

Linear Regression: Explained Variance

Probability and Statistics for Data Science

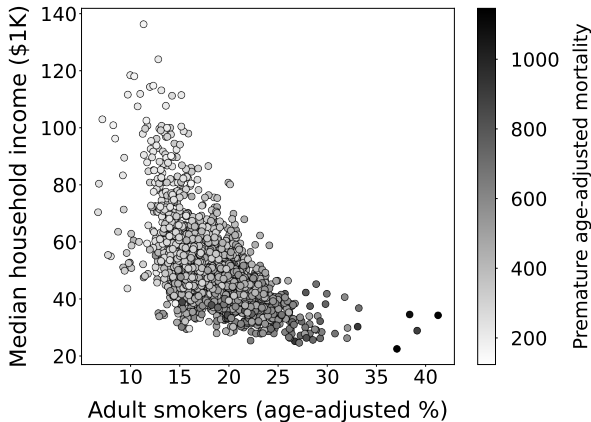
Carlos Fernandez-Granda



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 response from features



Response:

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

Features:

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

Linear regression

Linear minimum MSE estimator of response \tilde{y} given features \tilde{x}

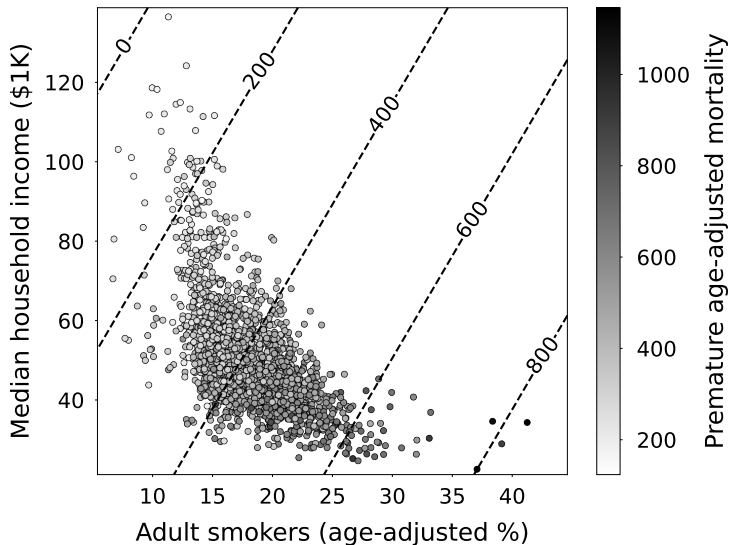
$$\ell_{\text{MMSE}}(\tilde{x}) = \Sigma_{\tilde{x}\tilde{y}}^T \Sigma_{\tilde{x}}^{-1} (\tilde{x} - \mu_{\tilde{x}}) + \mu_{\tilde{y}}$$

