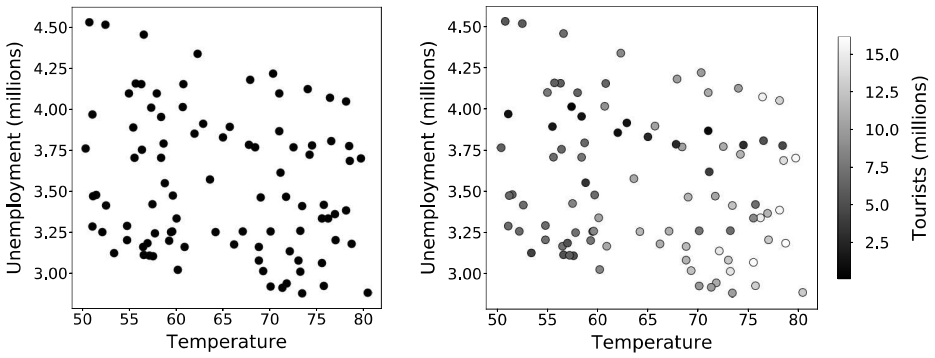

Figure 8.12 Unemployment and temperature. The scatterplot on the left shows the number of unemployed people and the temperature in Spain for each month between 2015 and 2022. The correlation coefficient is -0.21. The scatterplot on the right shows the number of tourists visiting in Spain (indicated by the color of each marker). When the temperature is high, there are more tourists, which explains the negative correlation between unemployment and temperature. We exclude April and May 2020 from our analysis because there were no tourists due to the COVID-19 pandemic.

length of the projection equals

$$||\ell_{\text{MMSE}}(\tilde{a})||^2 = ||\beta\tilde{a}||^2 \tag{8.196}$$

$$= \beta^2 ||\tilde{a}||^2 \tag{8.197}$$

$$= \frac{\rho^2_{\tilde{a},\tilde{b}}\text{Var}[\tilde{b}]}{\text{Var}[\tilde{a}]}\text{Var}[\tilde{a}] \tag{8.198}$$

$$= \rho^2_{\tilde{a},\tilde{b}}\text{Var}[\tilde{b}], \tag{8.199}$$

so the fraction of the squared length of $\tilde{b}$ covered by the linear estimate is the coefficient of determination $R^2 = \rho^2_{\tilde{a},\tilde{b}}$, as established in Theorem 8.27 and depicted in the right image of Figure 8.11.

8.8 Correlation (Usually) Does Not Imply Causation

Great care must be taken when interpreting correlation in terms of causal effects. For example, unemployment in Spain is negatively correlated with temperature, as shown in the left graph of Figure 8.12. This means that unemployment is lower, on average, when the temperatures are higher. However, it does *not* mean that higher temperatures *cause* the unemployment to decrease. If the temperature in Spain suddenly rose by several degrees, this would probably not decrease unemployment.

The goal of causal inference is to determine whether a variable, usually known as the *treatment*, has a causal effect on another variable, called the *outcome*. The

potential outcome $\widetilde{\text{po}}_t$ represents the distribution of the outcome of interest in a hypothetical scenario where the treatment $\tilde{t}$ equals t. In Sections 4.6.1 and 7.9, we study causal effects of binary treatments. Here, we consider discrete or continuous treatments with multiple possible values. In our example, the treatment $\tilde{t}$ is the temperature, which is typically modeled as continuous. The corresponding potential outcome $\widetilde{\text{po}}_t$ represents unemployment in a hypothetical situation where the temperature in Spain is always equal to t.

The problem that we encounter when trying to characterize causal effects from data is that our measurements of the potential outcome $\widetilde{\text{po}}_t$ are extremely incomplete. As illustrated in Figure 8.13, for each data point, the observed outcome is equal to the potential outcome for one specific value of the treatment:

$$\tilde{y} := \widetilde{\text{po}}_t \qquad \text{if} \qquad \tilde{t} = t. \tag{8.200}$$

The potential outcomes associated to all other values of the treatment are unobserved *counterfactuals*. In our example, if the temperature is 60 degrees, then we only get to see $\widetilde{\text{po}}_{60}$. We don't know what would have happened if the temperature had been equal to another value, such as 50 or 70, instead.

