Simple Linear Regression

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

Regression

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

Simple linear regression

Single feature

Affine estimator

We model the feature $\tilde{a}$ and the response $\tilde{b}$ as random variables

$$\tilde{b} \approx \beta \tilde{a} + \alpha$$

Plan

Derive optimal linear estimator

Explain how to compute it from data

Compare it to nonlinear estimator

Standardized variable

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

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

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

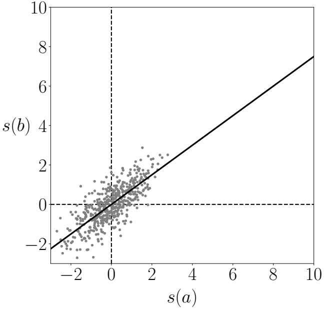
$$\mathrm{Var}\left[s(\tilde{a})\right] = 1$$

Linear dependence between random variables

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

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

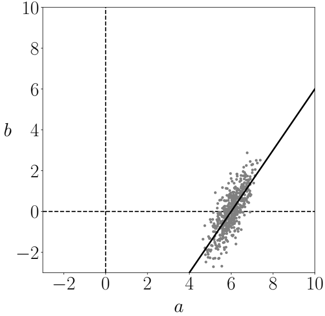
$\mu_{\tilde{b}} + \sigma_{\tilde{b}}\, \rho_{\tilde{a},\tilde{b}}\, s(\tilde{a})$



$$s(b) = \rho_{\tilde{a},\tilde{b}}\, s(a)$$

$\mu_{\tilde{a}} := 6,\ \sigma_{\tilde{a}} := 0.5$

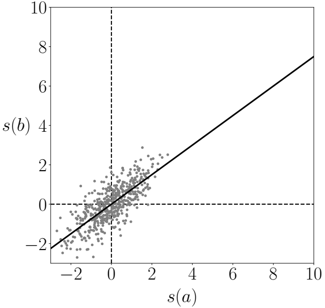
$\mu_{\tilde{b}} := 0,\ \sigma_{\tilde{b}} := 1$



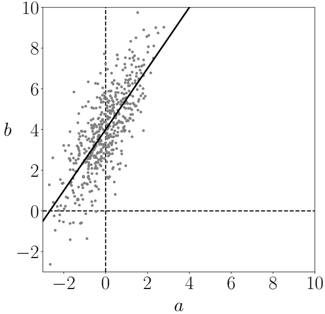
$$b = \frac{\rho_{\tilde{a},\tilde{b}}\,(a - \mu_{\tilde{a}})}{\sigma_{\tilde{a}}}$$

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

$\mu_{\tilde{a}} := 0,\ \sigma_{\tilde{a}} := 1$
$\mu_{\tilde{b}} := 4,\ \sigma_{\tilde{b}} := 2$



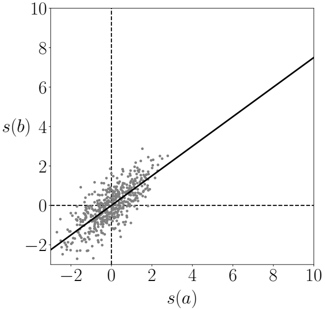
$s(b) = \rho_{\tilde{a},\tilde{b}}\, s(a)$

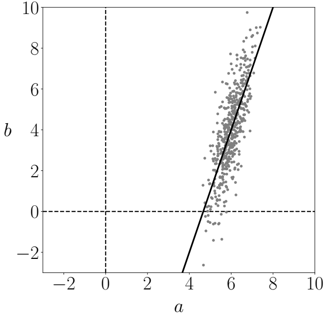


$b = \sigma_{\tilde{b}}\, \rho_{\tilde{a},\tilde{b}}\, a + \mu_{\tilde{b}}$

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

$\mu_{\tilde{a}} := 6,\ \sigma_{\tilde{a}} := 0.5$
$\mu_{\tilde{b}} := 4,\ \sigma_{\tilde{b}} := 2$





$s(b) = \rho_{\tilde{a},\tilde{b}}\, s(a)$

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

Linear MMSE estimator

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

Minimum MSE constant estimate

Best constant estimate of $\tilde{x}$?

$$\arg\min_{c \in \mathbb{R}} \mathrm{E}\left[(c - \tilde{x})^2\right] = \mathrm{E}[\tilde{x}]$$