Ordinary-least-squares estimator from dataset

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

$$\ell_{\text{OLS}}(x_i) = \Sigma_{XY}^T \Sigma_X^{-1} (x_i - m(X)) + m(Y)$$

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



Goal: Evaluate the estimator

One feature \tilde{a}

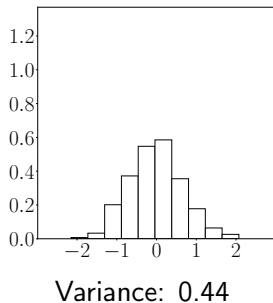
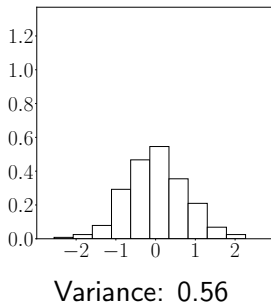
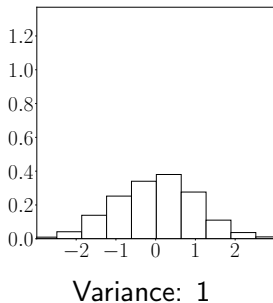
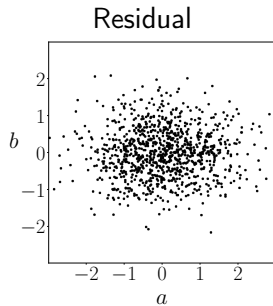
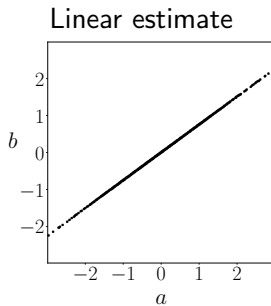
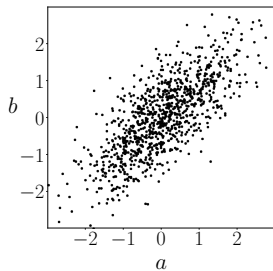
$$\tilde{y} = \underbrace{\ell_{\text{MMSE}}(\tilde{a})}_{\text{Linear MMSE estimate}} + \underbrace{\tilde{y} - \ell_{\text{MMSE}}(\tilde{a})}_{\text{Residual}}$$

Residual uncorrelated with $\ell_{\text{MMSE}}(\tilde{a})$

$$\text{Var} [\tilde{y}] = \text{Var} [\ell_{\text{MMSE}}(\tilde{a})] + \text{Var} [\tilde{y} - \ell_{\text{MMSE}}(\tilde{a})]$$

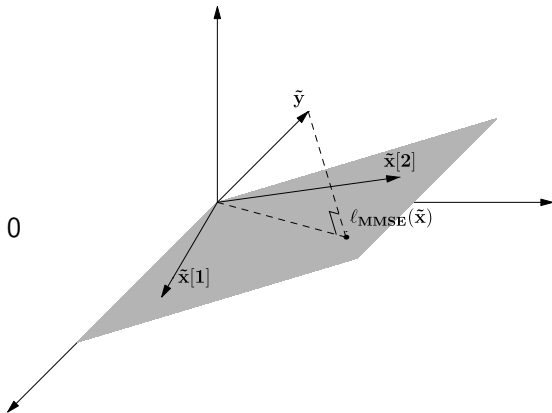
$$\text{Var} [\ell_{\text{MMSE}}(\tilde{a})] = \rho_{\tilde{a}, \tilde{y}}^2 \text{Var} [\tilde{y}] = R^2 \text{Var} [\tilde{y}]$$

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



Geometric intuition

$$\langle \ell_{\text{MMSE}}(\tilde{\mathbf{x}}), \tilde{\mathbf{y}} - \ell_{\text{MMSE}}(\tilde{\mathbf{x}}) \rangle = 0$$



Uncorrelated residual

$$\begin{aligned} & \text{Cov} [\ell_{\text{MMSE}}(\tilde{x}), \tilde{y} - \ell_{\text{MMSE}}(\tilde{x})] \\ &= \text{E} [\text{ct}(\ell_{\text{MMSE}}(\tilde{x})) \text{ct}(\tilde{y} - \ell_{\text{MMSE}}(\tilde{x}))] \\ &= \text{E} \left[\beta_{\text{MMSE}}^T \text{ct}(\tilde{x}) \left(\text{ct}(\tilde{y}) - \text{ct}(\tilde{x})^T \beta_{\text{MMSE}} \right) \right] \\ &= \beta_{\text{MMSE}}^T \text{E} [\text{ct}(\tilde{x}) \text{ct}(\tilde{y})] - \beta_{\text{MMSE}}^T \text{E} \left[\text{ct}(\tilde{x}) \text{ct}(\tilde{x})^T \right] \beta_{\text{MMSE}} \\ &= \Sigma_{\tilde{x}\tilde{y}}^T \Sigma_{\tilde{x}}^{-1} \Sigma_{\tilde{x}\tilde{y}} - \Sigma_{\tilde{x}\tilde{y}}^T \Sigma_{\tilde{x}}^{-1} \Sigma_{\tilde{x}} \Sigma_{\tilde{x}}^{-1} \Sigma_{\tilde{x}\tilde{y}} \\ &= \Sigma_{\tilde{x}\tilde{y}}^T \Sigma_{\tilde{x}}^{-1} \Sigma_{\tilde{x}\tilde{y}} - \Sigma_{\tilde{x}\tilde{y}}^T \Sigma_{\tilde{x}}^{-1} \Sigma_{\tilde{x}\tilde{y}} \\ &= 0 \end{aligned}$$

