Overview of Correlation

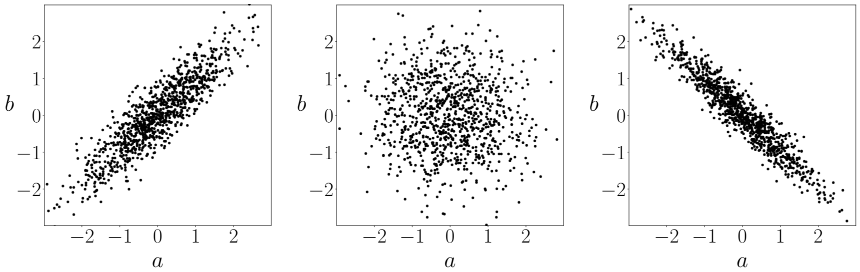
Probability and Statistics for Data Science

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

Quantify dependence between two quantities with a single number



Idea: Focus on *linear* dependence

Topics

Correlation coefficient and covariance

Geometric intuition about correlation

Simple linear regression

Causal inference

Linear dependence

How can we quantify linear dependence between random variables $\tilde{a}$ and $\tilde{b}$?

Approximate $\tilde{b}$ using linear function of $\tilde{a}$

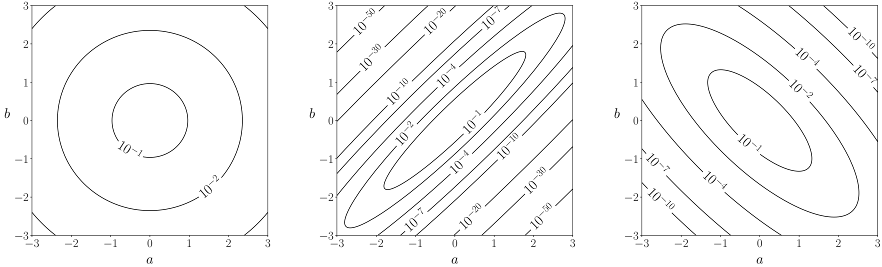
We first focus on random variables with zero mean and unit variance

Linear minimum mean-squared-error estimator of $\tilde{b}$ given $\tilde{a}$ is $\mathrm{E}[\tilde{a}\tilde{b}]$

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

Gaussian random variables

$\rho_{\tilde{a},\tilde{b}} = 0$ $\rho_{\tilde{a},\tilde{b}} = 0.95$ $\rho_{\tilde{a},\tilde{b}} = -0.75$



Correlation coefficient

What about random variables with non-zero mean or non-unit variance?

Standardized variable

To standardize a random variable $\tilde{a}$ we subtract its mean $\mu_{\tilde{a}}$ and divide by its standard deviation $\sigma_{\tilde{a}}$

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

$$\mathrm{E}\left[\mathrm{s}(\tilde{a})\right] = 0$$

$$\mathrm{Var}\left[\mathrm{s}(\tilde{a})\right] = 1$$

Linear dependence between random variables

Random variables $\tilde{a}$ and $\tilde{b}$ with means $\mu_{\tilde{a}}$ and $\mu_{\tilde{b}}$ and variances $\sigma_{\tilde{a}}^2$ and $\sigma_{\tilde{b}}^2$

Affine approximation of $\tilde{b}$ given $\tilde{a}$?

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

This is the minimum MSE linear estimator

Correlation coefficient

$$\rho_{\tilde{a},\tilde{b}} := \rho_{s(\tilde{a}),s(\tilde{b})}$$
$$= \frac{\mathrm{E}\left[(\tilde{a}-\mu_{\tilde{a}})(\tilde{b}-\mu_{\tilde{b}})\right]}{\sigma_{\tilde{a}}\,\sigma_{\tilde{b}}}$$

Invariant to positive scaling and shifts

Covariance

The covariance between $\tilde{a}$ and $\tilde{b}$ is

$$\begin{aligned} \mathrm{Cov}[\tilde{a}, \tilde{b}] &:= \mathrm{E}\left[(\tilde{a} - \mu_{\tilde{a}})\,(\tilde{b} - \mu_{\tilde{b}})\right] \\ &= \mathrm{E}[\tilde{a}\tilde{b}] - \mu_{\tilde{a}}\,\mu_{\tilde{b}} \end{aligned}$$

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

Correlation

If $\rho_{\tilde{a},\tilde{b}} > 0$ and $\mathrm{Cov}[\tilde{a},\tilde{b}] > 0$, $\tilde{a}$ and $\tilde{b}$ are positively correlated

If $\rho_{\tilde{a},\tilde{b}} = 0$ and $\mathrm{Cov}[\tilde{a},\tilde{b}] = 0$, $\tilde{a}$ and $\tilde{b}$ are uncorrelated

If $\rho_{\tilde{a},\tilde{b}} < 0$ and $\mathrm{Cov}[\tilde{a},\tilde{b}] < 0$, $\tilde{a}$ and $\tilde{b}$ are negatively correlated

Estimating covariance from data

Data: (x_1, y_1), (x_2, y_2), ..., (x_n, y_n)

$X := \{x_1, x_2, \ldots, x_n\}, \qquad Y := \{y_1, y_2, \ldots, y_n\}$

The sample covariance equals

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

where $m(X)$ and $m(Y)$ are the sample means of X and Y

Sample correlation coefficient

The sample correlation coefficient equals

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

where $v(X)$ and $v(Y)$ are the sample variances of X and Y

Height of NBA players

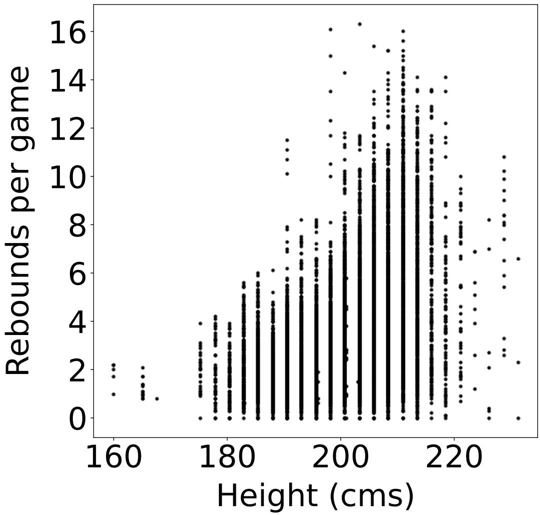
Data:

