Properties of the Correlation Coefficient

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

Properties of the correlation coefficient

1. The correlation coefficient is bounded between -1 and 1
2. If it equals ± 1, then there is complete linear dependence
3. Its square equals the fraction of variance explained by the linear minimum MSE estimator

Linear MMSE estimator

For random variables $\tilde{a}$ and $\tilde{b}$ with means $\mu_{\tilde{a}}$ and $\mu_{\tilde{b}}$, variances $\sigma_{\tilde{a}}^2$ and $\sigma_{\tilde{b}}$, and correlation coefficient $\rho_{\tilde{a},\tilde{b}}$

The linear minimum MSE estimator of $\tilde{b}$ given $\tilde{a} = a$ is

$$\begin{aligned} \ell_{\text{MMSE}}(a) &= \sigma_{\tilde{b}}\, \rho_{\tilde{a},\tilde{b}} \left(\frac{a - \mu_{\tilde{a}}}{\sigma_{\tilde{a}}} \right) + \mu_{\tilde{b}} \\ &= \sigma_{\tilde{b}}\, \rho_{\tilde{a},\tilde{b}}\, s(a) + \mu_{\tilde{b}} \end{aligned}$$

Mean squared error

$$
\begin{aligned}
&\mathrm{E}\left[(\ell_{\mathsf{MMSE}}(\tilde{a}) - \tilde{b})^2\right] \\
&= \mathrm{E}\left[(\sigma_{\tilde{b}}\,\rho_{\tilde{a},\tilde{b}}\,s(\tilde{a}) + \mu_{\tilde{b}} - \tilde{b})^2\right] \\
&= \sigma_{\tilde{b}}^2\;\mathrm{E}\left[(\rho_{\tilde{a},\tilde{b}}\,s(\tilde{a}) - s(\tilde{b}))^2\right] \\
&= \sigma_{\tilde{b}}^2\left(\rho_{\tilde{a},\tilde{b}}^2\mathrm{E}[s(\tilde{a})^2] + \mathrm{E}[s(\tilde{b})^2] - 2\rho_{\tilde{a},\tilde{b}}\mathrm{E}\left[s(\tilde{a})s(\tilde{b})\right]\right) \\
&= \sigma_{\tilde{b}}^2\left(1 - \rho_{\tilde{a},\tilde{b}}^2\right)
\end{aligned}
$$

Property 1: $-1 \leq \rho_{\tilde{a},\tilde{b}} \leq 1$

$$\sigma_{\tilde{b}}^2 \left(1 - \rho_{\tilde{a},\tilde{b}}^2\right) = \mathrm{E}\left[(\tilde{b} - \ell_{\mathsf{MMSE}}(\tilde{a}))^2\right] \geq 0$$

$$\rho_{\tilde{a},\tilde{b}}^2 \leq 1$$

Small detour

If $\mathrm{E}\left[\tilde{a}^2\right] = 0$, then $\tilde{a} = 0$ with probability one

