Uncorrelation and Independence

Probability and Statistics for Data Science

Carlos Fernandez-Granda

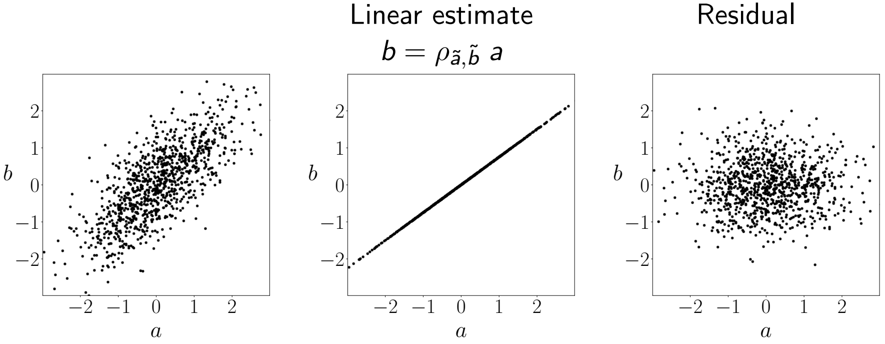
These slides are based on the book Probability and Statistics for Data Science by Carlos Fernandez-Granda, available for purchase here. A free preprint, videos, code, slides and solutions to exercises are available at https://www.ps4ds.net

Goal

Explain the difference between uncorrelation and independence

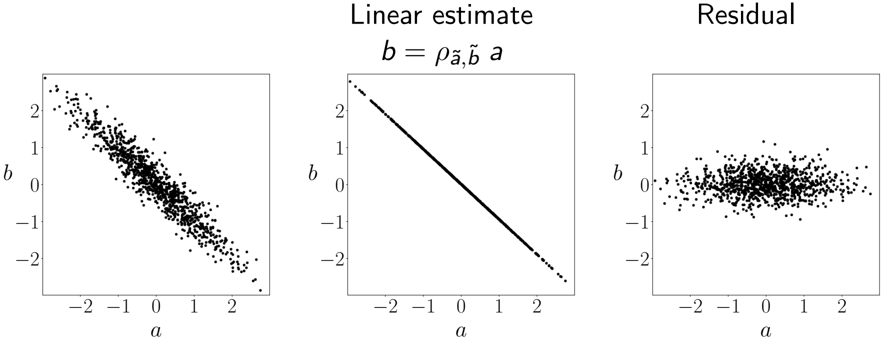
$\rho_{\tilde{a},\tilde{b}} = 0.75$

If $\rho_{\tilde{a},\tilde{b}}$ and $\mathrm{Cov}[\tilde{a},\tilde{b}]$ are positive, $\tilde{a}$ and $\tilde{b}$ are positively correlated



$\rho_{\tilde{a},\tilde{b}} = -0.95$

If $\rho_{\tilde{a},\tilde{b}}$ and $\mathrm{Cov}[\tilde{a}, \tilde{b}]$ are negative, $\tilde{a}$ and $\tilde{b}$ are negatively correlated

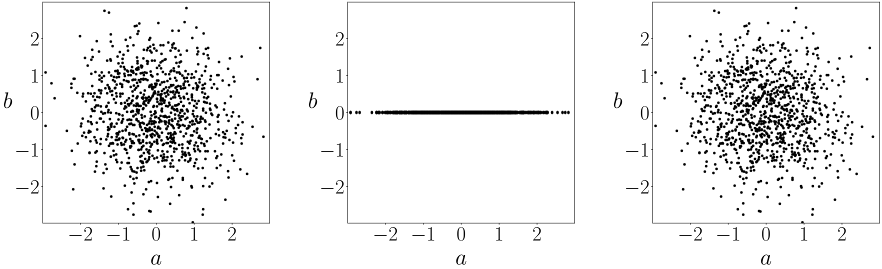


$\rho_{\tilde{a},\tilde{b}} = 0$

If $\rho_{\tilde{a},\tilde{b}}$ and $\mathrm{Cov}[\tilde{a}, \tilde{b}]$ are zero, $\tilde{a}$ and $\tilde{b}$ are uncorrelated