A common assumption when analyzing nonbinary treatments is that the average causal dependence between the treatment and the outcome is *linear*. More precisely, for a certain constant β_{true},

$$\text{E}\left[\widetilde{\text{po}}_t\right] = \beta_{\text{true}} t. \tag{8.201}$$

In our example, if $\beta_{\text{true}} = -1$, increasing temperatures by one degree decreases unemployment on average by one unit (one million people).

In Theorems 4.23 and 7.62, we establish that randomization enables us to estimate causal effects in the case of binary treatments. The key insight is that if the treatment and the potential outcomes are independent, then the observed outcome in the treatment and control groups is an accurate proxy for the corresponding potential outcome. The independence assumption is key, because it ensures that the treatment is the only systematic difference between the treatment and control groups. The following theorem extends this result beyond binary treatments, showing that we can estimate linear causal effects from data by computing the covariance between the observed output and the treatment, as long as the treatment and the potential outcomes are independent.

Theorem 8.36 (Estimation of causal linear effects)**.** *Let $\widetilde{\text{po}}_t$ denote the potential outcomes associated to a discrete or continuous treatment $\tilde{t}$, which we assume is standardized, so that $\text{E}[\tilde{t}] = 0$ and $\text{E}[\tilde{t}^2] = 1$. The observed outcome $\tilde{y}$ is equal to $\widetilde{\text{po}}_t$ if $\tilde{t} = t$. We assume that the treatment has a linear effect on the mean of $\widetilde{\text{po}}_t$: $\text{E}\left[\widetilde{\text{po}}_t\right] = \beta_{\text{true}} t$, for some real constant β_{true}. If $\widetilde{\text{po}}_t$ and $\tilde{t}$ are independent for all t, then*

$$\text{Cov}\left[\tilde{y}, \tilde{t}\right] = \beta_{\text{true}}. \tag{8.202}$$

Proof Without loss of generality, we assume that all variables are continuous. In order to derive the covariance of $\tilde{y}$ and $\tilde{t}$, we first compute the conditional mean

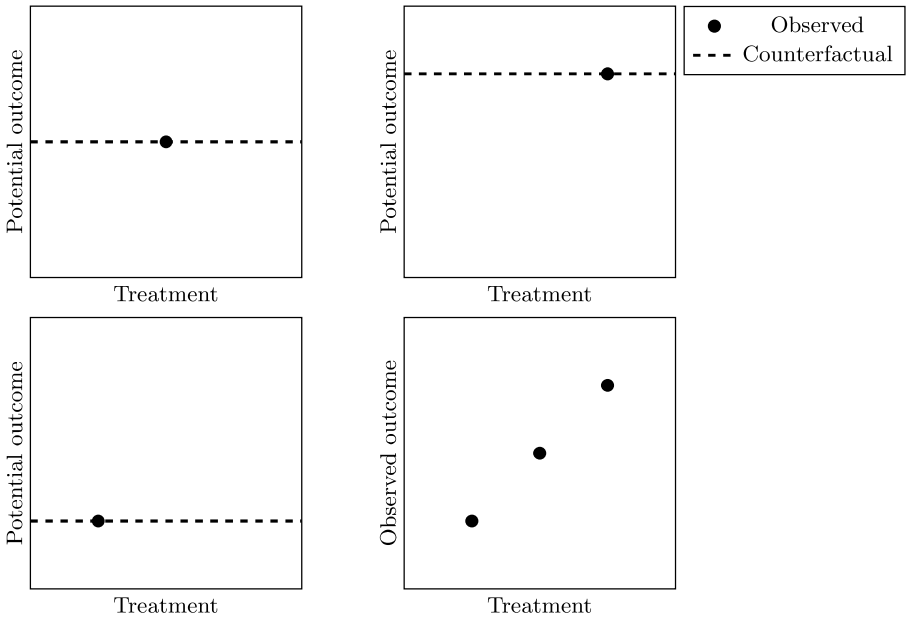

Figure 8.13 Observed and counterfactual outcomes. The top two plots and the bottom left plot show three instances of a potential outcome associated to the treatment. The observed outcome corresponds to only one value of the treatment (indicated by a marker). The remaining values (dashed lines) are unobserved. In every case, the potential outcome is constant with respect to the treatment, so the treatment has no causal effect. However, the treatment is proportional to the potential outcome, which induces a spurious correlation between the observed outcome and the treatment, as shown in the lower right plot.

