Chapter 08

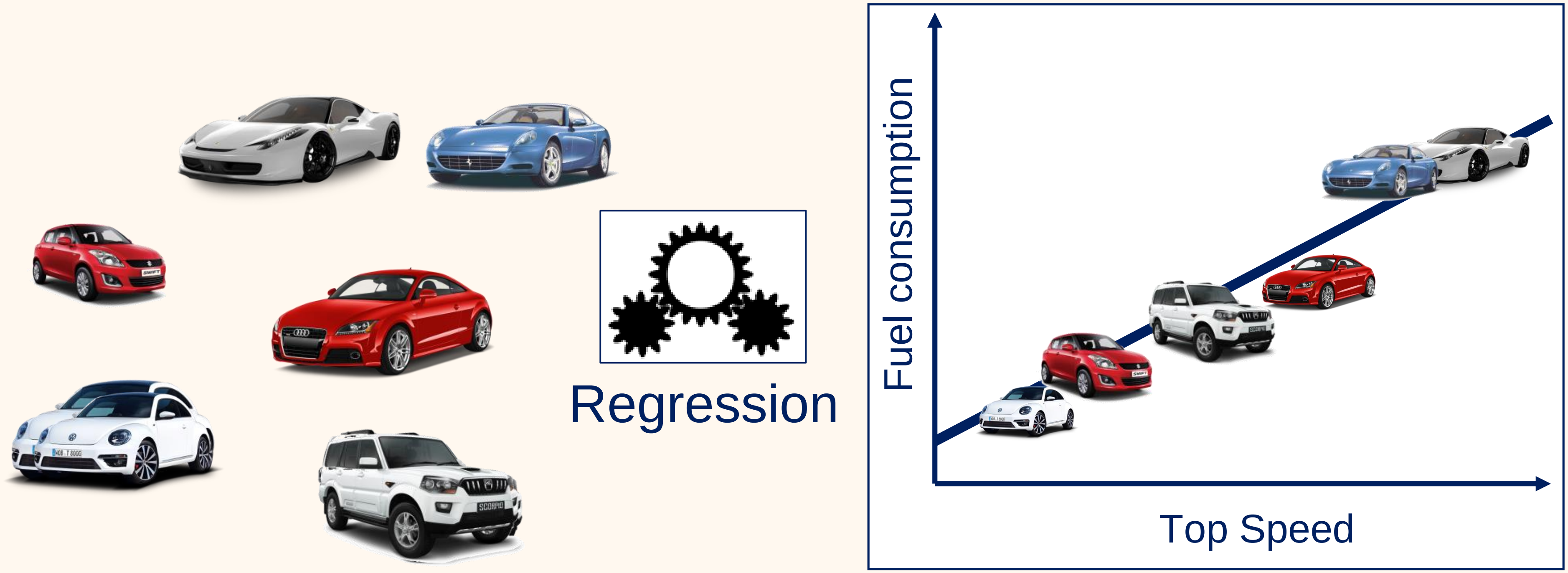
Regression

Dr. Steffen Herbold
herbold@cs.uni-goettingen.de

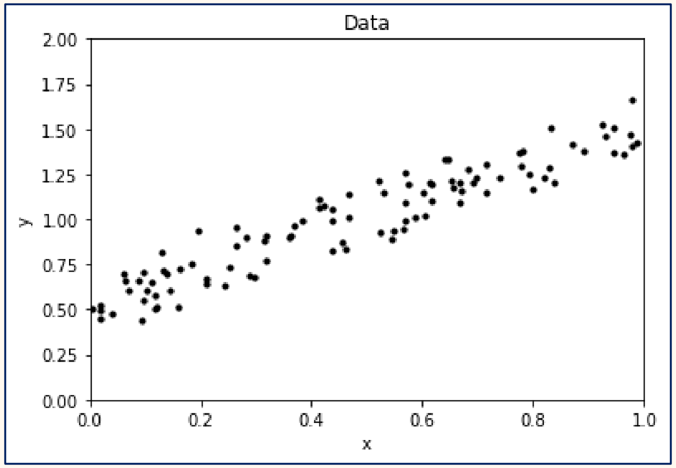
Outline

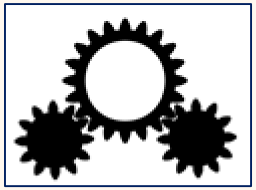
- Overview
- Linear Regression Models
- Comparison of Regression Models
- Summary