Linear estimate
$b = \rho_{\tilde{a},\tilde{b}}\ a$

Residual



Independence implies uncorrelation

If $\tilde{a}$ and $\tilde{b}$ are independent, then

$$\begin{aligned}
\operatorname{Cov}[\tilde{a}, \tilde{b}] &= \mathrm{E}[\tilde{a}\tilde{b}] - \mathrm{E}[\tilde{a}]\,\mathrm{E}[\tilde{b}] \\
&= \mathrm{E}[\tilde{a}]\,\mathrm{E}[\tilde{b}] - \mathrm{E}[\tilde{a}]\,\mathrm{E}[\tilde{b}] \\
&= 0
\end{aligned}$$

Gaussian random variables

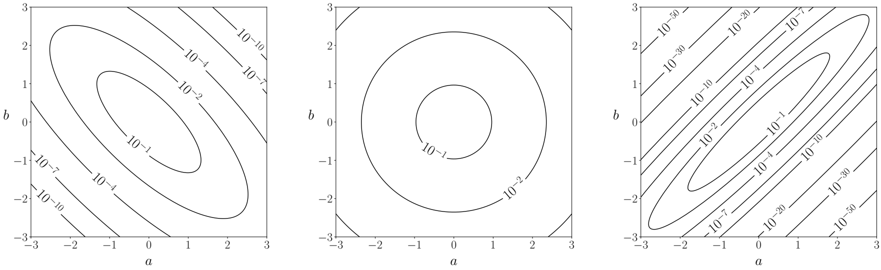
Covariance matrix for uncorrelated Gaussian random variables with zero mean and unit variance

$$\Sigma := \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \qquad \Sigma^{-1} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

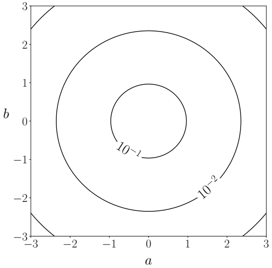
$$\begin{aligned} f_{\tilde{a},\tilde{b}}(a,b) &= \frac{1}{2\pi\sqrt{|\Sigma|}} \exp\left(-\frac{1}{2} \begin{bmatrix} a \\ b \end{bmatrix}^T \Sigma^{-1} \begin{bmatrix} a \\ b \end{bmatrix} \right) \\ &= \frac{1}{2\pi} \exp\left(-\frac{a^2+b^2}{2} \right) \\ &= \frac{1}{2\pi} \exp\left(-\frac{a^2}{2} \right) \frac{1}{2\pi} \exp\left(-\frac{b^2}{2} \right) \\ &= f_{\tilde{a}}(a) f_{\tilde{b}}(b) \qquad \text{Independent} \end{aligned}$$

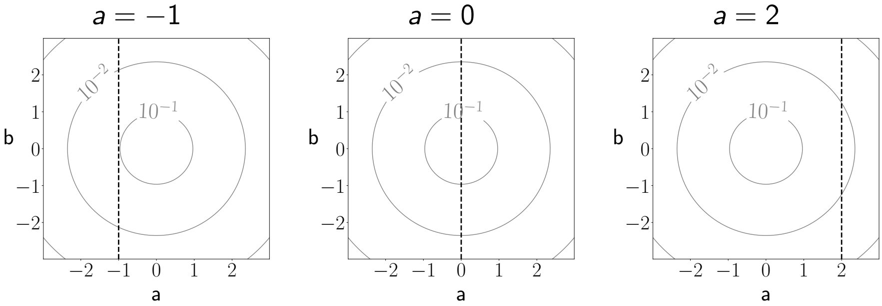
Gaussian random variables

$\rho = -0.75$ $\rho = 0$ $\rho = 0.95$

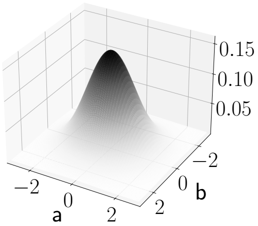


$\rho = 0$

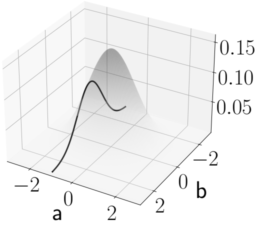




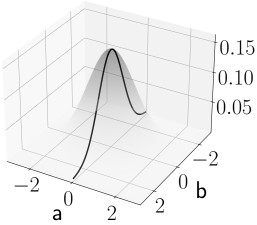
Conditional distribution given $\tilde{a} = a$