function of their product given $\tilde{t} = t$, and then apply iterated expectation. By the independence assumption, the conditional mean function equals

$$\mu_{\tilde{y}\tilde{t}\,|\,\tilde{t}}(t) = \int_{y=-\infty}^{\infty} yt f_{\tilde{y}\,|\,\tilde{t}}(y\,|\,t)\ \mathrm{d}y \tag{8.203}$$

$$= \int_{y=-\infty}^{\infty} yt f_{\widetilde{\mathrm{po}}_t\,|\,\tilde{t}}(y\,|\,t)\ \mathrm{d}y \tag{8.204}$$

$$= t\int_{y=-\infty}^{\infty} y f_{\widetilde{\mathrm{po}}_t}(y)\ \mathrm{d}y \tag{8.205}$$

$$= t\mathrm{E}\left[\widetilde{\mathrm{po}}_t\right] \tag{8.206}$$

$$= \beta_{\mathrm{true}} t^2. \tag{8.207}$$

By iterated expectation and the assumption that $\mathrm{E}\left[\tilde{t}\right] = 0$,

$$\begin{aligned}
\mathrm{Cov}\left[\tilde{y}, \tilde{t}\right] = \mathrm{E}\left[\tilde{y}\tilde{t}\right] &= \mathrm{E}\left[\mu_{\tilde{y}\tilde{t}\,|\,\tilde{t}}(\tilde{t})\right] && (8.208)\\
&= \mathrm{E}\left[\beta_{\text{true}}\tilde{t}^{\,2}\right] && (8.209)\\
&= \beta_{\text{true}}\mathrm{E}\left[\tilde{t}^{\,2}\right] && (8.210)\\
&= \beta_{\text{true}}. && (8.211)
\end{aligned}$$

■

Randomizing the treatment guarantees that the potential outcomes and the treatment are independent, but in many cases randomization is not possible. This includes our motivating example: we cannot set the temperature in Spain to a random value and then measure the corresponding unemployment! Consequently, we cannot guarantee that the potential outcomes associated with unemployment are independent from the temperatures, so the observed correlation does not necessarily imply a causal linear effect.

In fact, the main reason why temperature and unemployment in Spain are correlated is that more tourists visit Spain when the temperatures are higher, as shown in the right graph in Figure 8.12. The number of tourists does have a causal effect on unemployment. If one million more tourists were suddenly to arrive in Spain, then unemployment would surely decrease as a consequence. The number of tourists is a confounding factor or confounder, which induces a spurious correlation between temperature and unemployment. A similar effect occurs in Figure 8.13: the potential outcome is constant with respect to the treatment for each observation, but there is a spurious correlation between the observed outcome and the treatment. This is caused by the choice of treatment, which is proportional to the potential outcome. Randomization avoids such situations by rendering the treatment and the potential outcome independent.

The following example further illustrates the influence of confounders on the correlation between the observed outcome and the treatment.

Example 8.37 (Guinea-pig rescue: Spurious correlation)**.** Imara runs a guinea-pig rescue center. Rescued guinea pigs are typically underweight, so it is important to fatten them up. Imara is interested in determining whether a certain nutritional supplement is effective. We model the amount of supplement provided to a guinea pig over a week as a treatment random variable $\tilde{t}$. The observed outcome of interest, modeled by the random variable $\tilde{y}$, is the resulting change in the weight of the guinea pig.

Being familiar with Theorem 8.36, Imara assumes that the causal effect is linear, and estimates it by computing the covariance between the treatment $\tilde{t}$ and the observed outcome $\tilde{y}$, $\mathrm{E}\left[\tilde{y}\tilde{t}\right]$ (after standardizing the treatment). She finds that the covariance is positive, which suggests that the weight gain is proportional to the supplement eaten by the pigs.

To double check, Imara asks her friend Christina to repeat the experiment the following week while she's on a trip to Ohio. Puzzlingly, the covariance turns out to be negative, which suggests that the supplement is making the pigs lose

weight! Imara and Christina consult their friend Dimitris, who advises them to repeat the experiment once again, but randomizing the treatment to ensure that the guarantees in Theorem 8.36 are valid. Making each guinea pig eat a random amount of supplement is hard, but Imara and Christina manage to do it. This time the covariance is zero, indicating that the supplement is useless.