Height and offensive statistics of NBA players between 1996 and 2019

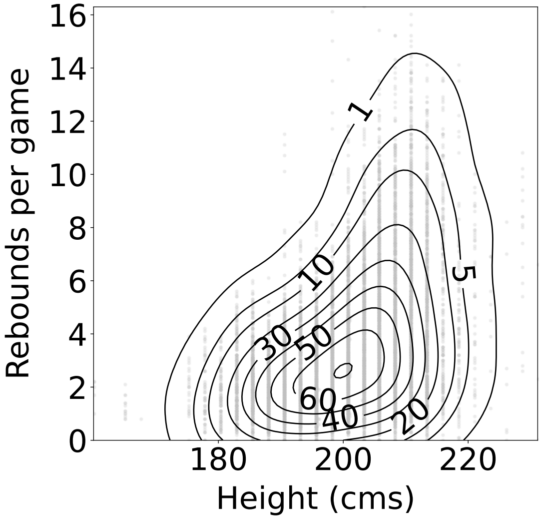
Goal:

Quantify linear dependence between rebounds/assists/points and height

Height and rebounds

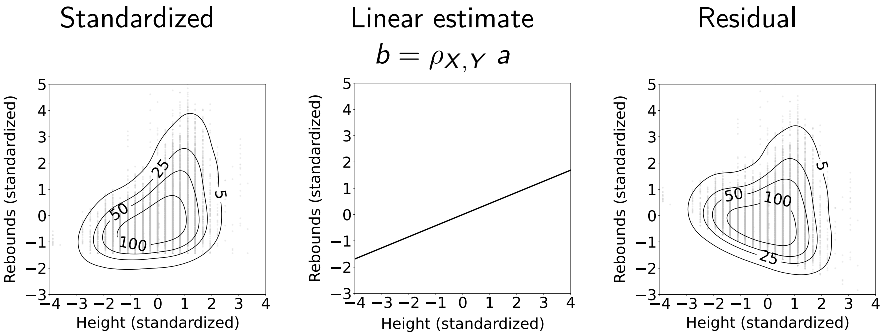


Height and rebounds

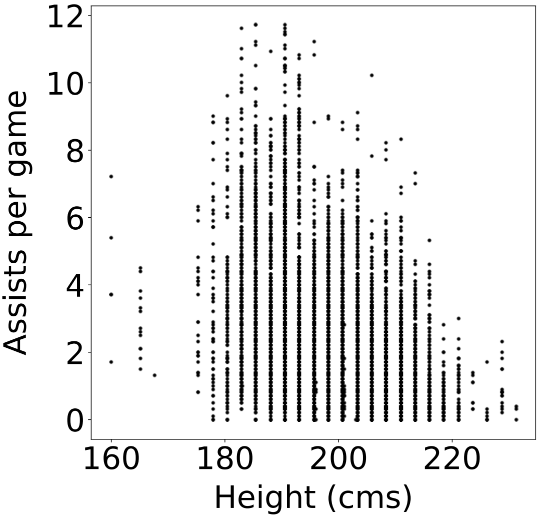


Height and rebounds

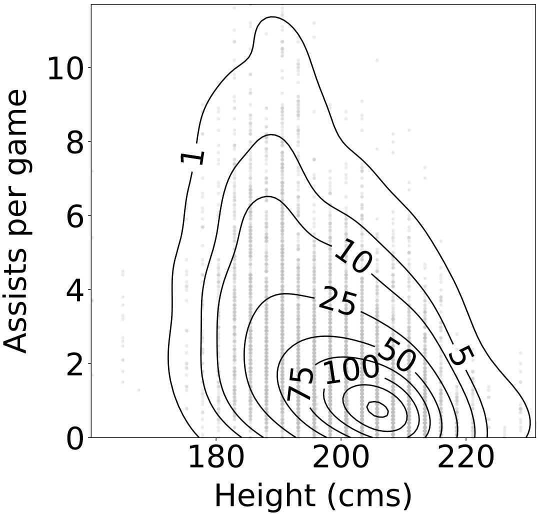
$$\rho_{\text{height,rebounds}} = 0.42$$