$$\begin{aligned} \text{ct}(\ell_{\text{MMSE}}(\tilde{x})) &= \ell_{\text{MMSE}}(\tilde{x}) - \text{E}[\ell_{\text{MMSE}}(\tilde{x})] \\ &= \beta_{\text{MMSE}}^T \tilde{x} + \alpha_{\text{MMSE}} - \beta_{\text{MMSE}}^T \mu_{\tilde{x}} - \alpha_{\text{MMSE}} \\ &= \beta_{\text{MMSE}}^T \text{ct}(\tilde{x}) \end{aligned}$$

Decomposition of variance

$$\begin{aligned}\text{Var} [\tilde{y}] &= \text{Var} [\ell_{\text{MMSE}}(\tilde{x}) + \tilde{y} - \ell_{\text{MMSE}}(\tilde{x})] \\ &= \text{Var} [\ell_{\text{MMSE}}(\tilde{x})] + \text{Var} [\tilde{y} - \ell_{\text{MMSE}}(\tilde{x})] \\ &= \text{Var} [\ell_{\text{MMSE}}(\tilde{x})] + \text{MSE}\end{aligned}$$

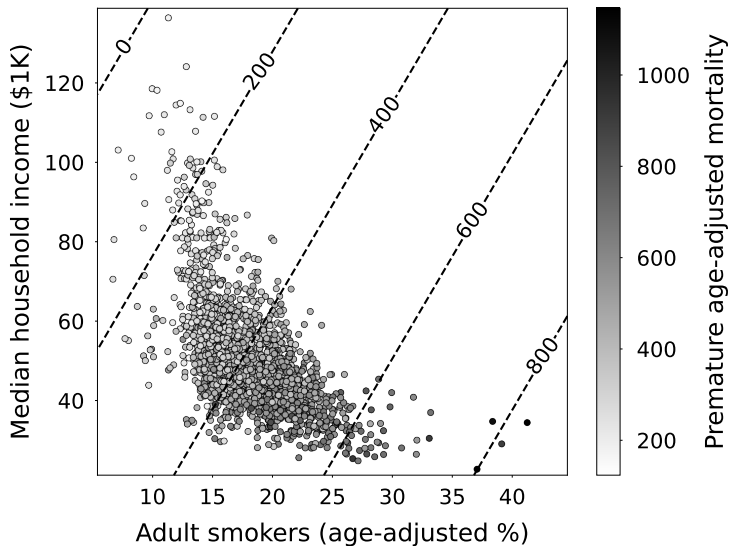
$$\begin{aligned}\text{E} [\tilde{y} - \ell_{\text{MMSE}}(\tilde{x})] &= \text{E} [\tilde{y} - \beta_{\text{MMSE}}^T \tilde{x} - \alpha_{\text{MMSE}}] \\ &= \text{E} [\tilde{y}] - \beta_{\text{MMSE}}^T \text{E} [\tilde{x}] - \alpha_{\text{MMSE}} \\ &= 0\end{aligned}$$