Additive constant

$$\begin{aligned}\alpha^*(\beta) &:= \arg\min_{\alpha} \mathrm{E}\left[(\tilde{b} - \beta\tilde{a} - \alpha)^2\right] \\ &= \mathrm{E}\left[\tilde{b} - \beta\tilde{a}\right] \\ &= \mathrm{E}\left[\tilde{b}\right] - \beta\mathrm{E}\left[\tilde{a}\right] \\ &= \mu_{\tilde{b}} - \beta\mu_{\tilde{a}}\end{aligned}$$

Linear coefficient

For any β and any α, $\mathsf{MSE}(\beta, \alpha) \geq \mathsf{MSE}(\beta, \alpha^*(\beta))$

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

Mean squared error

$$\begin{aligned}
&\mathsf{MSE}(\beta, \alpha^*(\beta)) \\
&= \mathrm{E}\left[(\tilde{b} - \beta\tilde{a} - \alpha^*(\beta))^2\right] \\
&= \mathrm{E}\left[(\tilde{b} - \beta\tilde{a} - \mu_{\tilde{b}} + \beta\mu_{\tilde{a}})^2\right] \\
&= \mathrm{E}\left[(\tilde{b} - \mu_{\tilde{b}})^2\right] + \beta^2\mathrm{E}\left[(\tilde{a} - \mu_{\tilde{a}})^2\right] - 2\beta\mathrm{E}\left[(\tilde{a} - \mu_{\tilde{a}})(\tilde{b} - \mu_{\tilde{b}})\right] \\
&= \sigma_{\tilde{b}}^2 + \sigma_{\tilde{a}}^2\beta^2 - 2\mathrm{Cov}[\tilde{a}, \tilde{b}]\beta
\end{aligned}$$

Mean squared error

$$\mathsf{MSE}(\beta, \alpha^*(\beta)) = \sigma_{\tilde{b}}^2 + \sigma_{\tilde{a}}^2 \beta^2 - 2\mathrm{Cov}[\tilde{a}, \tilde{b}]\beta$$

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

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

$$\beta_{\mathsf{MMSE}} = \frac{\mathrm{Cov}[\tilde{a}, \tilde{b}]}{\sigma_{\tilde{a}}^2} = \frac{\rho_{\tilde{a},\tilde{b}}\,\sigma_{\tilde{b}}}{\sigma_{\tilde{a}}}$$

Linear MMSE estimator

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

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

$$\mathrm{E}[\tilde{c}] = 0.63 \qquad \mathrm{E}[\tilde{d}] = 0.42 \qquad \mathrm{Var}[\tilde{c}] = 0.793 \qquad \mathrm{Var}[\tilde{d}] = 0.383$$

$$\mathrm{Cov}[\tilde{c}, \tilde{d}] = -0.115 \qquad \rho_{\tilde{c},\tilde{d}} := \frac{\mathrm{Cov}[\tilde{c}, \tilde{d}]}{\sqrt{\mathrm{Var}[\tilde{c}]\mathrm{Var}[\tilde{d}]}} = -0.208$$

$$\ell_{\mathrm{MMSE}}(d) = \sigma_{\tilde{c}}\, \rho_{\tilde{c},\tilde{d}} \left(\frac{d - \mathrm{E}[\tilde{d}]}{\sqrt{\mathrm{Var}[\tilde{d}]}} \right) + \mathrm{E}[\tilde{c}] = -0.3d + 0.756$$

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

$$\ell_{\text{MMSE}}(d) = -0.3d + 0.756$$

$$\ell_{\text{MMSE}}(0) = 0.756 \qquad \ell_{\text{MMSE}}(1) = 0.456 \qquad \ell_{\text{MMSE}}(2) = 0.156$$

$$\mathrm{E}\left[\left(\tilde{c} - \ell_{\text{MMSE}}(\tilde{d})\right)^2\right] = \sum_{c=0}^{3}\sum_{d=0}^{2} p_{\tilde{c},\tilde{d}}(c,d)\,(c + 0.3d - 0.756)^2 = 0.759$$

Minimum MSE estimator? Conditional mean

$p_{\tilde{c}\,|\,\tilde{d}}(c\,|\,d)$

Dogs (d) \ Cats (c)	**0**	**1**	**2**	**3**
0	0.54	0.23	0.15	0.08
1	0.71	0.18	0.11	0
2	0.71	0.29	0	0