Property 2: $\rho_{\tilde{a},\tilde{b}} = \pm 1$ implies linear dependence

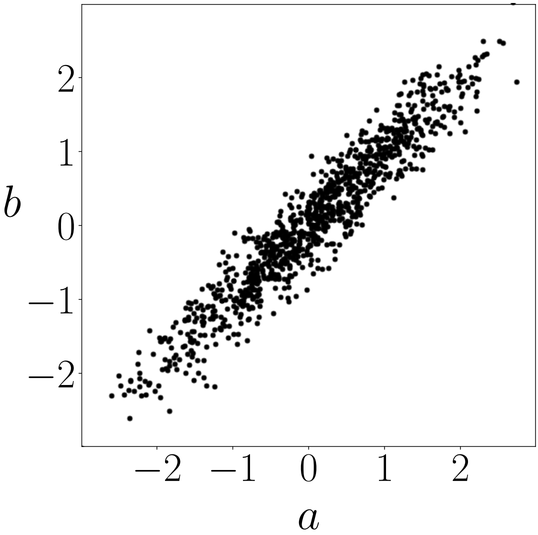
If $|\rho_{\tilde{a},\tilde{b}}| = 1$

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

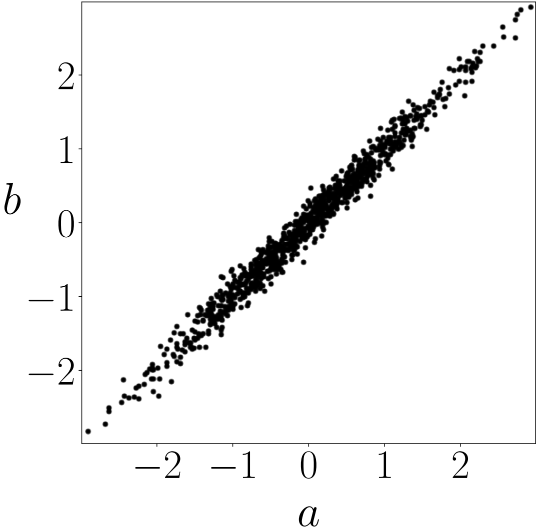
$$\ell_{\mathsf{MMSE}}(\tilde{a}) - \tilde{b} = 0$$

$$\tilde{b} = \ell_{\mathsf{MMSE}}(\tilde{a}) = \sigma_{\tilde{b}}\, \rho_{\tilde{a},\tilde{b}} \left(\frac{\tilde{a} - \mu_{\tilde{a}}}{\sigma_{\tilde{a}}}\right) + \mu_{\tilde{b}}$$

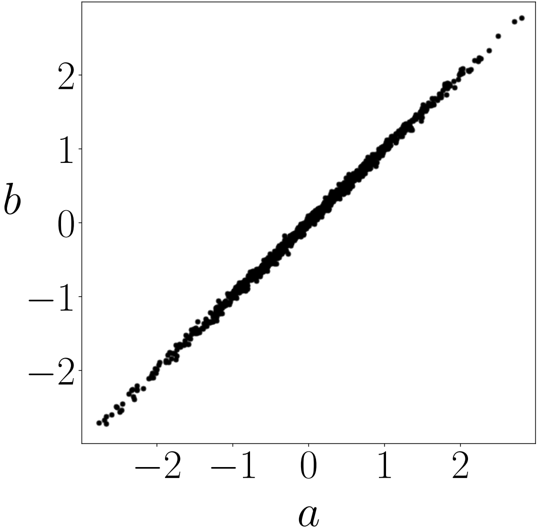
$\rho_{\tilde{a},\tilde{b}} = 0.95$



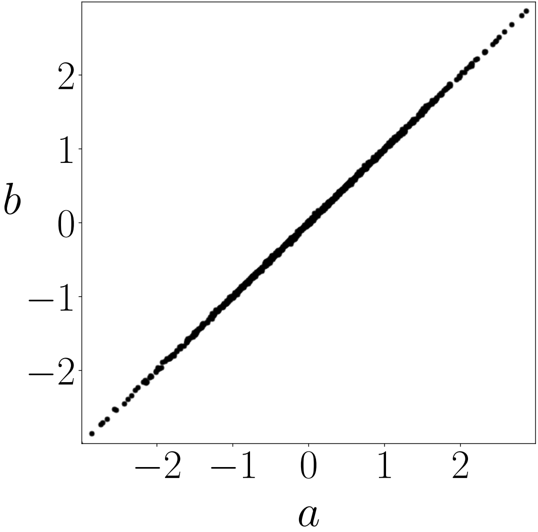
$\rho_{\tilde{a},\tilde{b}} = 0.99$



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



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



Property 3

Goal: Quantify how much variance is *explained* by linear MMSE estimator

$$\tilde{b} = \underbrace{\ell_{\mathsf{MMSE}}(\tilde{a})}_{\text{Linear MMSE estimate}} + \underbrace{\tilde{b} - \ell_{\mathsf{MMSE}}(\tilde{a})}_{\text{Residual}}$$

