# Correlation and Covariance

**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

# Correlation coefficient

Quantifies linear dependence between random variables with zero mean and unit variance

What about other random variables?

How do we compute it from data?

# Standardized variable

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

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

$$\mathrm{E}\left[\mathsf{s}(\tilde{a})\right] = \mathrm{E}\left[\frac{\tilde{a} - \mu_{\tilde{a}}}{\sigma_{\tilde{a}}}\right] = \frac{\mathrm{E}\left[\tilde{a}\right] - \mu_{\tilde{a}}}{\sigma_{\tilde{a}}} = 0$$

$$\begin{aligned} \mathrm{Var}\left[\mathsf{s}(\tilde{a})\right] = \mathrm{E}\left[\mathsf{s}(\tilde{a})^2\right] &= \mathrm{E}\left[\frac{(\tilde{a} - \mu_{\tilde{a}})^2}{\sigma_{\tilde{a}}^2}\right] \\ &= \frac{\mathrm{E}\left[(\tilde{a} - \mu_{\tilde{a}})^2\right]}{\sigma_{\tilde{a}}^2} = 1 \end{aligned}$$

# Linear dependence between random variables

The best linear approximation of $s(\tilde{b})$ given $s(\tilde{a})$ is $\rho_{s(\tilde{a}),s(\tilde{b})} s(\tilde{a})$

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

This turns out to be optimal!

$\rho_{s(\tilde{a}),s(\tilde{b})}$ quantifies affine dependence between $\tilde{a}$ and $\tilde{b}$

# Correlation coefficient

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

Invariant to scaling and shifts:

For any $\beta_1 > 0$, $\beta_2 > 0$, $\alpha_1$, $\alpha_2$, correlation coefficient between $\beta_1\tilde{a} + \alpha_1$ and $\beta_2\tilde{b} + \alpha_2$ is the same

# Covariance

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

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

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

# Correlation

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

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

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

# Cats and dogs

| | | Cats | | | |
|---|---|---|---|---|---|
| | | **0** | **1** | **2** | **3** |
| Dogs | **0** | 0.35 | 0.15 | 0.1 | 0.05 |
| | **1** | 0.2 | 0.05 | 0.03 | 0 |
| | **2** | 0.05 | 0.02 | 0 | 0 |

$$\mathrm{E}[\tilde{c}\,\tilde{d}] := \sum_{c=0}^{3} \sum_{d=0}^{2} c\, d\, p_{\tilde{c},\tilde{d}}(c,d) = 1 \cdot 0.05 + 2(0.03 + 0.02) = 0.15$$

$$\mathrm{E}[\tilde{c}] = 0.63 \qquad \mathrm{E}[\tilde{d}] = 0.42$$

$$\mathrm{Cov}[\tilde{c},\tilde{d}] = \mathrm{E}[\tilde{c}\,\tilde{d}] - \mathrm{E}[\tilde{c}]\mathrm{E}[\tilde{d}] = -0.115$$

# Cats and dogs

| | | Cats | | | |
|---|---|---|---|---|---|
| | | **0** | **1** | **2** | **3** |
| Dogs | **0** | 0.35 | 0.15 | 0.1 | 0.05 |
| | **1** | 0.2 | 0.05 | 0.03 | 0 |
| | **2** | 0.05 | 0.02 | 0 | 0 |

$$\mathrm{Cov}[\tilde{c}, \tilde{d}] = -0.115$$

$$\mathrm{Var}[\tilde{c}] = 0.793 \qquad \mathrm{Var}[\tilde{d}] = 0.383$$

$$\rho_{\tilde{c},\tilde{d}} := \frac{\mathrm{Cov}[\tilde{c}, \tilde{d}]}{\sqrt{\mathrm{Var}[\tilde{c}]\mathrm{Var}[\tilde{d}]}} = -0.208$$

# 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 $Y$

Correlation coefficient is optimal linear scaling between standardized random variables

# Standardized data

Data: $X := \{x_1, x_2, \ldots, x_n\}$

Standardized data:

$$s(x_i) := \frac{x_i - m(X)}{\sqrt{v(X)}} \qquad 1 \leq i \leq n$$

$$\rho_{X,Y} = \frac{1}{n-1} \sum_{i=1}^{n} \frac{(x_i - m(X))(y_i - m(Y))}{\sqrt{v(X)v(Y)}} = \frac{1}{n-1} \sum_{i=1}^{n} s(x_i)s(y_i)$$

For standardized data, sample mean $= 0$, sample variance $= 1$, so

$$\frac{1}{n-1} \sum_{i=1}^{n} s(x_i)^2 = 1$$

# Residual sum of squares

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

Goal: Approximate $s(y_i)$ by scaling $s(x_i)$

$$\begin{aligned}
\mathrm{RSS}(\beta) &:= \sum_{i=1}^{n} (s(y_i) - \beta s(x_i))^2 \\
&= \sum_{i=1}^{n} s(y_i)^2 + \beta^2 \sum_{i=1}^{n} s(x_i)^2 - 2\beta \sum_{i=1}^{n} s(x_i) s(y_i) \\
&= (n-1)(1 + \beta^2 - 2\beta \rho_{X,Y})
\end{aligned}$$