$\mu_{\tilde{c}\,|\,\tilde{d}}(0) = 0.77$ $\quad\mu_{\tilde{c}\,|\,\tilde{d}}(1) = 0.4$ $\quad\mu_{\tilde{c}\,|\,\tilde{d}}(2) = 0.29$

$\ell_{\text{MMSE}}(0) = 0.756$ $\quad\ell_{\text{MMSE}}(1) = 0.456$ $\quad\ell_{\text{MMSE}}(2) = 0.156$

$$\mathrm{E}\left[\left(\tilde{c} - \mu_{\tilde{c}\,|\,\tilde{d}}(\tilde{d})\right)^2\right] = 0.76 < 0.79 = \mathrm{E}\left[\left(\tilde{c} - \ell_{\text{MMSE}}(\tilde{d})\right)^2\right]$$

Gaussian random variables

Gaussian random vector $\begin{bmatrix} \tilde{a} \\ \tilde{b} \end{bmatrix}$ with mean $\begin{bmatrix} \mu_{\tilde{a}} \\ \mu_{\tilde{b}} \end{bmatrix}$ and covariance matrix

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

Minimum MSE estimator of $\tilde{b}$ given $\tilde{a} = a$?

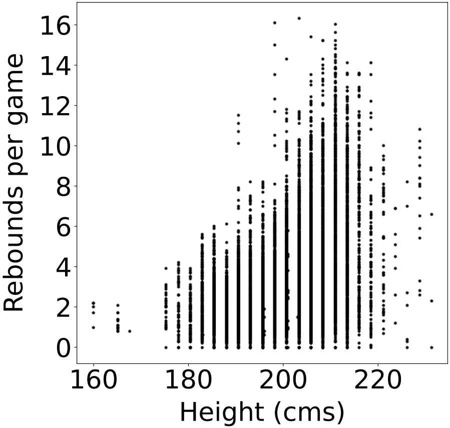
Conditional mean function

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

Equal to linear MMSE estimator

Simple linear regression from data

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



Goal: Estimate response y from features x

First idea

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 $\tilde{b}$

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

Second idea

Equivalent!

$$\begin{aligned}\ell_{\mathsf{OLS}}(x_i) &:= \beta_{\mathsf{OLS}}\, x_i + \alpha_{\mathsf{OLS}} \\ &= \sqrt{v(Y)}\rho_{X,Y}\left(\frac{x - m(X)}{\sqrt{v(X)}}\right) + m(Y)\end{aligned}$$

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

Height of NBA players

Data:

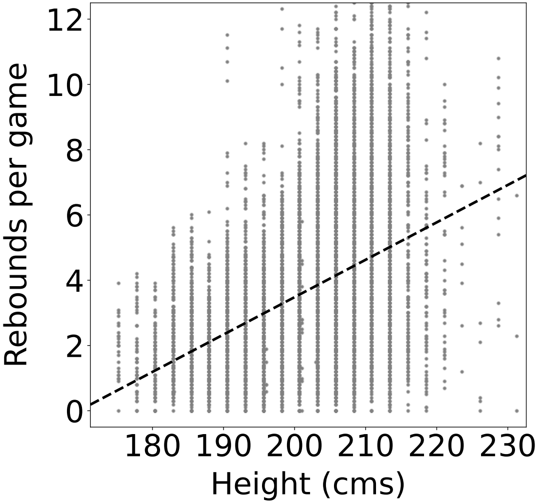
Height and offensive statistics of NBA players between 1996 and 2019

Goal:

Quantify linear dependence between rebounds/assists/points and height

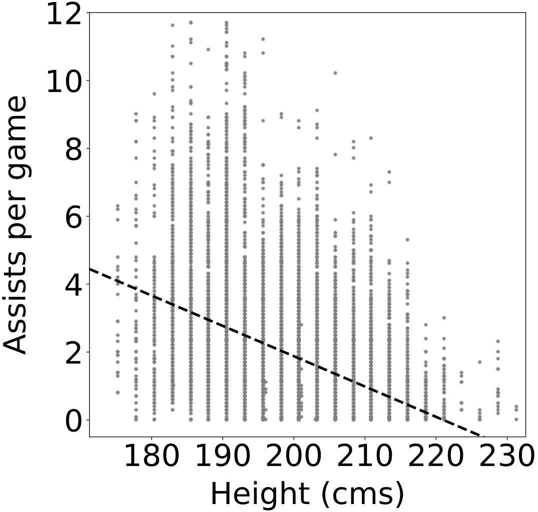
Height and rebounds

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



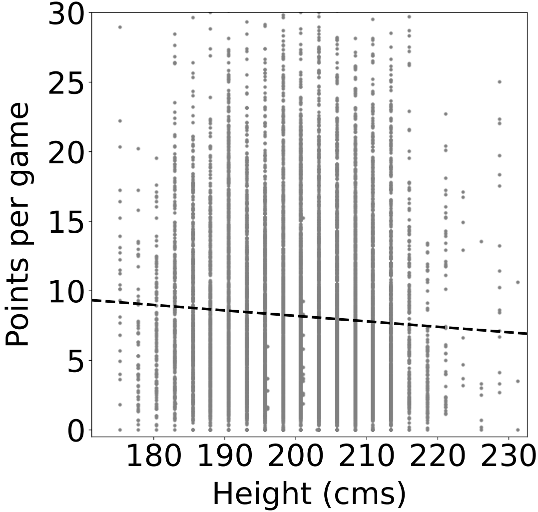
Height and assists

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



Height and points

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

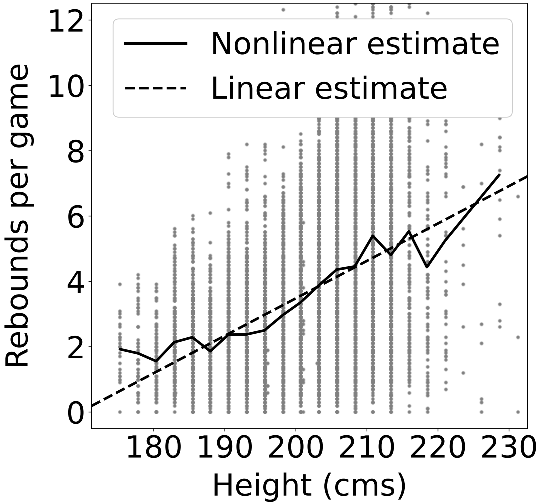


Best nonlinear estimate

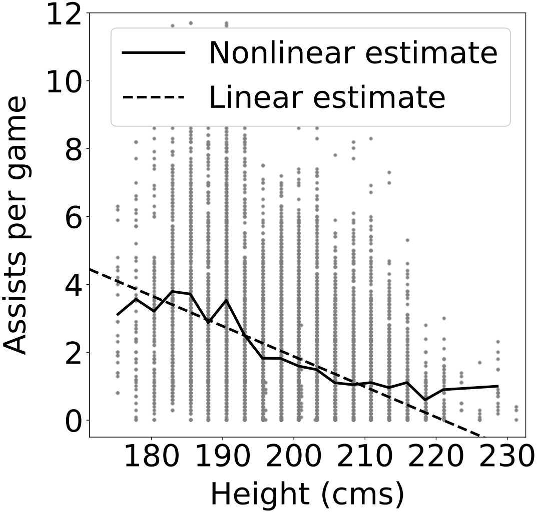
Best estimate of $\tilde{b}$ given $\tilde{a} = a$ in mean squared error?

Conditional mean function of $\tilde{b}$ given $\tilde{a} = a$

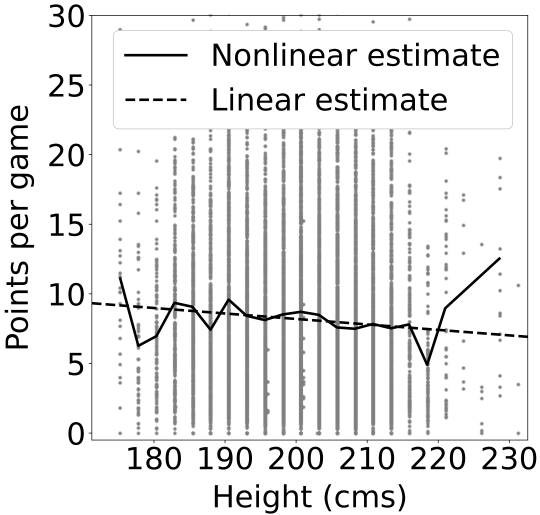
Height and rebounds



Height and assists



Height and points



What have we learned

Linear MMSE estimator for simple linear regression

How to compute it from data

Comparison to nonlinear MMSE estimator