Height and assists

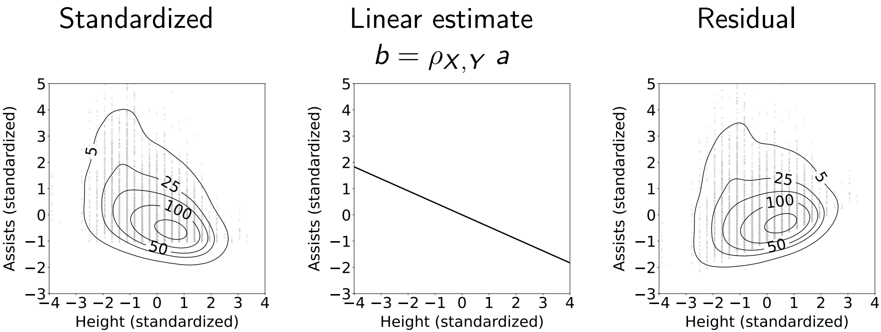


Height and assists

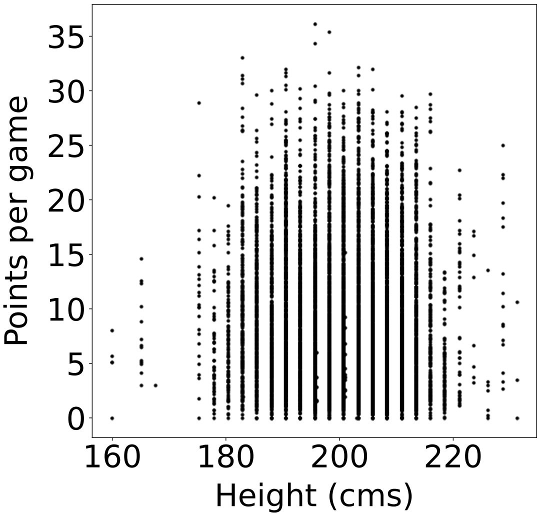


Height and assists

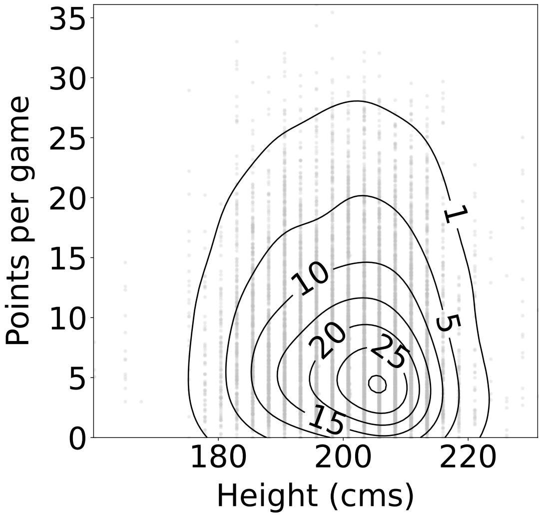
$\rho_{\text{height,assists}} = -0.46$



Height and points



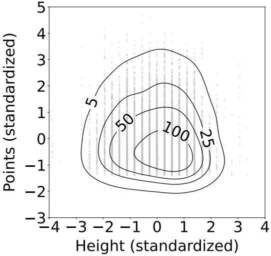
Height and points



Height and points

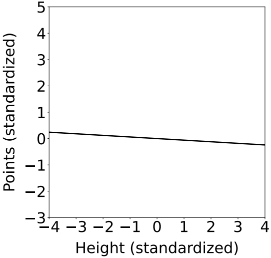
$$\rho_{\text{height,points}} = -0.06$$

Standardized

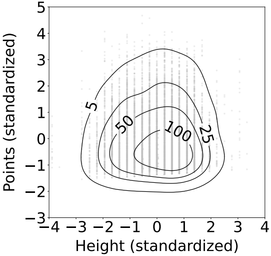


Linear estimate

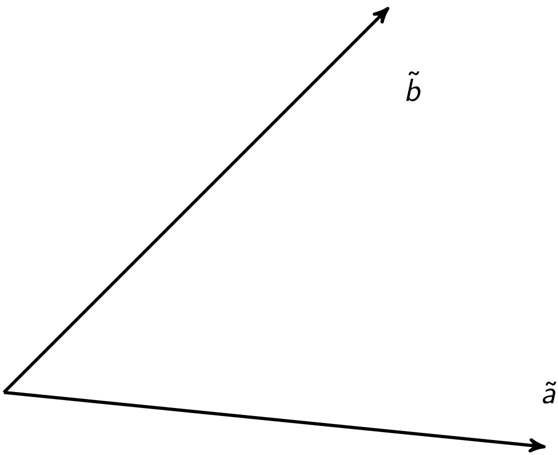
$b = \rho_{X,Y}\, a$



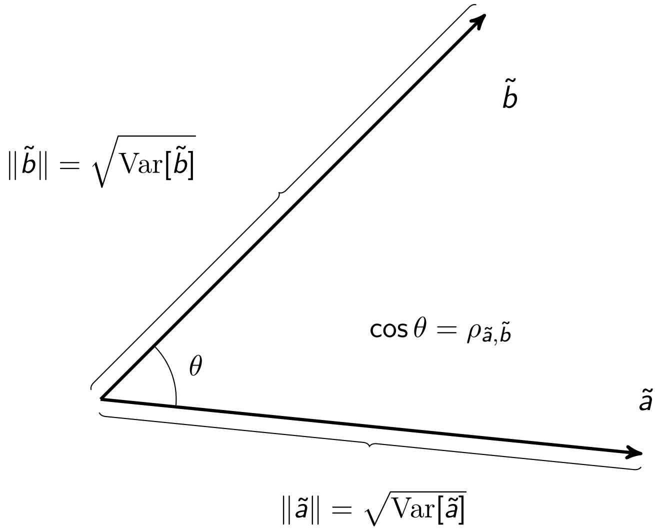
Residual



Geometric analysis of correlation



Covariance as an inner product



$-1 \leq \cos\theta \leq 1$ $\quad$ $-1 \leq \rho_{\tilde{a},\tilde{b}} \leq 1$

If $\cos\theta > 0$ vectors point in the same direction

If $\rho_{\tilde{a},\tilde{b}} > 0$ $\tilde{a}$ and $\tilde{b}$ are positively correlated

If $\cos\theta < 0$ vectors point in opposite directions

If $\rho_{\tilde{a},\tilde{b}} < 0$ $\tilde{a}$ and $\tilde{b}$ are negatively correlated

If $\cos\theta = 0$ vectors are orthogonal

If $\rho_{\tilde{a},\tilde{b}} = 0$ $\tilde{a}$ and $\tilde{b}$ are uncorrelated

Regression

Goal: Estimate quantity of interest (response) from observed features

Simple linear regression

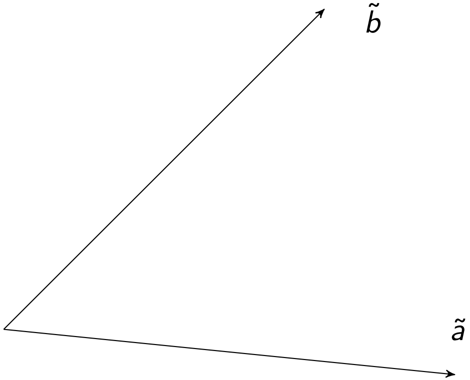
Single feature

Linear MMSE estimator:

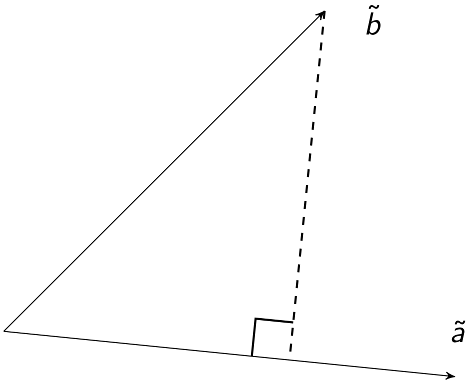
$$\tilde{b} = \sigma_{\tilde{b}} s(\tilde{b}) + \mu_{\tilde{b}} \approx \sigma_{\tilde{b}}\, \rho_{\mathsf{s}(\tilde{a}),\mathsf{s}(\tilde{b})}\, s(\tilde{a}) + \mu_{\tilde{b}}$$

$$= \frac{\sigma_{\tilde{b}}\, \rho_{\mathsf{s}(\tilde{a}),\mathsf{s}(\tilde{b})}\, (\tilde{a} - \mu_{\tilde{a}})}{\sigma_{\tilde{a}}} + \mu_{\tilde{b}}$$

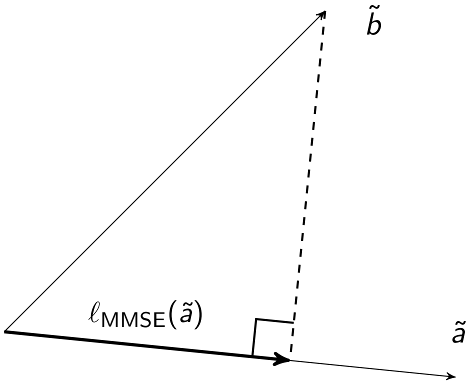
Vector collinear with $\tilde{a}$ closest to $\tilde{b}$?



