Linear Regression: Explained Variance

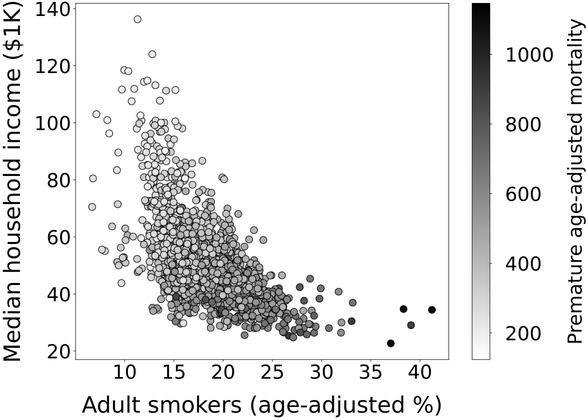
Probability and Statistics for Data Science

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

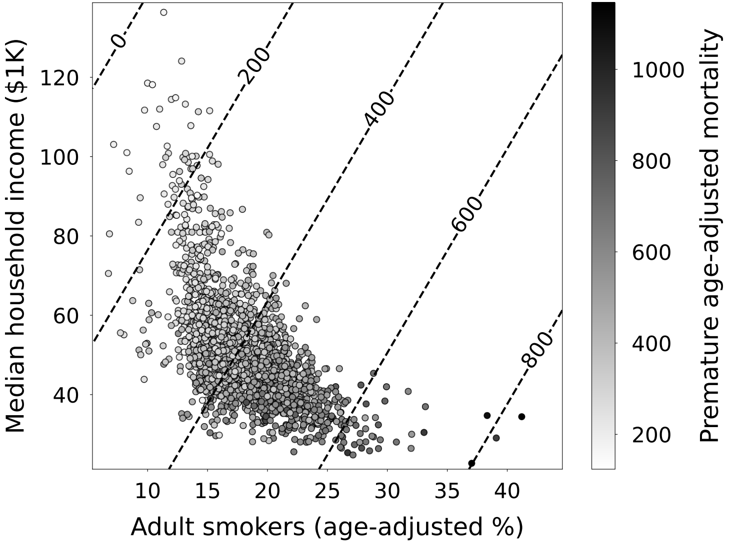
Linear minimum MSE estimator of response $\tilde{y}$ given features $\tilde{x}$

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

Ordinary-least-squares estimator from dataset
(x_1, y_1), (x_2, y_2), ..., (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 $\tilde{a}$

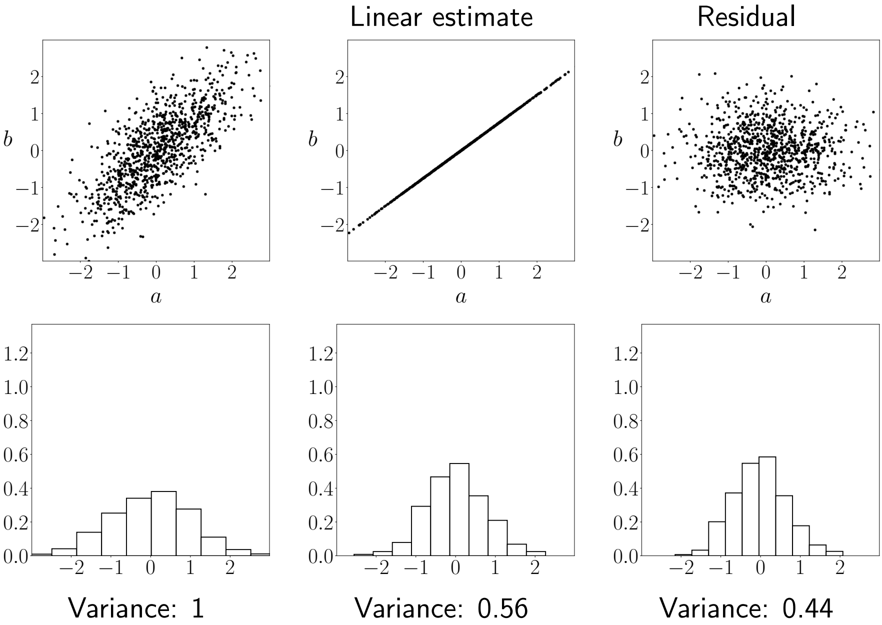
$$\tilde{y} = \underbrace{\ell_{\text{MMSE}}(\tilde{a})}_{\text{Linear MMSE estimate}} + \underbrace{\tilde{y} - \ell_{\text{MMSE}}(\tilde{a})}_{\text{Residual}}$$

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

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

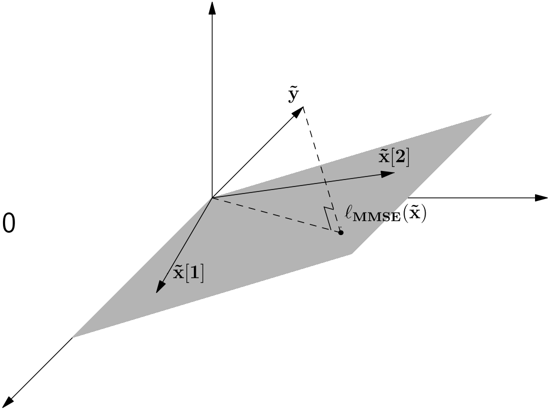
$$\mathrm{Var}\left[\ell_{\text{MMSE}}(\tilde{a})\right] = \rho^2_{\tilde{a},\tilde{y}} \mathrm{Var}\left[\tilde{y}\right] = R^2 \mathrm{Var}\left[\tilde{y}\right]$$