The reason for the contradictory results is that Imara and Christina fed the guinea pigs differently: Imara mixed it in with the food, but Christina gave them the supplement after their meal. In the first case, the pigs that eat more food also eat more supplement. Consequently, the pigs that eat more supplement gain more weight. In the second case, the pigs that eat more food are full by the time the supplement is offered, so they eat less supplement. As a result, the pigs that eat less supplement gain more weight. The food intake is a confounder that produces a spurious correlation between the treatment and the observed outcome. Randomizing the treatment avoids this by ensuring that there is no systematic difference in food intake between the pigs that eat more supplement and the ones that eat less.

The following theorem shows that, in general, we cannot estimate the linear causal effect of the treatment using the covariance between the observed outcome and the treatment, as proposed in Theorem 8.36, because the covariance can be distorted by confounders.

Theorem 8.38 (Unobserved confounders produce spurious correlations)**.** *Let* $\widetilde{\text{po}}_{t,c}$ *denote the potential outcome associated to a treatment* $\tilde{t}$ *and a confounding variable* $\tilde{c}$*. The observed outcome* $\tilde{y}$ *is equal to* $\widetilde{\text{po}}_{t,c}$ *if* $\tilde{t} = t$ *and* $\tilde{c} = c$*. We assume that the treatment and confounder have a linear effect on the mean of* $\widetilde{\text{po}}_{t,c}$*:*

$$\text{E}\left[\widetilde{\text{po}}_{t,c}\right] = \beta_{\text{true}} t + \gamma c, \tag{8.212}$$

for some real constants β_{true} *and* γ*, where* $\tilde{t}$ *and* $\tilde{c}$ *are standardized. In addition, we assume that the dependence between the potential outcome and the treatment is due only to the confounder. For all possible values of* t *and* c*, the distribution of* $\widetilde{\text{po}}_{t,c}$ *is the same as the distribution of* $\widetilde{\text{po}}_{t,c}$ *conditioned on the intersection of the events* $\tilde{t} = t$ *and* $\tilde{c} = c$*.*

In that case, the covariance between the observed outcome and the treatment is dictated by the confounder. It equals

$$\text{Cov}\left[\tilde{y}, \tilde{t}\right] = \beta_{\text{true}} + \gamma \rho_{\tilde{t},\tilde{c}}, \tag{8.213}$$

where $\rho_{\tilde{t},\tilde{c}}$ *is the correlation coefficient of* $\tilde{t}$ *and the confounder* $\tilde{c}$*.*

Proof Without loss of generality, we assume that all variables are continuous. We apply a similar approach to the proof of Theorem 8.36. Our first step is to derive the conditional mean function of $\tilde{y}\tilde{t}$ given $\tilde{t} = t$ and $\tilde{c} = c$. Under the assumption that the distribution of $\widetilde{\text{po}}_{t,c}$ is the same as the distribution of $\widetilde{\text{po}}_{t,c}$ conditioned on the intersection of the events $\tilde{t} = c$ and $\tilde{c} = c$, the pdf $f_{\widetilde{\text{po}}_{t,c}}$ of $\widetilde{\text{po}}_{t,c}$

is equal to its conditional pdf $f_{\widetilde{\text{po}}_{t,c}\,|\,\tilde{t},\tilde{c}}(\cdot\,|\,c,t)$ given $\tilde{t}=t$ and $\tilde{c}=c$. Therefore,

$$\mu_{\tilde{y}\tilde{t}\,|\,\tilde{t},\tilde{c}}(t,c) = \int_{y=-\infty}^{\infty} yt\, f_{\tilde{y}\,|\,\tilde{t},\tilde{c}}(y\,|\,t,c)\;\mathrm{d}y \tag{8.214}$$

$$= \int_{y=-\infty}^{\infty} yt\, f_{\widetilde{\text{po}}_{t,c}\,|\,\tilde{t},\tilde{c}}(y\,|\,t,c)\;\mathrm{d}y \tag{8.215}$$

$$= t\int_{y=-\infty}^{\infty} y\, f_{\widetilde{\text{po}}_{t,c}}(y)\;\mathrm{d}y \tag{8.216}$$