Orthogonal projection



Linear minimum MSE estimator



Simple linear regression from data

Data: (x_1, y_1), (x_2, y_2), ..., (x_n, y_n)

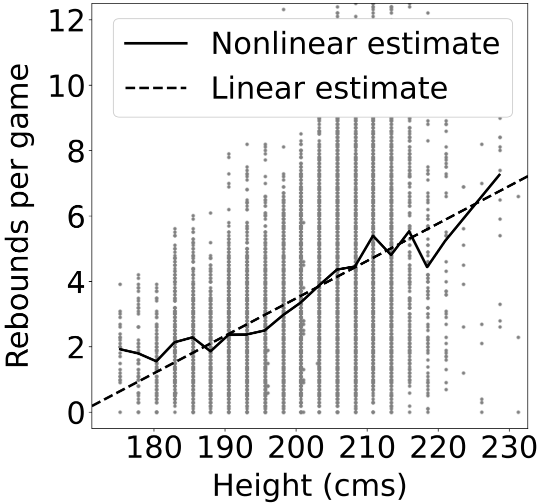
$$X := \{x_1, x_2, \ldots, x_n\}, \qquad Y := \{y_1, y_2, \ldots, y_n\}$$

Interpret x_i as sample from $\tilde{a}$, and y_i as sample from $\tilde{b}$

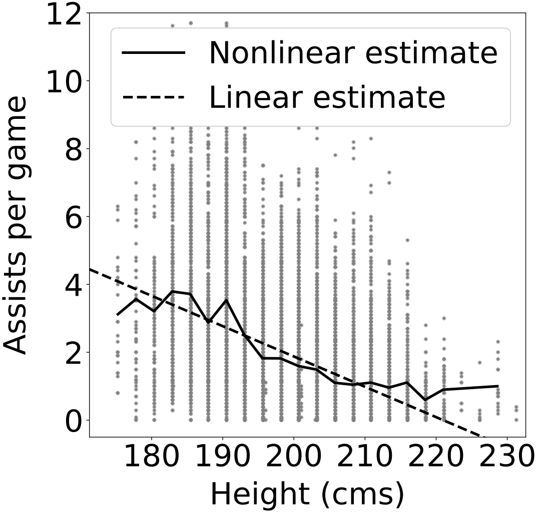
$$\begin{aligned}\ell_{\text{MMSE}}(a) &= \sigma_{\tilde{b}}\, \rho_{\tilde{a},\tilde{b}} \left(\frac{a - \mu_{\tilde{a}}}{\sigma_{\tilde{a}}} \right) + \mu_{\tilde{b}} \\ &\approx \sqrt{v(Y)} \rho_{X,Y} \left(\frac{x - m(X)}{\sqrt{v(X)}} \right) + m(Y)\end{aligned}$$

This is the ordinary least squares (OLS) estimator because it minimizes the residual sum of squares

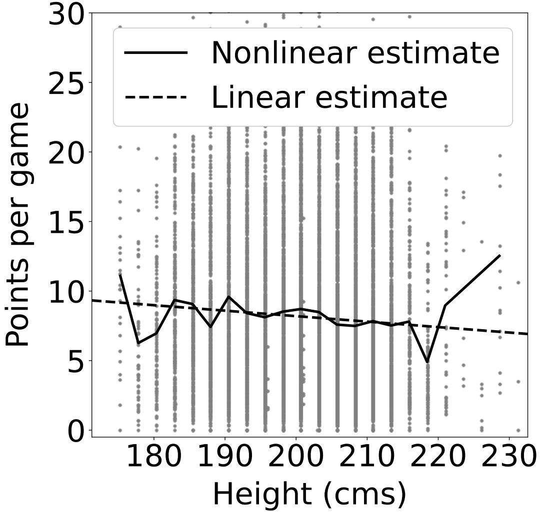
Height and rebounds



Height and assists



Height and points



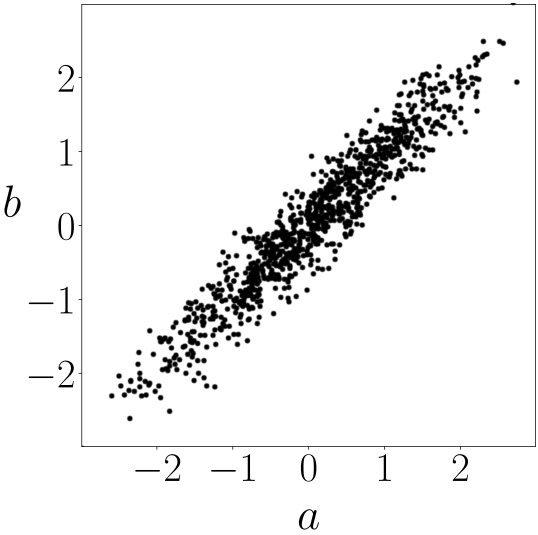
Properties of the correlation coefficient

The correlation coefficient is bounded between -1 and 1

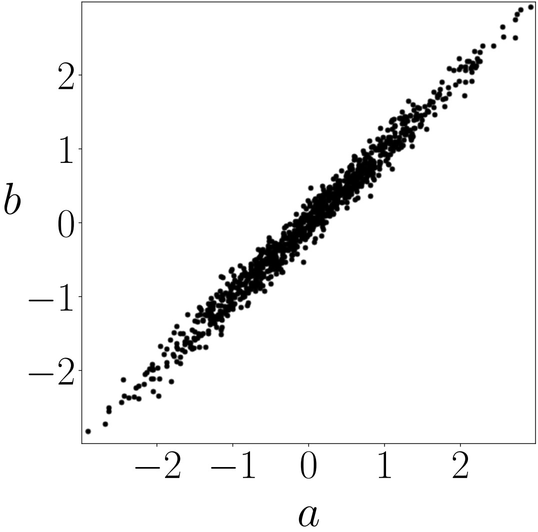
Properties of the correlation coefficient

If it equals ± 1, then there is complete linear dependence

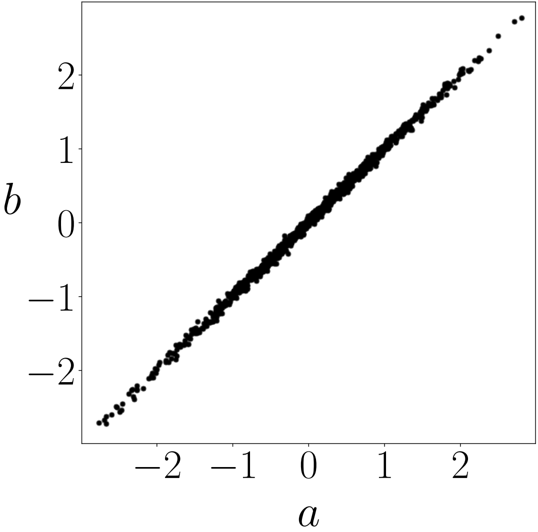
$\rho_{\tilde{a},\tilde{b}} = 0.95$



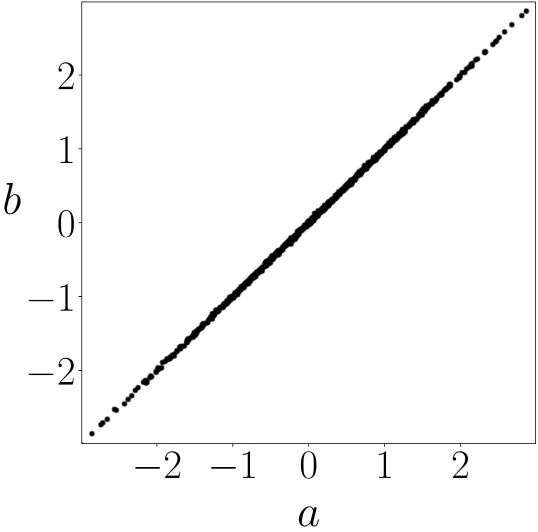
$\rho_{\tilde{a},\tilde{b}} = 0.99$



$\rho_{\tilde{a},\tilde{b}} = 0.999$



$\rho_{\tilde{a},\tilde{b}} = 0.9999$



Variance decomposition

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

$$\mathrm{Var}[\tilde{b} - \ell_{\mathsf{MMSE}}(\tilde{a})] = (1 - \rho^2_{\tilde{a},\tilde{b}})\mathrm{Var}\left[\tilde{b}\right]$$

$$\mathrm{Var}\left[\ell_{\mathsf{MMSE}}(\tilde{a})\right] = \rho^2_{\tilde{a},\tilde{b}}\mathrm{Var}\left[\tilde{b}\right]$$