# Linear estimator

$$\mathrm{RSS}(\beta) = (n-1)(1 + \beta^2 - 2\beta\rho_{X,Y})$$

$$\mathrm{RSS}'(\beta) = 2(n-1)(\beta - \rho_{X,Y})$$

$$\mathrm{RSS}''(\beta) = 2(n-1)$$

$$\beta_{\mathrm{OLS}} = \rho_{X,Y}$$

# Height of NBA players

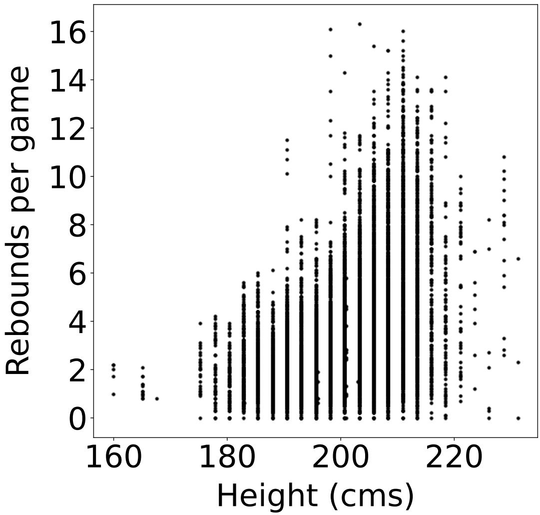
**Data:**

Height and offensive statistics of NBA players between 1996 and 2019

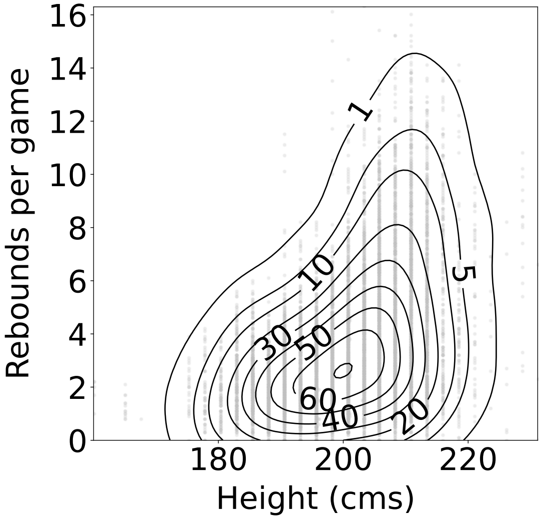
**Goal:**

Quantify linear dependence between rebounds/assists/points and height

# Height and rebounds

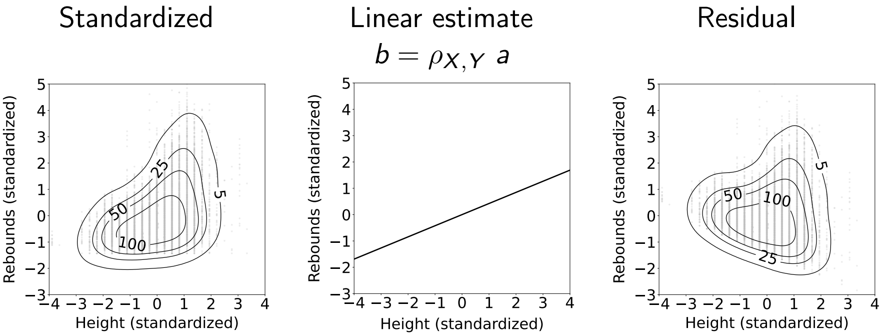


# Height and rebounds

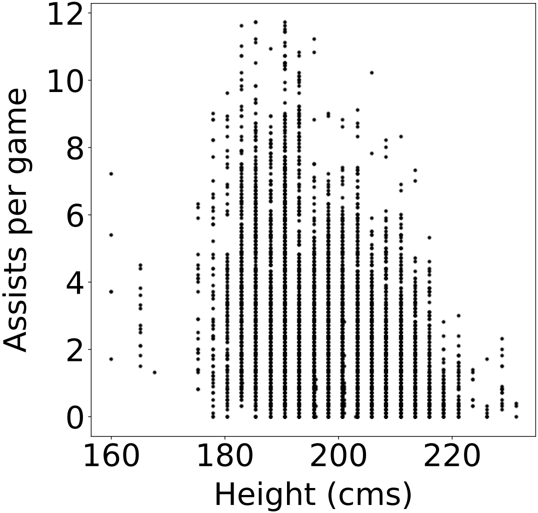