$$= t\mathrm{E}\left[\widetilde{\text{po}}_{t,c}\right] \tag{8.217}$$

$$= \beta_{\text{true}}t^2 + \gamma ct. \tag{8.218}$$

The covariance can now be computed via iterated expectation,

$$\mathrm{Cov}\left[\tilde{y},\tilde{t}\right] = \mathrm{E}\left[\tilde{y}\tilde{t}\right] = \mathrm{E}\left[\mu_{\tilde{y}\tilde{t}\,|\,\tilde{t},\tilde{c}}(\tilde{t},\tilde{c})\right] \tag{8.219}$$

$$= \mathrm{E}\left[\beta_{\text{true}}\tilde{t}^{\,2} + \gamma\tilde{t}\tilde{c}\right] \tag{8.220}$$

$$= \beta_{\text{true}}\mathrm{E}\left[\tilde{t}^{\,2}\right] + \gamma\mathrm{E}\left[\tilde{t}\tilde{c}\right] \tag{8.221}$$

$$= \beta_{\text{true}} + \gamma\rho_{\tilde{t},\tilde{c}}. \tag{8.222}$$

■

Theorem 8.38 does not contradict Theorem 8.36. If we randomize the treatment, then this will render it independent from $\tilde{c}$, so $\rho_{\tilde{t},\tilde{c}}=0$, and the covariance is indeed equal to β_{true}.

Example 8.39 (Guinea-pig rescue: Quantitative analysis)**.** In this example we provide a more quantitative analysis of Example 8.37. We model the treatment (supplement eaten by the guinea pigs) and the confounder (food intake) as the random variables $\tilde{t}$ and $\tilde{c}$, both standardized to have zero mean and unit variance. The potential outcome $\widetilde{\text{po}}_{t,c}$ representing the weight change of the guinea pigs depends on both the treatment and the confounder. The observed outcome $\tilde{y}$ equals $\widetilde{\text{po}}_{t,c}$ when $\tilde{t}=t$ and $\tilde{c}=c$.

Since the supplement is useless and weight gain only depends on the food intake, we set the mean of the potential outcome $\widetilde{\text{po}}_{t,c}$ equal to c,

$$\mathrm{E}\left[\widetilde{\text{po}}_{t,c}\right] = c. \tag{8.223}$$

We model $\widetilde{\text{po}}_{t,c}$ as being independent from $\tilde{t}$ and $\tilde{c}$, which means that there is no systematic difference between the guinea pigs receiving a certain amount of supplement and food (i.e. there are no additional confounders). By Theorem 8.38, the covariance between $\tilde{y}$ and $\tilde{t}$ equals $\rho_{\tilde{t},\tilde{c}}$. This is illustrated in Figure 8.14, assuming that the joint distribution of the treatment and the confounder is Gaussian, and that the distribution of $\widetilde{\text{po}}_{t,c}$ is Gaussian with mean c and unit variance.

When Imara feeds the pigs and the supplement is mixed with the food (left column), the correlation between the confounder and the treatment is positive (first row), which produces a spurious positive correlation between the treatment and the observed outcome (second row) completely driven by the confounder (third