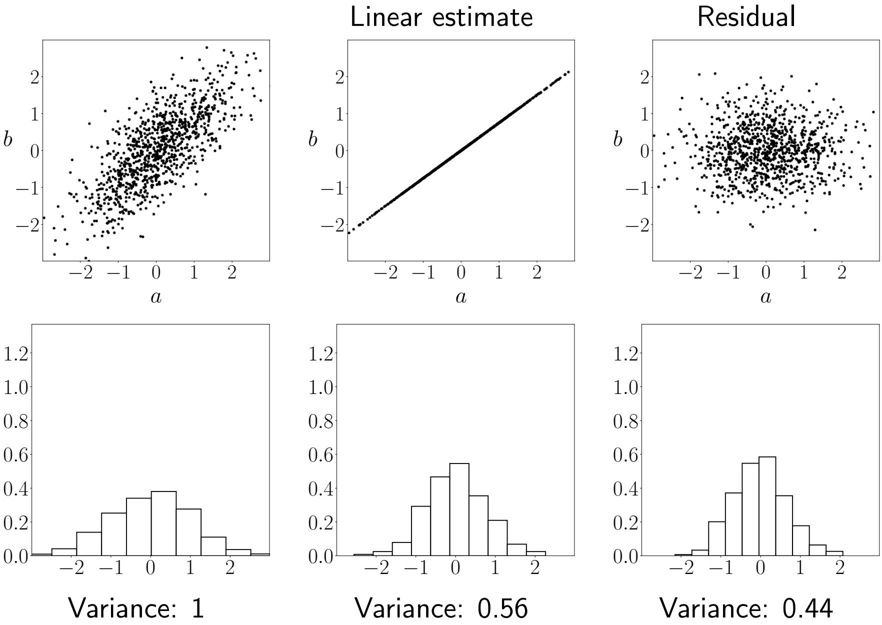
Coefficient of determination

$$R^2 := \frac{\mathrm{Var}\left[\ell_{\mathsf{MMSE}}(\tilde{a})\right]}{\mathrm{Var}[\tilde{b}]}$$