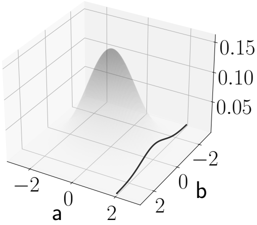
$a = -1$



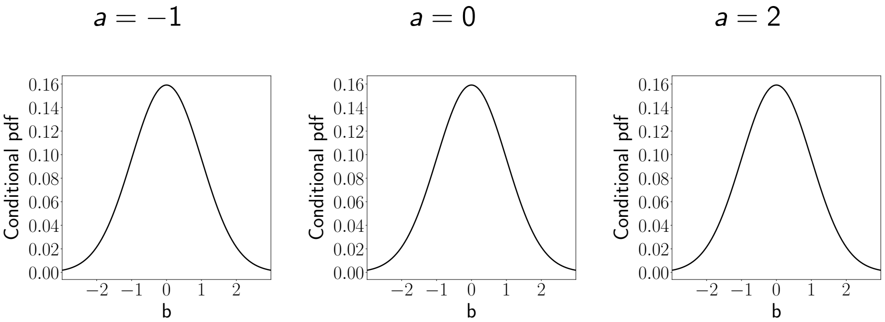
$a = 0$



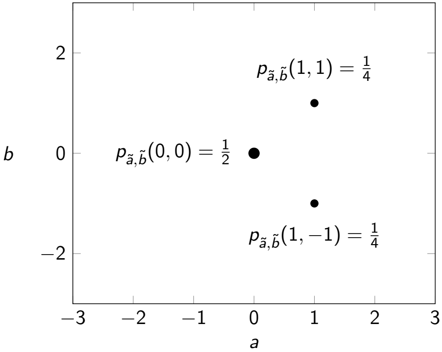
$a = 2$



Conditional distribution given $\tilde{a} = a$



Example

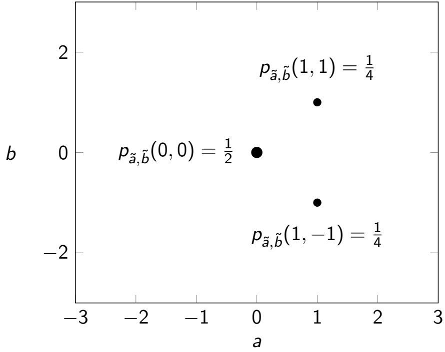


$$\mathrm{E}[\tilde{b}] = \sum_{a=0}^{1} \sum_{b=-1}^{1} b\, p_{\tilde{a},\tilde{b}}(a,b) = 0 \cdot \frac{1}{2} - 1 \cdot \frac{1}{4} + 1 \cdot \frac{1}{4} = 0$$

$$\mathrm{Cov}[\tilde{a},\tilde{b}] = \mathrm{E}[\tilde{a}\tilde{b}] - \mathrm{E}[\tilde{a}]\mathrm{E}[\tilde{b}] = \mathrm{E}[\tilde{a}\tilde{b}] \qquad \text{Uncorrelated}$$

$$= \sum_{a=0}^{1} \sum_{b=-1}^{1} ab\, p_{\tilde{a},\tilde{b}}(a,b) = 0 \cdot \frac{1}{2} - 1 \cdot \frac{1}{4} + 1 \cdot \frac{1}{4} = 0$$

Example



Conditional pmf of $\tilde{b}$ given $\tilde{a} = 0$?

$$p_{\tilde{b}\,|\,\tilde{a}}(0\,|\,0) = 1$$

Conditional pmf of $\tilde{b}$ given $\tilde{a} = 1$?