Coefficient of determination

$$\text{Var} [\tilde{y}] = \text{Var} [\ell_{\text{MMSE}}(\tilde{x})] + \text{MSE}$$

$$\begin{aligned} R^2 &:= \frac{\text{Var} [\ell_{\text{MMSE}}(\tilde{x})]}{\text{Var} [\tilde{y}]} \\ &= \frac{\text{Var} [\tilde{y}] - \text{MSE}}{\text{Var} [\tilde{y}]} \\ &= 1 - \frac{\text{MSE}}{\text{Var} [\tilde{y}]} \end{aligned}$$

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



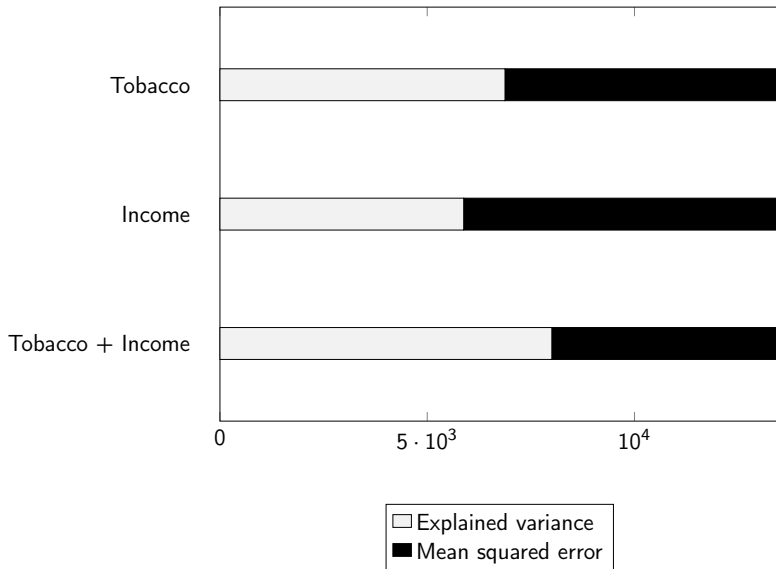
Counties in the United States

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

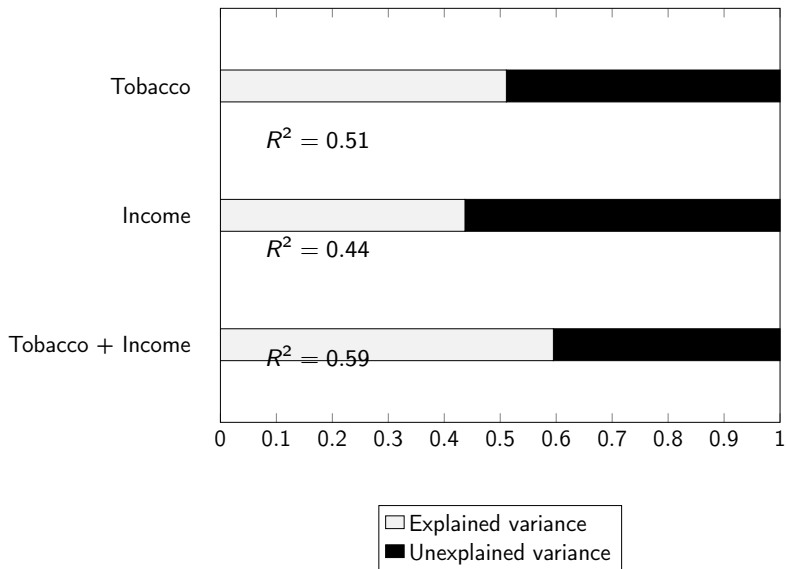
$$\ell_{\text{OLS}}(x_{\text{tobacco}}) = 22.52 x_{\text{tobacco}} + 2$$

$$\ell_{\text{OLS}}(x_{\text{income}}) = -5.57 x_{\text{income}} + 692$$

Variance of the response: $1.35 \cdot 10^4$



Explained variance



Sample correlation coefficient between tobacco and income: -0.6

What have we learned?

How to evaluate a linear-regression model using its explained variance

Connection between explained variance and mean squared error