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



Geometric intuition

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



Uncorrelated residual

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

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

Decomposition of variance

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

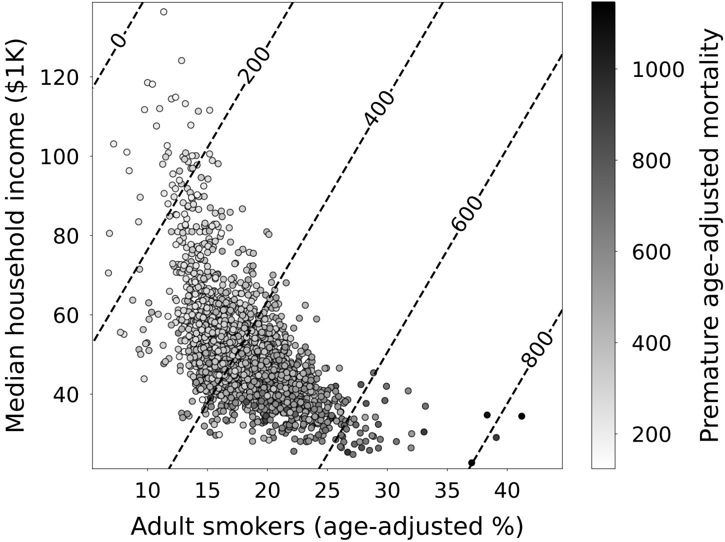
$$
\begin{aligned}
\mathrm{E}\left[\tilde{y} - \ell_{\text{MMSE}}\left(\tilde{x}\right)\right] &= \mathrm{E}\left[\tilde{y} - \beta_{\text{MMSE}}^{T}\tilde{x} - \alpha_{\text{MMSE}}\right] \\
&= \mathrm{E}\left[\tilde{y}\right] - \beta_{\text{MMSE}}^{T}\mathrm{E}\left[\tilde{x}\right] - \alpha_{\text{MMSE}}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

Coefficient of determination

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

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

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



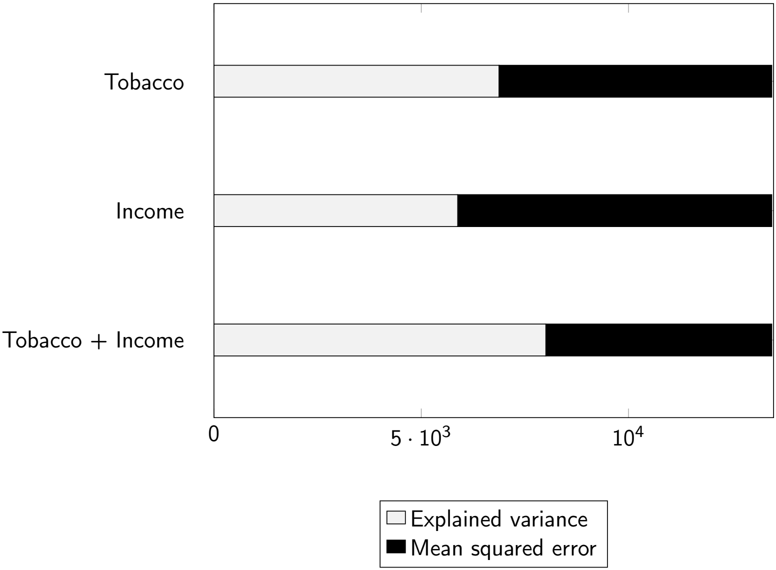
Counties in the United States

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

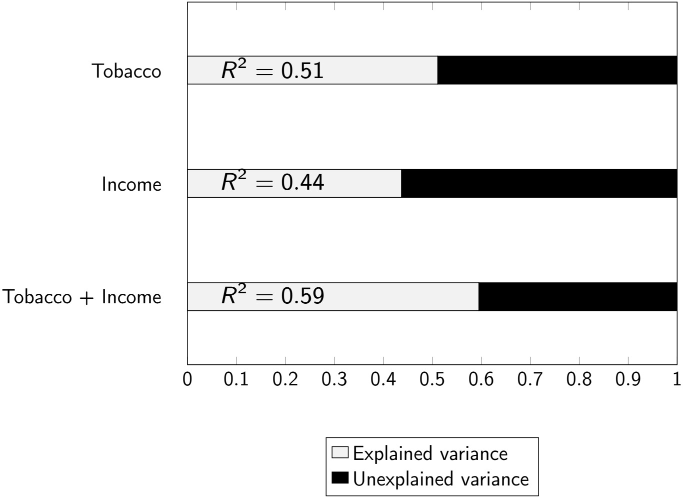
$$\ell_{\text{OLS}}\left(x_{\text{tobacco}}\right) = 22.52\, x_{\text{tobacco}} + 2$$

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

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