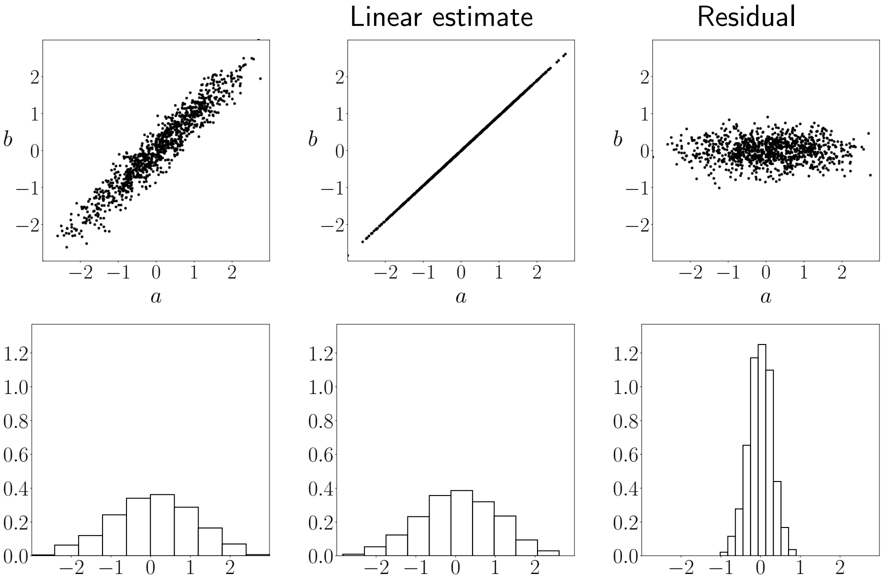
$$= \rho^2_{\tilde{a},\tilde{b}}$$

$$0 \leq R^2 \leq 1$$

$\rho_{\tilde{a},\tilde{b}} = 0.75,\ R^2 = 0.56$



$\rho_{\tilde{a},\tilde{b}} = 0.95,\ R^2 = 0.90$

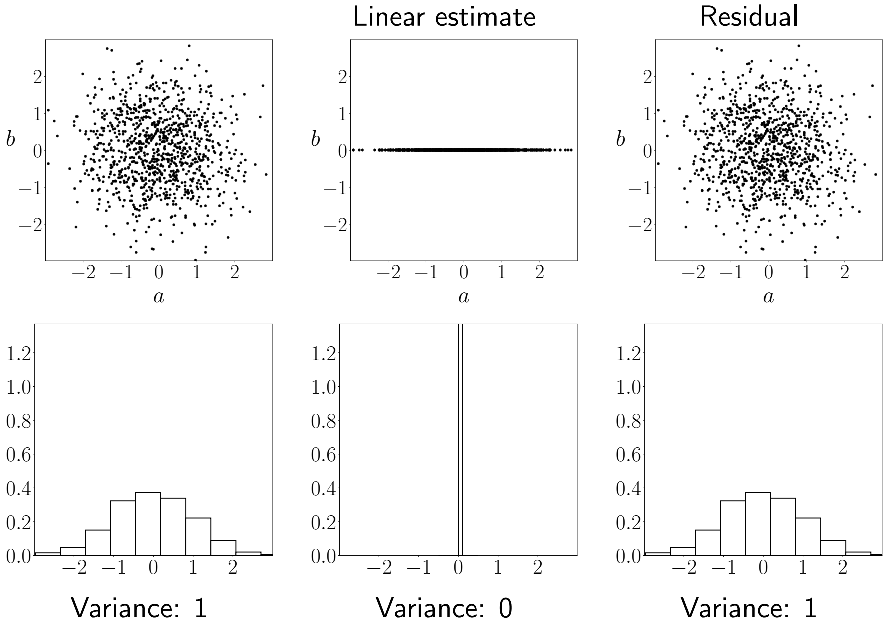


Variance: 1

Variance: 0.90

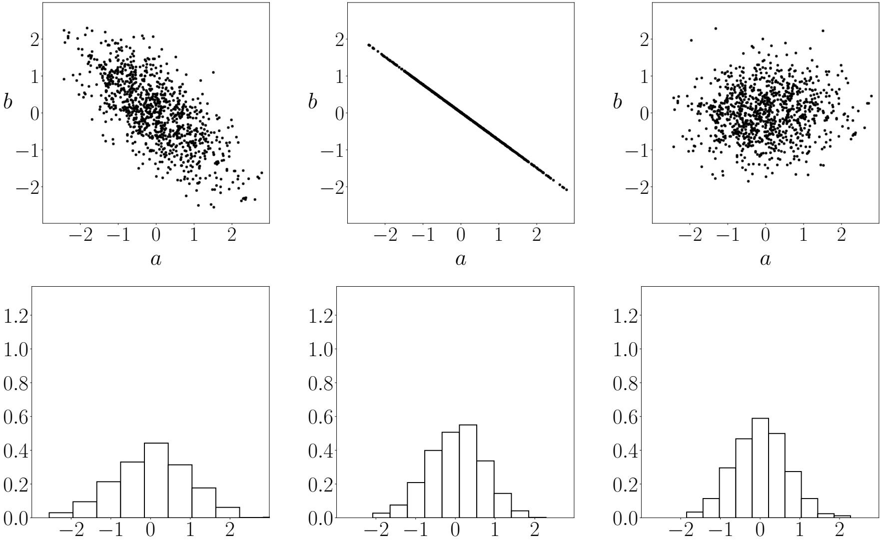
Variance: 0.10

$\rho_{\tilde{a},\tilde{b}} = 0,\ R^2 = 0$



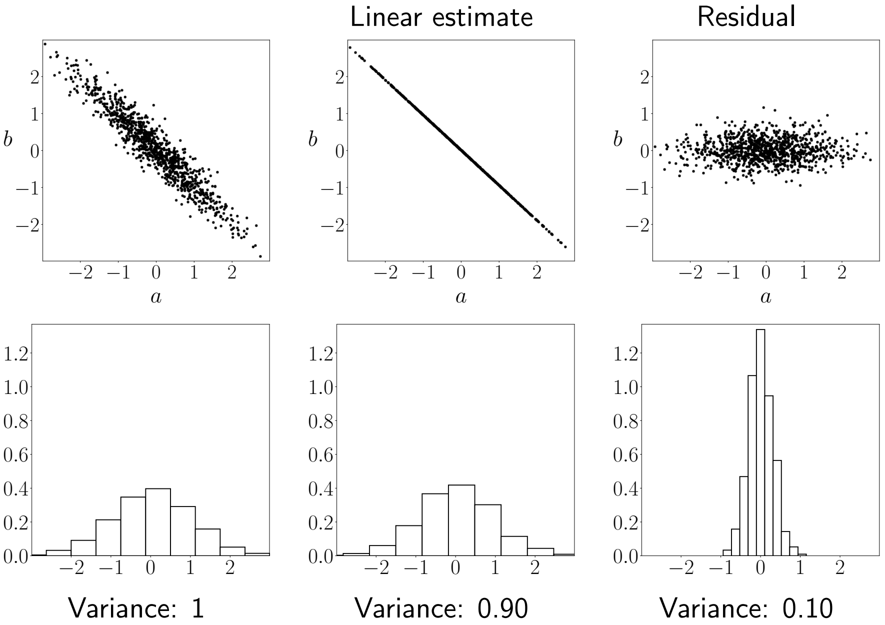
$\rho_{\tilde{a},\tilde{b}} = -0.75, \ R^2 = 0.56$

Linear estimate | Residual

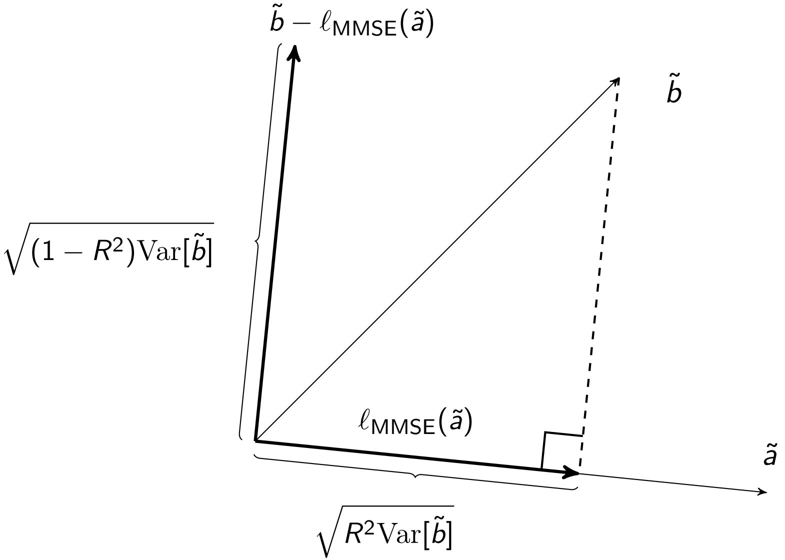


Variance: 1 | Variance: 0.56 | Variance: 0.44

$\rho_{\tilde{a},\tilde{b}} = -0.95,\ R^2 = 0.90$



Decomposition of variance

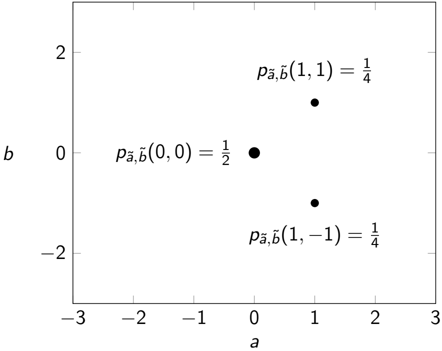


Independence implies uncorrelation

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

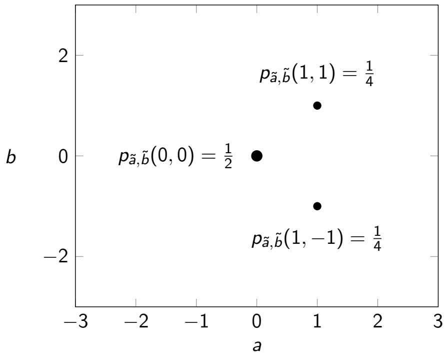
$$\mathrm{Cov}[\tilde{a}, \tilde{b}] = 0$$