$$p_{\tilde{b}\,|\,\tilde{a}}(1\,|\,1) = \frac{1}{2} \qquad p_{\tilde{b}\,|\,\tilde{a}}(-1\,|\,1) = \frac{1}{2}$$

Not independent

Uncorrelated residual

Let $\ell_{\mathsf{MMSE}}(\tilde{a})$ be the linear MMSE estimate of $\tilde{b}$ given $\tilde{a}$

$$\mathrm{Cov}\left[\tilde{a}, \tilde{b} - \ell_{\mathsf{MMSE}}(\tilde{a})\right] = 0$$

Height of NBA players

Data:

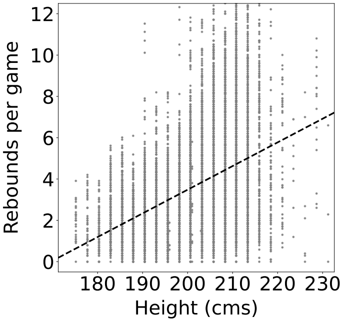
Height and offensive statistics of NBA players between 1996 and 2019

Goal:

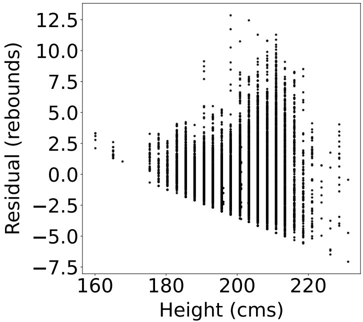
Quantify linear dependence between rebounds/assists/points and height

Rebounds and height

OLS estimator

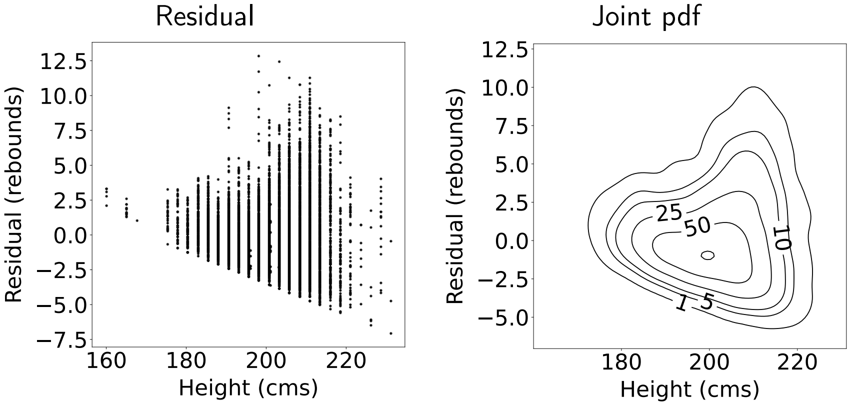


Residual

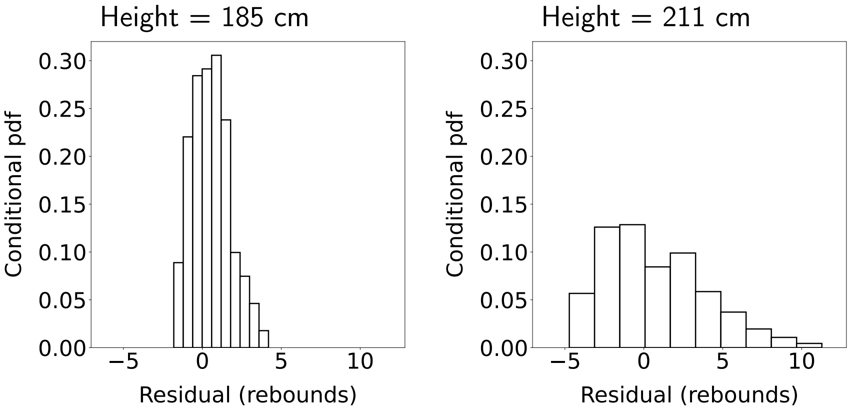


Correlation coefficient between residual and height: 0

Independent?



Conditional distribution of residual given height



What have we learned

- Independence implies uncorrelation
- Uncorrelation does not imply independence
- But it does for Gaussian random variables