Variance of a sum

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

Uncorrelated random variables

If $\tilde{a}$ and $\tilde{b}$ are uncorrelated

$$\begin{aligned}
\operatorname{Var}[\tilde{a} + \tilde{b}] &= \operatorname{Var}[\tilde{a}] + \operatorname{Var}[\tilde{b}] + 2\operatorname{Cov}[\tilde{a}, \tilde{b}] \\
&= \operatorname{Var}[\tilde{a}] + \operatorname{Var}[\tilde{b}]
\end{aligned}$$

Residual

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

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

Decomposition of variance

$$\tilde{b} = \underbrace{\ell_{\text{MMSE}}(\tilde{a})}_{\text{Linear MMSE estimate}} + \underbrace{\tilde{b} - \ell_{\text{MMSE}}(\tilde{a})}_{\text{Residual}}$$

The residual is uncorrelated with $\tilde{a}$, so it is uncorrelated with any affine function of $\tilde{a}$, including $\ell_{\text{MMSE}}(\tilde{a})$

Property 3: Explained variance

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

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

$$\begin{aligned}\mathrm{Var}\left[\ell_{\mathsf{MMSE}}(\tilde{a})\right] &= \mathrm{Var}[\tilde{b}] - \mathrm{Var}[\tilde{b} - \ell_{\mathsf{MMSE}}(\tilde{a})] \\ &= \rho_{\tilde{a},\tilde{b}}^2\mathrm{Var}\left[\tilde{b}\right]\end{aligned}$$

