Linear Regression: Ordinary Least Squares

Probability and Statistics for Data Science

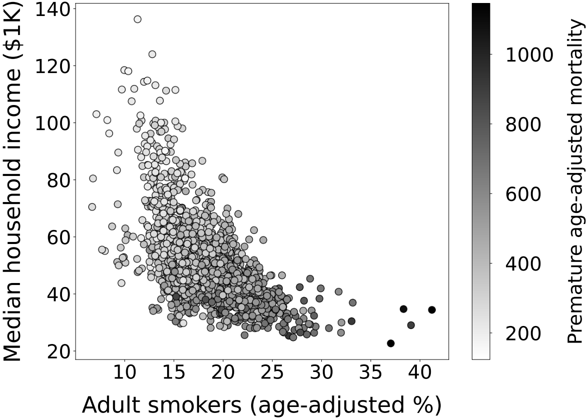
Carlos Fernandez-Granda

NYU Courant Institute of Mathematical Sciences
NYU Data Science

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 response from features



Response:

Premature mortality (deaths < age 75 per 10^4 people)

Features:

(1) Fraction of adult smokers (2) Median household income

Probabilistic formulation

Goal: Find function h, such that $h(x)$ approximates the response $\tilde{y}$ when the features $\tilde{x} = x$

How do we evaluate the estimator?

Mean squared error (MSE): $\mathrm{E}\left[(\tilde{y} - h(\tilde{x}))^2\right]$

MMSE estimator

The conditional mean is the minimum MSE estimator

$$\mu_{\tilde{y} \mid \tilde{x}}(\tilde{x}) = \arg\min_{h(\tilde{x})} \mathrm{E}\left[(\tilde{y} - h(\tilde{x}))^2\right]$$

Often impossible to compute due to curse of dimensionality

Linear regression

We approximate the response as an affine function of the features

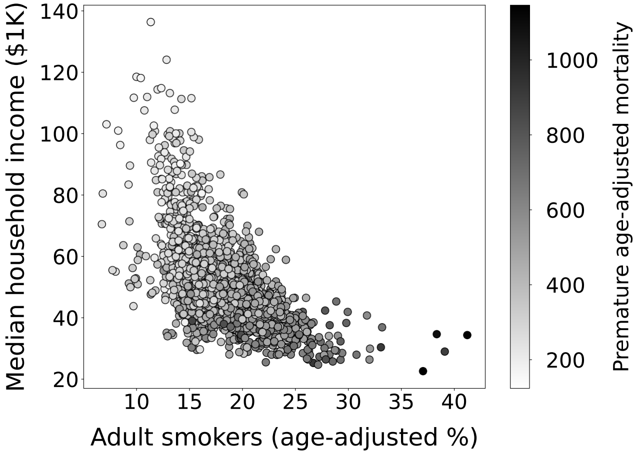
$$\tilde{y} \approx \ell(\tilde{x}) := \sum_{i=1}^{d} \beta[i]\tilde{x}[i] + \alpha$$
$$= \beta^T \tilde{x} + \alpha$$

Linear minimum MSE (MMSE) estimator

$$\ell_{\text{MMSE}}(\tilde{x}) := \Sigma_{\tilde{x}\tilde{y}}^T \Sigma_{\tilde{x}}^{-1} (\tilde{x} - \mu_{\tilde{x}}) + \mu_{\tilde{y}}$$

Data

$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$



Interpretation:

Samples from random vector $\tilde{x}$ and random variable $\tilde{y}$

Approximating the linear minimum MSE estimator

Response: $Y := \{y_1, y_2, \ldots, y_n\}$

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

$$\ell_{\text{MMSE}}(\tilde{x}) = \Sigma_{\tilde{x}\tilde{y}}^T \Sigma_{\tilde{x}}^{-1} (\tilde{x} - \mu_{\tilde{x}}) + \mu_{\tilde{y}}$$

$\mu_{\tilde{x}} \rightarrow$ sample mean $m(X)$

$\mu_{\tilde{y}} \rightarrow$ sample mean $m(Y)$

$\Sigma_{\tilde{x}}$? $\Sigma_{\tilde{x}\tilde{y}}$?

Sample covariance matrix of the features

jth feature: $X[j] := \{x_1[j], \ldots, x_n[j]\}$

$v(X[j])$: sample variance of $X[j]$

$c(X[j], X[k])$: sample covariance of $X[j]$ and $X[k]$

$$\Sigma_X := \begin{bmatrix} v(X[1]) & c(X[1], X[2]) & \cdots & c(X[1], X[d]) \\ c(X[1], X[2]) & v(X[2]) & \cdots & c(X[2], X[d]) \\ \vdots & \vdots & \ddots & \vdots \\ c(X[1], X[d]) & c(X[2], X[d]) & \cdots & v(X[d]) \end{bmatrix}$$

Sample cross-covariance

$$\Sigma_{XY} := \begin{bmatrix} c(X[1], Y) \\ c(X[2], Y) \\ \cdots \\ c(X[d], Y) \end{bmatrix}$$