Example



$$\mathrm{Cov}[\tilde{a}, \tilde{b}] = 0$$

Example



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

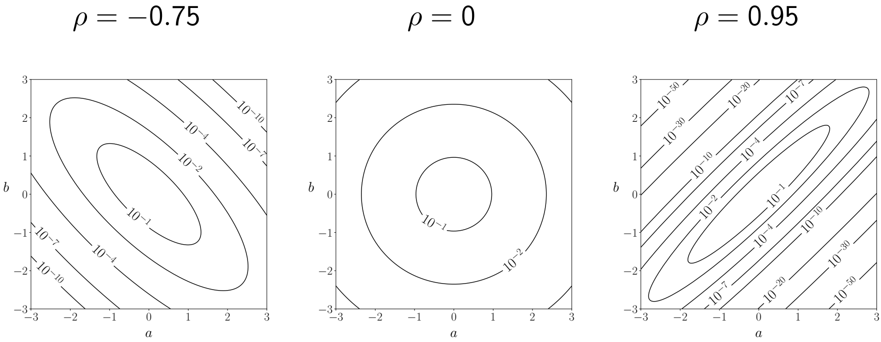
$$p_{\tilde b \,|\, \tilde a}(0 \,|\, 0) = 1$$

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

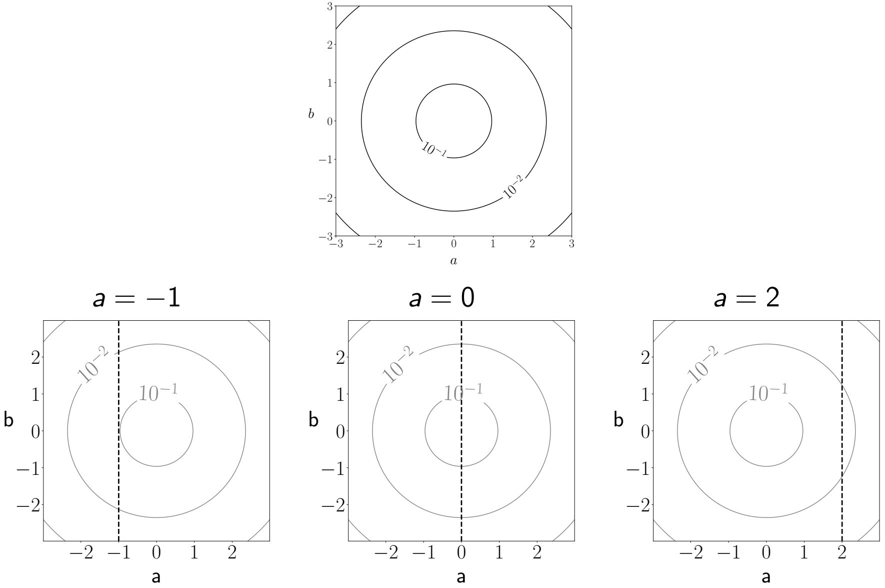
$$p_{\tilde b \,|\, \tilde a}(1 \,|\, 1) = \frac{1}{2} \qquad p_{\tilde b \,|\, \tilde a}(-1 \,|\, 1) = \frac{1}{2}$$

Not independent

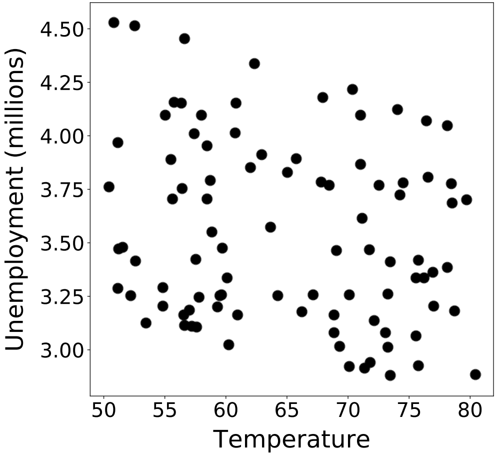
Gaussian random variables



Uncorrelation implies independence



Unemployment and temperature in Spain (2015-2022)



Correlation coefficient: -0.21

Would an increase in temperature decrease unemployment?

Causal inference

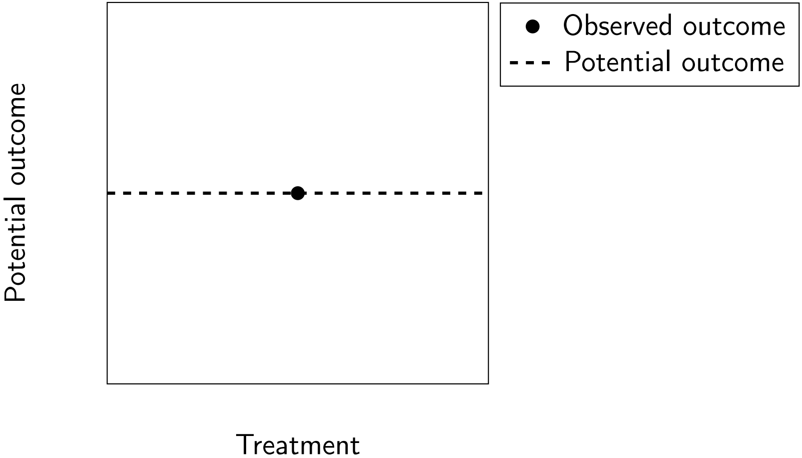
Key question: Does a treatment $\tilde{t}$ cause a certain outcome?

Potential outcome: $\widetilde{\mathrm{po}}_t$

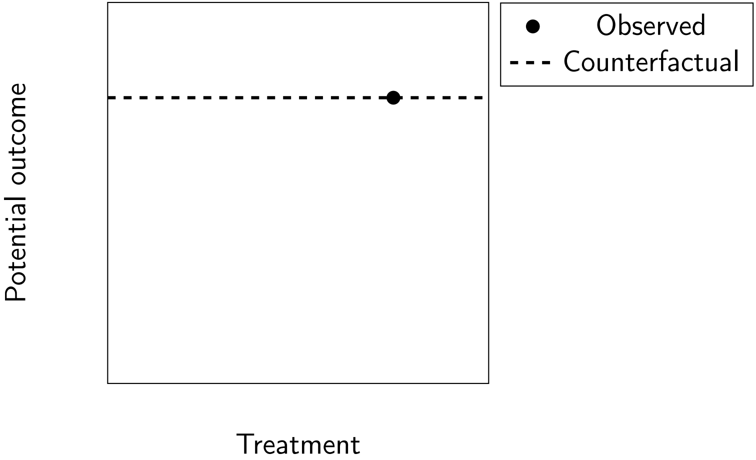
Observed data:

$$\tilde{y} := \widetilde{\mathrm{po}}_t \qquad \text{if} \qquad \tilde{t} = t$$

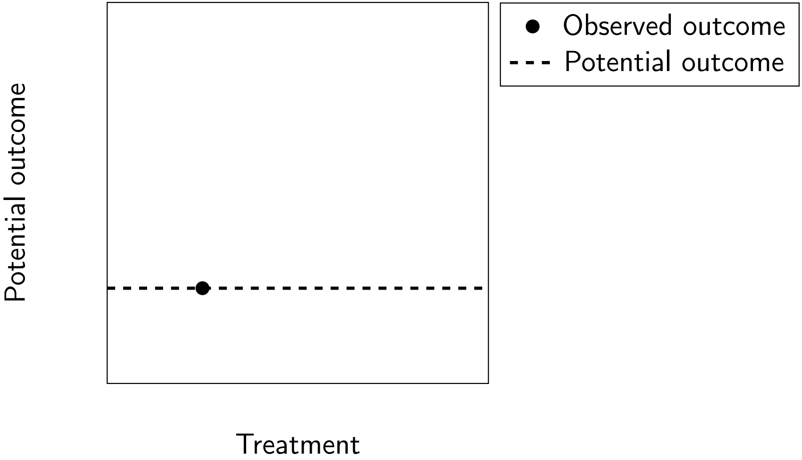
Potential outcomes



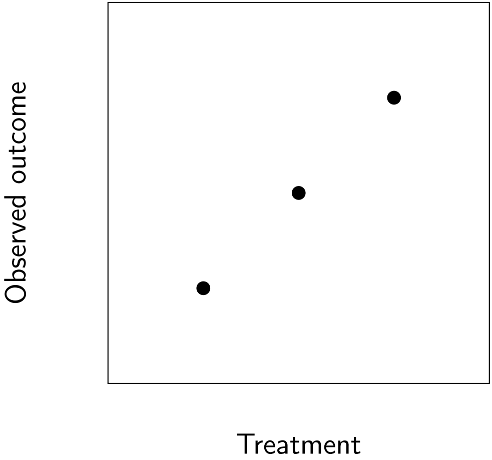
Potential outcomes



Potential outcomes



Observed data



Linear causal effect

For some constant $\beta \in \mathbb{R}$

$$\mathrm{E}\left[\widetilde{\mathsf{po}}_t\right] = \beta t$$

Key question: Can we estimate linear causal effects from data?

Idea

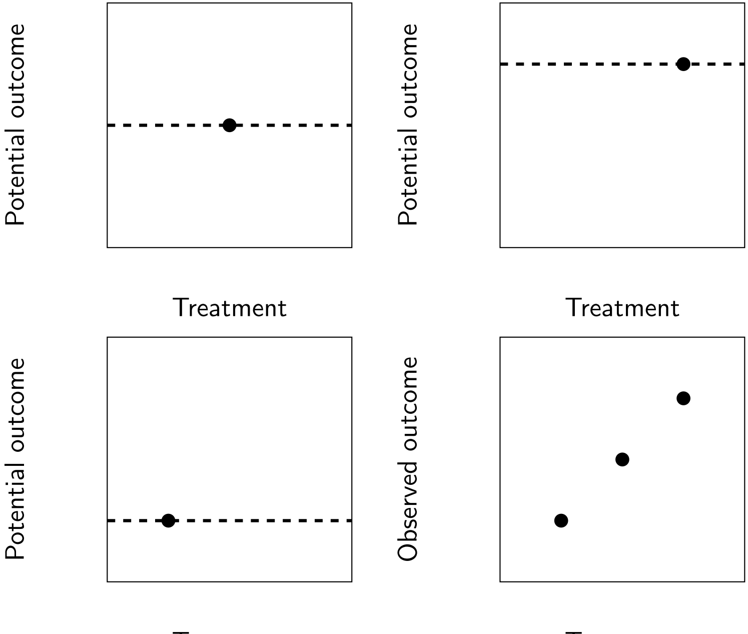
Use covariance between observed outcome $\tilde{y}$ and the treatment $\tilde{t}$

If $\widetilde{\text{po}}_t$ and $\tilde{t}$ are independent for all t

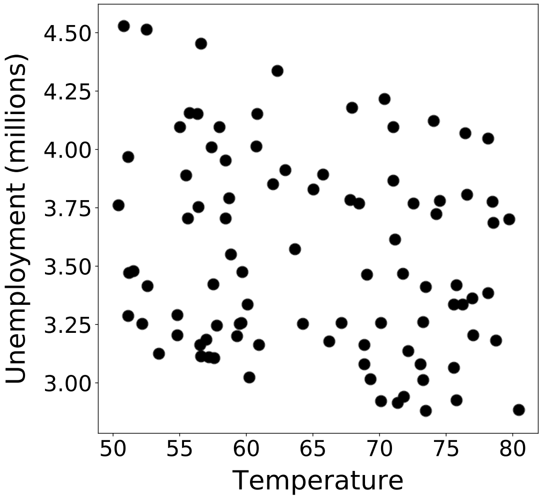
$$\operatorname{Cov}[\tilde{y}, \tilde{t}\,] = \beta$$

Assuming $\mathrm{E}[\tilde{t}] = 0$ and $\mathrm{E}[\tilde{t}^2] = 1$

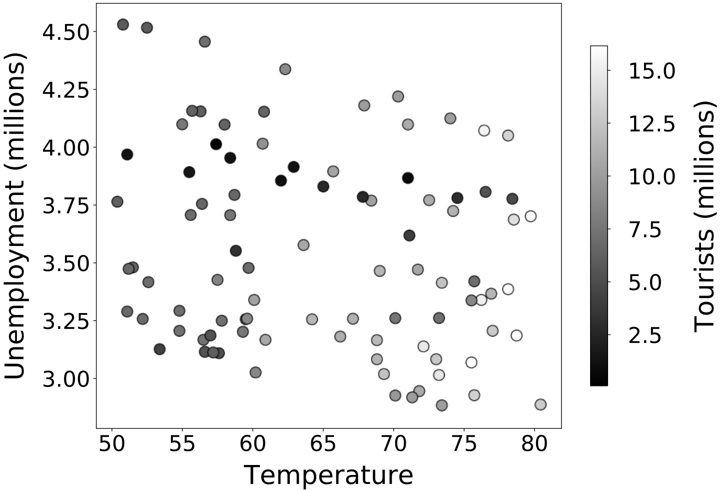
Why do we need independence?



Unemployment and temperature in Spain (2015-2022)



Unemployment and temperature in Spain (2015-2022)



Unobserved confounder

Potential outcome $\widetilde{\text{po}}_{t,c}$ depends on treatment $\tilde{t}$ and on confounder $\tilde{c}$

Observed data:

$$\tilde{y} := \widetilde{\text{po}}_{t,c} \qquad \text{if} \qquad \tilde{t} = t, \tilde{c} = c$$

For some constants $\beta, \gamma \in \mathbb{R}$

$$\mathrm{E}\left[\widetilde{\text{po}}_{t,c}\right] = \beta t + \gamma c$$

If $\widetilde{\text{po}}_{t,c}$ is independent from $(\tilde{t}, \tilde{c})$

$$\mathrm{Cov}\left[\tilde{y}, \tilde{t}\,\right] = \beta + \gamma \rho_{\tilde{t},\tilde{c}}$$

where $\tilde{t}$ and $\tilde{c}$ are standardized