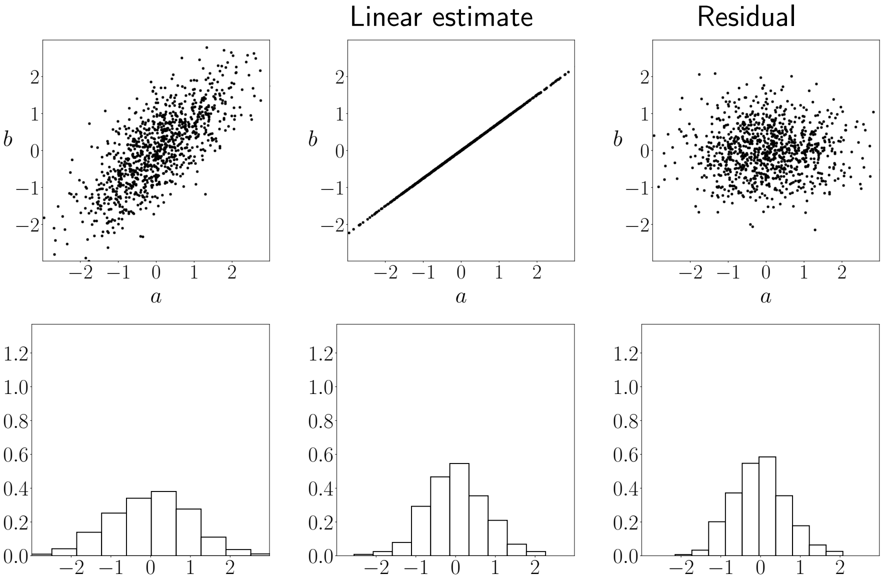
Coefficient of determination

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

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

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

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

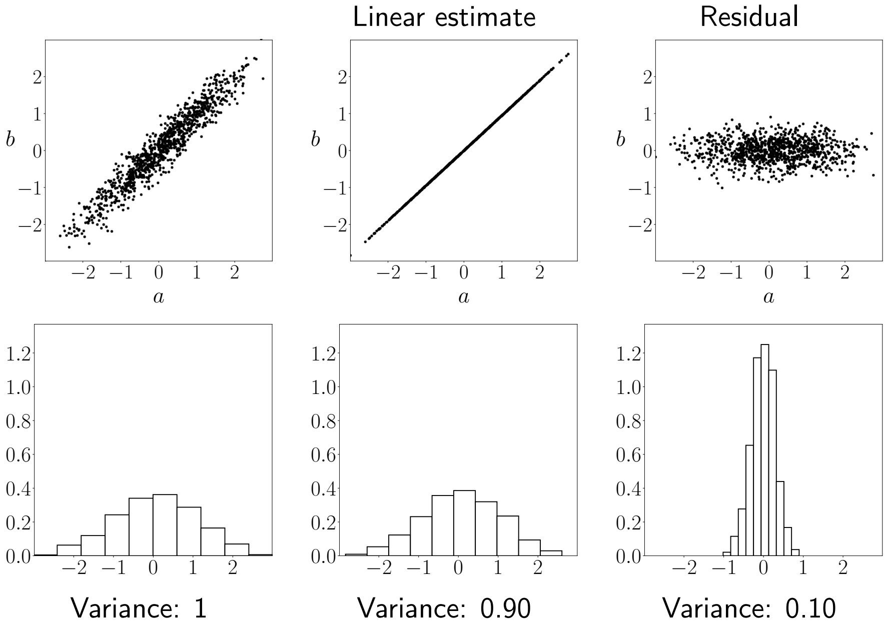


Variance: 1

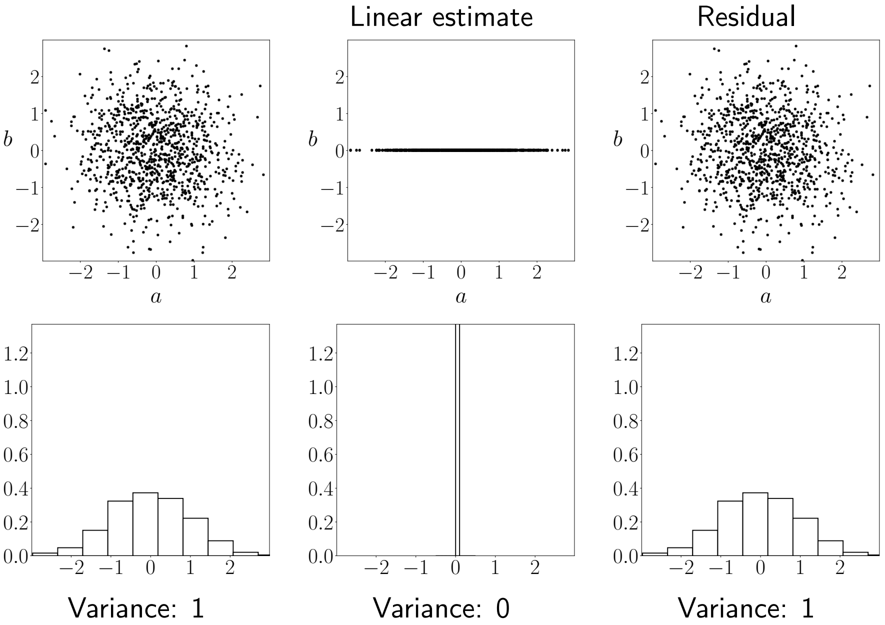
Variance: 0.56

Variance: 0.44

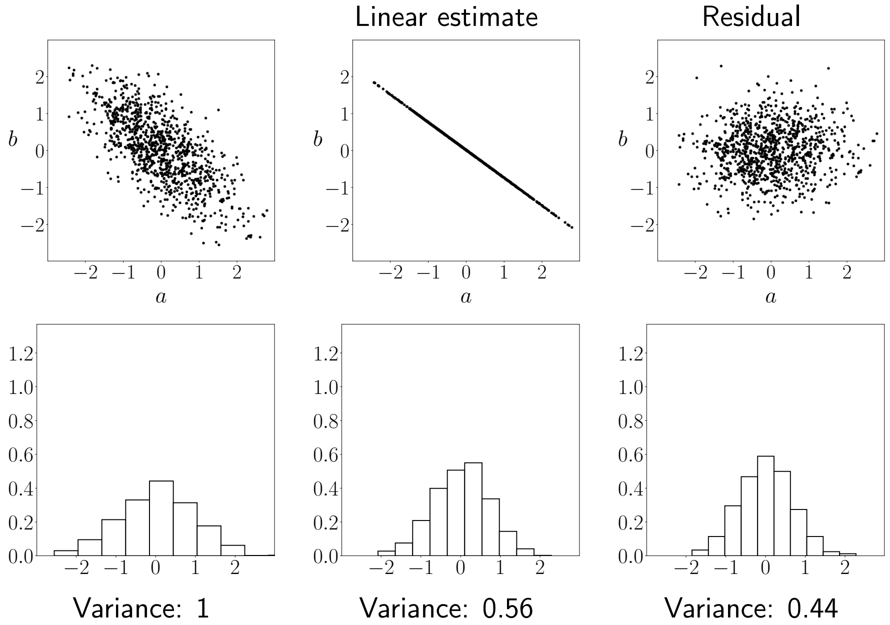
$\rho_{\tilde{a},\tilde{b}} = 0.95,\ R^2 = 0.90$