Ordinary-least-squares (OLS) estimator

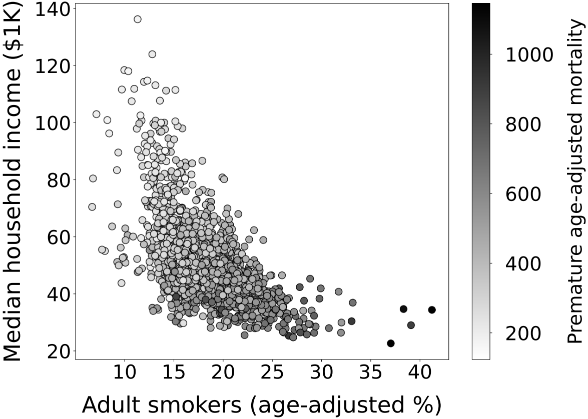
$$\ell_{\text{MMSE}}(\tilde{x}) = \Sigma_{\tilde{x}\tilde{y}}^T \Sigma_{\tilde{x}}^{-1} (\tilde{x} - \mu_{\tilde{x}}) + \mu_{\tilde{y}}$$

$$\begin{aligned} \ell_{\text{OLS}}(x_i) &= \Sigma_{XY}^T \Sigma_X^{-1} (x_i - m(X)) + m(Y) \\ &= \beta_{\text{OLS}}^T x_i + \alpha_{\text{OLS}} \end{aligned}$$

Alternative strategy:

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

Counties in the United States



Response:

Premature mortality (deaths < age 75 per 10^4 people)

Features:

(1) Fraction of adult smokers (2) Median household income

Counties in the United States

$$\beta_{\text{OLS}} = \Sigma_X^{-1}\Sigma_{XY} = \begin{bmatrix} 15.7 \\ -3.04 \end{bmatrix}$$

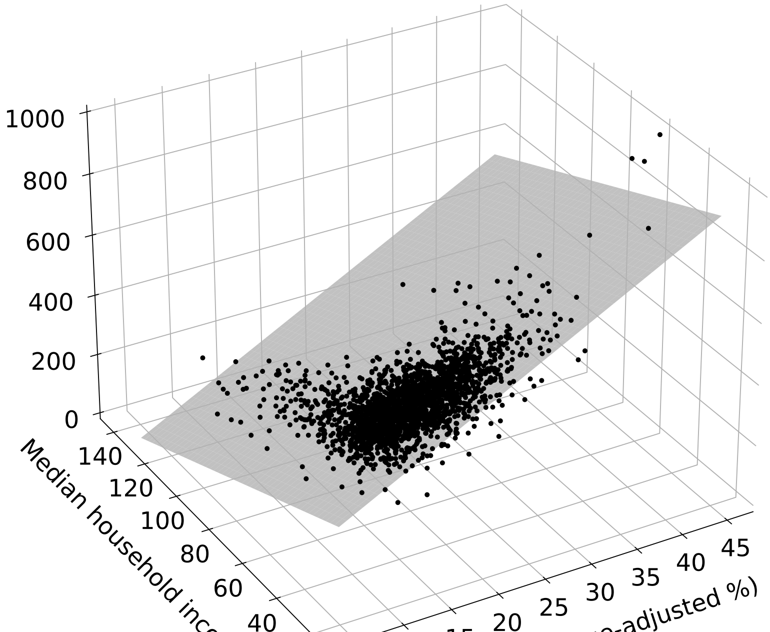
$$\alpha_{\text{OLS}} = m(Y) - \Sigma_{XY}^T \Sigma_X^{-1} m(X) = 281$$

$$\ell_{\text{OLS}}\left(x_{\text{tobacco}}, x_{\text{income}}\right) = 15.7\, x_{\text{tobacco}} - 3.04\, x_{\text{income}} + 281$$

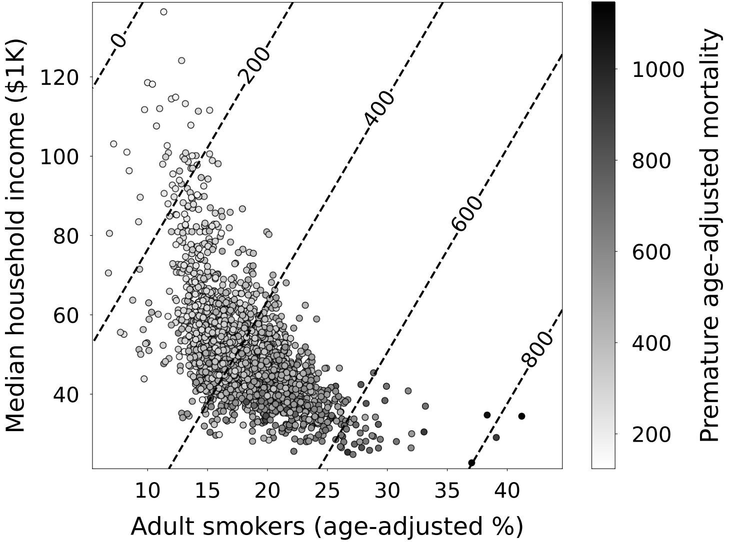
$$\Sigma_X = \begin{bmatrix} 13.6 & -30.6 \\ -30.6 & 190 \end{bmatrix} \qquad \Sigma_{XY} = \begin{bmatrix} 306 \\ -1057 \end{bmatrix}$$

$$m(X) = \begin{bmatrix} 18 \\ 50.9 \end{bmatrix} \qquad m(Y) = 408$$

$$15.7\,x_{\text{tobacco}} - 3.04\,x_{\text{income}} + 281$$



$$15.7\, x_{\text{tobacco}} - 3.04\, x_{\text{income}} + 281$$



Interpreting the coefficients

$$\beta_{\text{OLS}} = \begin{bmatrix} 15.7 \\ -3.04 \end{bmatrix}$$

Rate of change with respect to each feature assuming that remaining features are fixed

If median income is fixed:

+1% adult smokers $\implies$ +15.7 premature deaths per 10^4

If tobacco use is fixed:

+$1,000 in income $\implies$ -3.04 premature deaths per 10^4

What have we learned?

How to compute and interpret the ordinary-least-squares estimator