# Height and rebounds

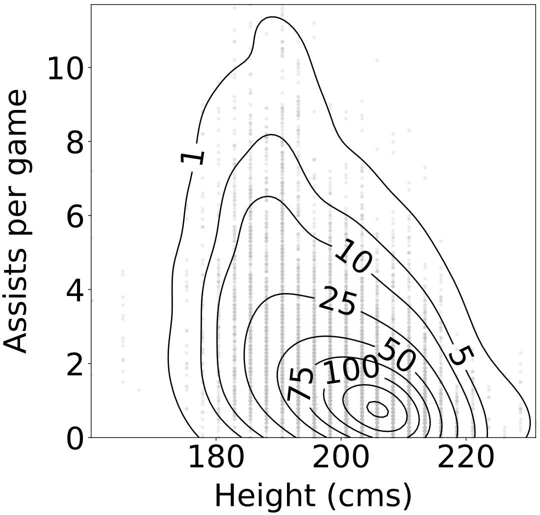
$\rho_{\text{height,rebounds}} = 0.42$



# Height and assists

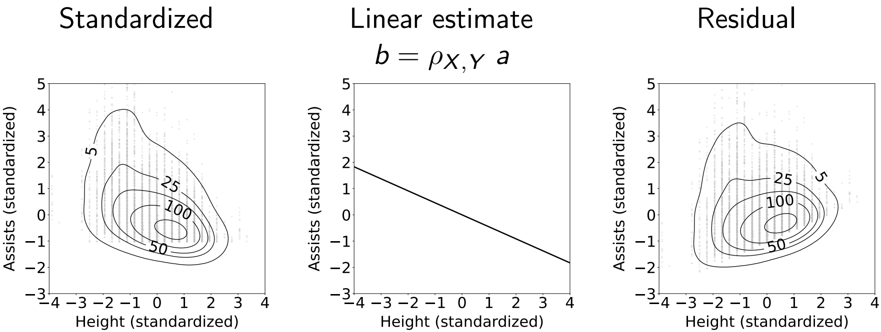


# Height and assists

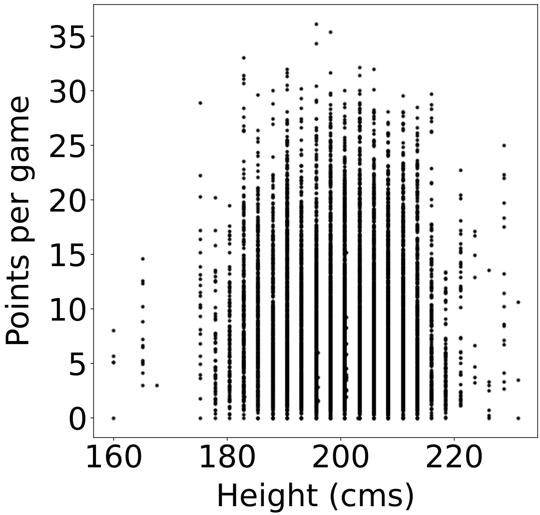


# Height and assists

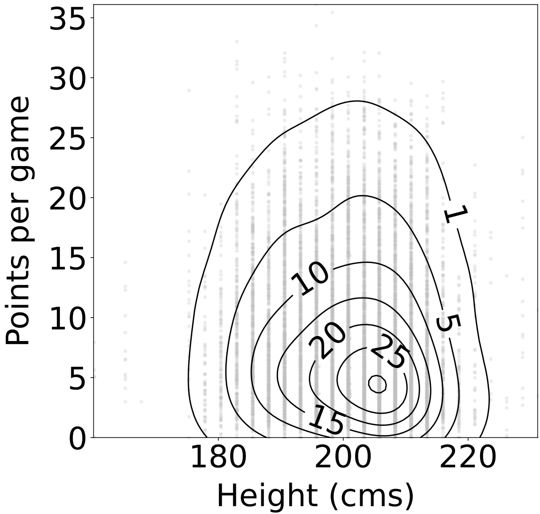
$$\rho_{\text{height,assists}} = -0.46$$



# Height and points

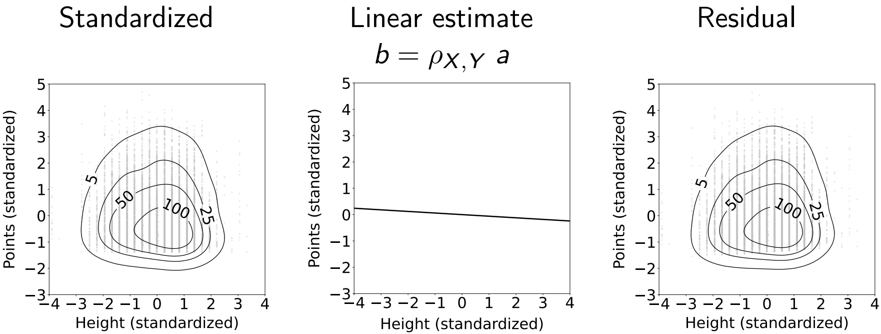


# Height and points



# Height and points

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



# What have we learned

General definition of correlation coefficient

Definition of covariance

How to estimate correlation from data