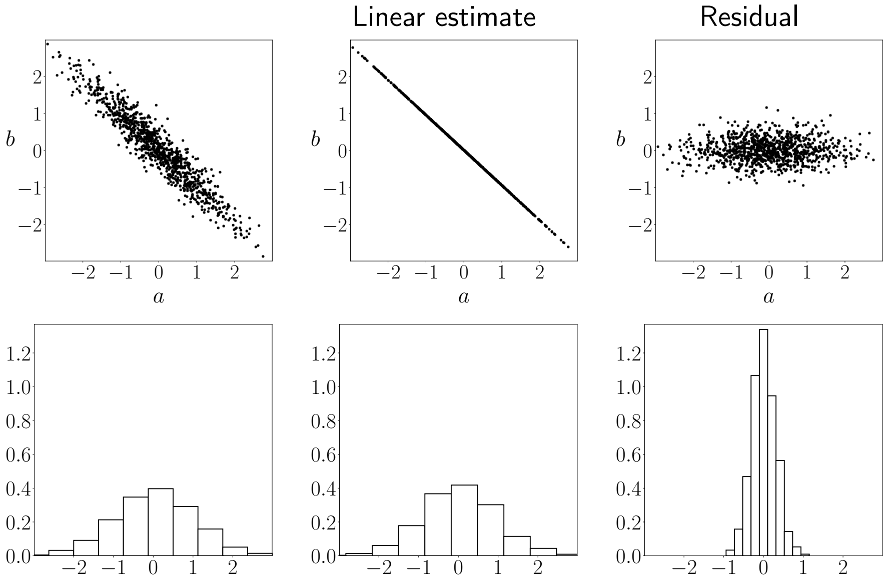
$\rho_{\tilde{a},\tilde{b}} = 0, \; R^2 = 0$



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



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



Variance: 1

Variance: 0.90

Variance: 0.10

Height of NBA players

Data:

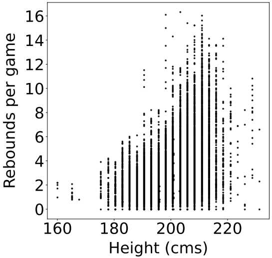
Height and offensive statistics of NBA players between 1996 and 2019

Goal:

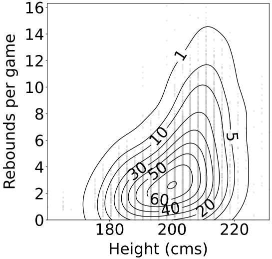
Quantify linear dependence between rebounds/assists/points and height