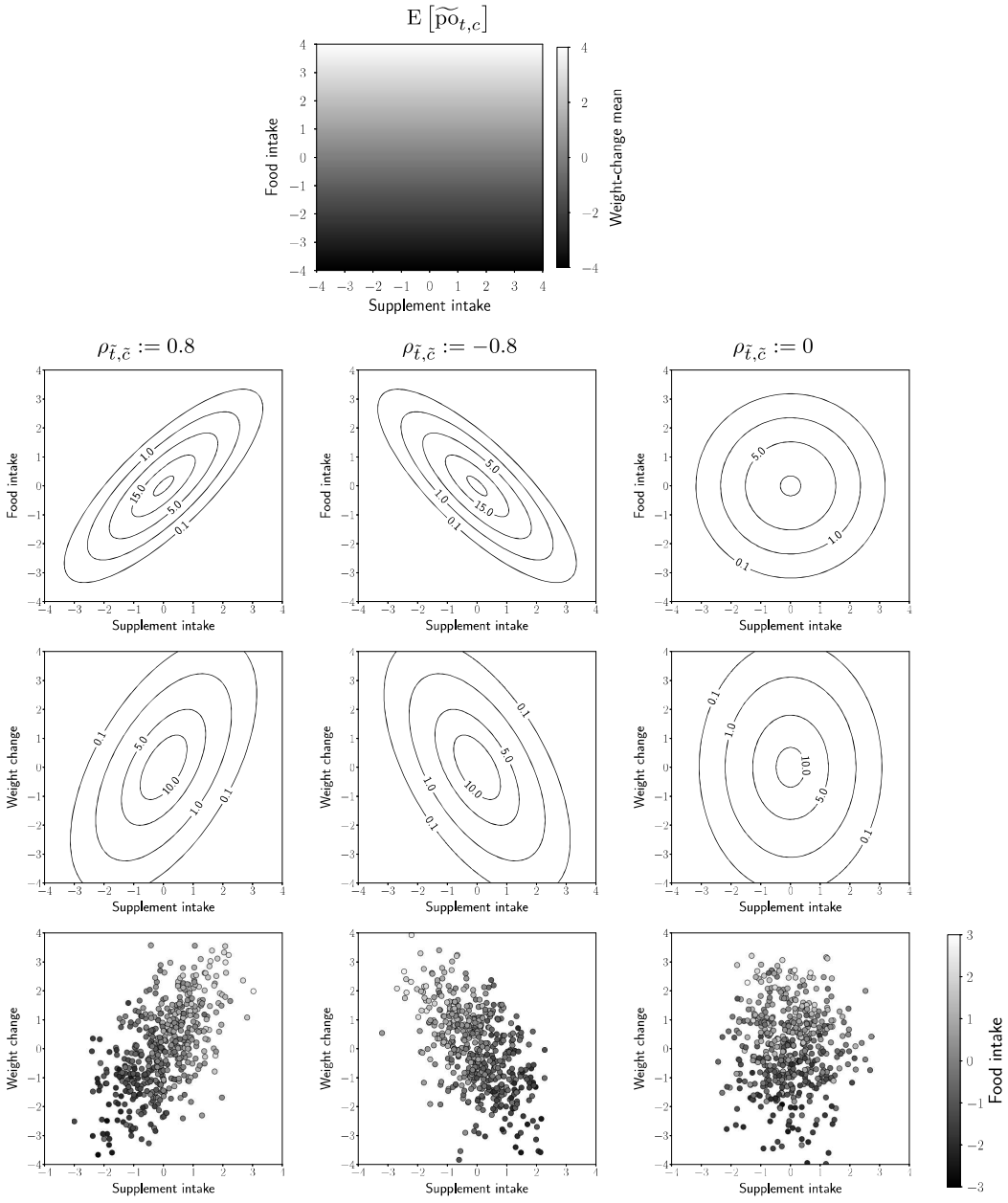

Figure 8.14 Unobserved confounders distort the observed correlation. The heatmap at the top shows the mean of the potential outcome $\widetilde{\mathrm{po}}_{t,c}$ associated to the weight change of the guinea pigs in Example 8.39. $\mathrm{E}[\widetilde{\mathrm{po}}_{t,c}]$ only depends on the food intake c and not at all on the supplement intake t. Below, the first row shows the contour lines of the joint pdf between the random variables representing the treatment ($\tilde{t}$) and the food intake ($\tilde{c}$) for different values of the correlation coefficient between them. The second row shows the contour lines of the corresponding joint pdf between $\tilde{t}$ and the observed outcome $\tilde{y}$ (see Exercise 5.13 for a derivation of the joint pdf). As revealed by the scatterplot in the third row, the observed correlation between $\tilde{t}$ and $\tilde{y}$ is completely dictated by the confounding factor $\tilde{c}$, as established in Theorem 8.38.

row). When Christina feeds them the supplement after the food (middle column), the correlation between the confounder and the treatment is negative (first row), which produces a spurious negative correlation between the treatment and the observed outcome (second row) again completely driven by the confounder (third row). When they randomize the treatment, this renders it independent from the confounder (first row), so there is no spurious correlation (second row) because the confounder is neutralized (third row).

In conclusion, correlation computed from data cannot be interpreted causally, unless the treatment is randomized. Otherwise, unobserved confounders can completely distort it. In Section 12.1.4, we discuss how to adjust for known confounders using linear regression.