Rebounds and height

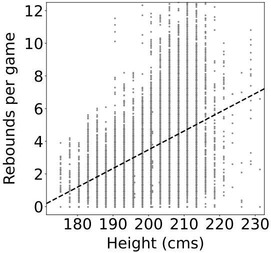
Scatterplot



Density estimate

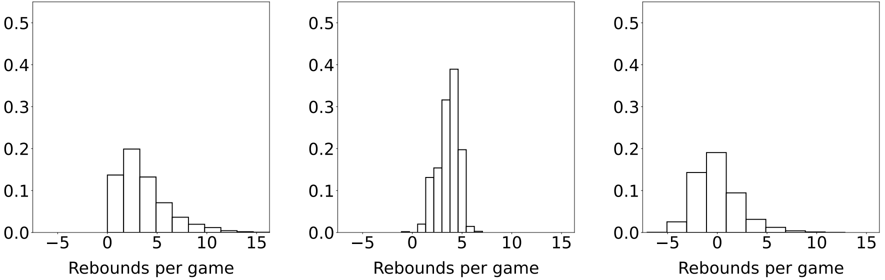


OLS estimator



Rebounds and height: $R^2 = 0.176$

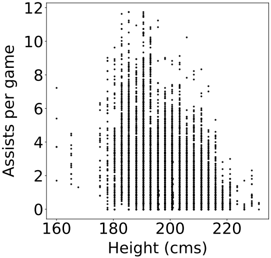
OLS estimator | Residual



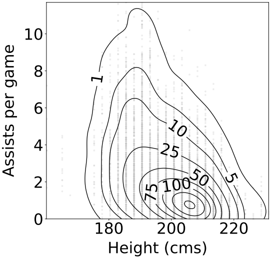
Variance: 6.19 | Variance: 1.10 | Variance: 5.09

Assists and height

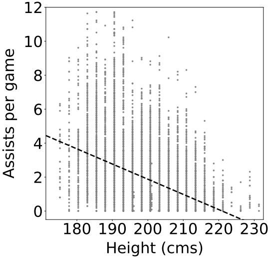
Scatterplot



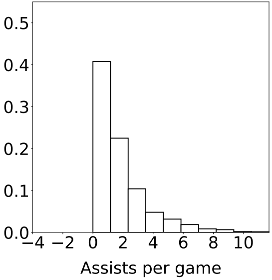
Density estimate



OLS estimator

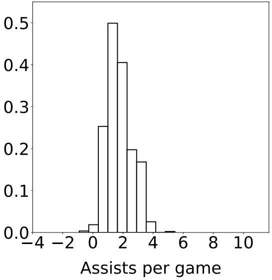


Assists and height: $R^2 = 0.212$



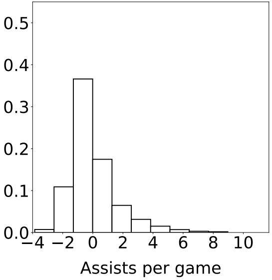
Variance: 3.21

OLS estimator



Variance: 0.67

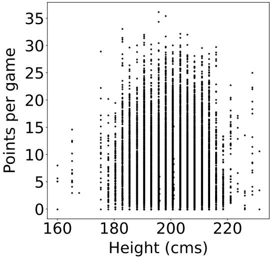
Residual



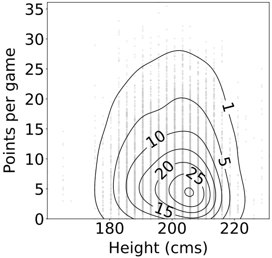
Variance: 2.54

Points and height

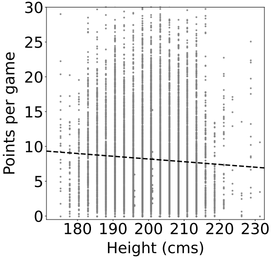
Scatterplot



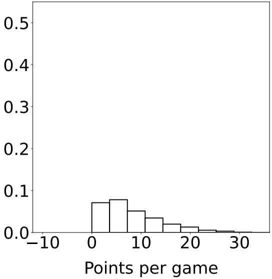
Density estimate



OLS estimator

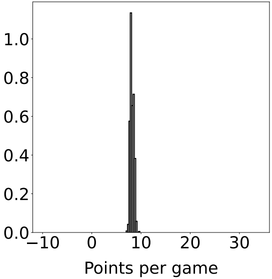


Points and height: $R^2 = 0.036$



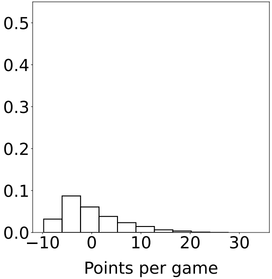
Variance: 35.47

OLS estimator



Variance: 0.13

Residual



Variance: 35.34

What have we learned

1. The correlation coefficient is bounded between -1 and 1
2. If it equals ± 1, this implies complete linear dependence
3. Its square equals the fraction of variance explained by the linear minimum MSE estimator