Example of Regression

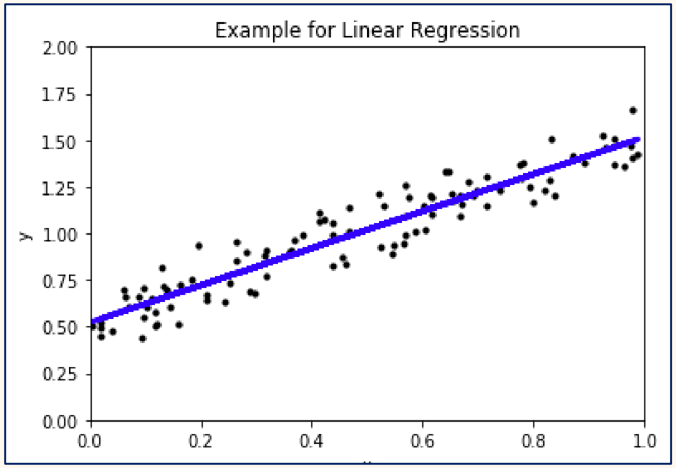


The General Problem



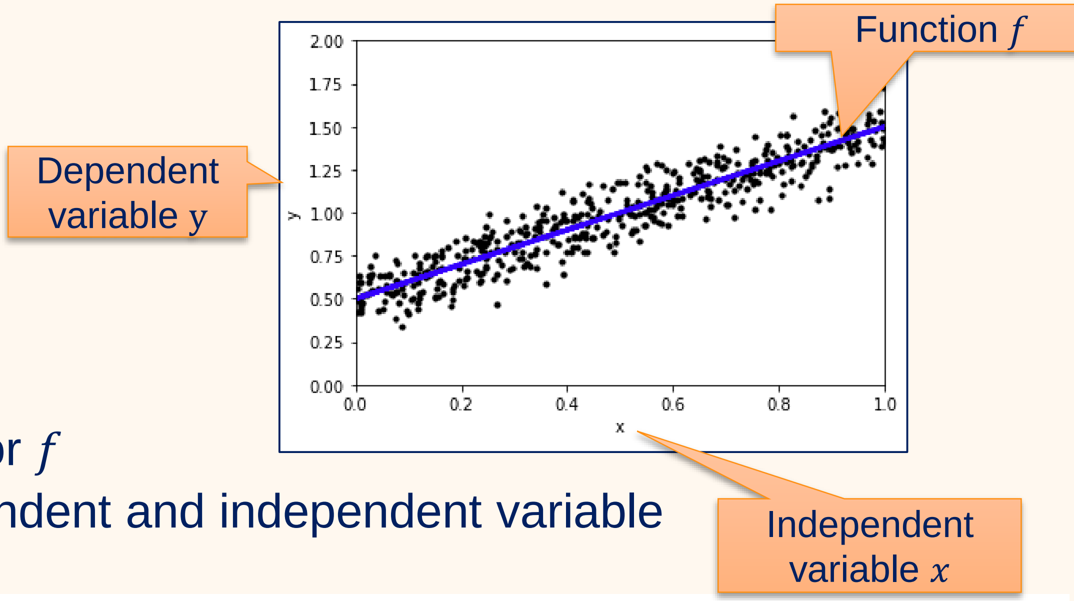


Regression



The Formal Problem

- Object space
  - $O = \{object_1, object_2, \dots\}$
  - Often infinite
- Representations of the objects in a (real valued) feature space
  - $\mathcal{F} = \{\phi(o), o \in O\} = \{(x_1, \dots, x_m) \in \mathbb{R}^m\} = X$
  - „Independent“ variables
- Dependent variable
  - $f^*(o) = y \in \mathbb{R}$
- A regression function
  - $f: \mathbb{R}^m \to \mathbb{R}$
- Regression
  - Finding an approximation for f
  - Relationship between dependent and independent variable



Quality of Regressions

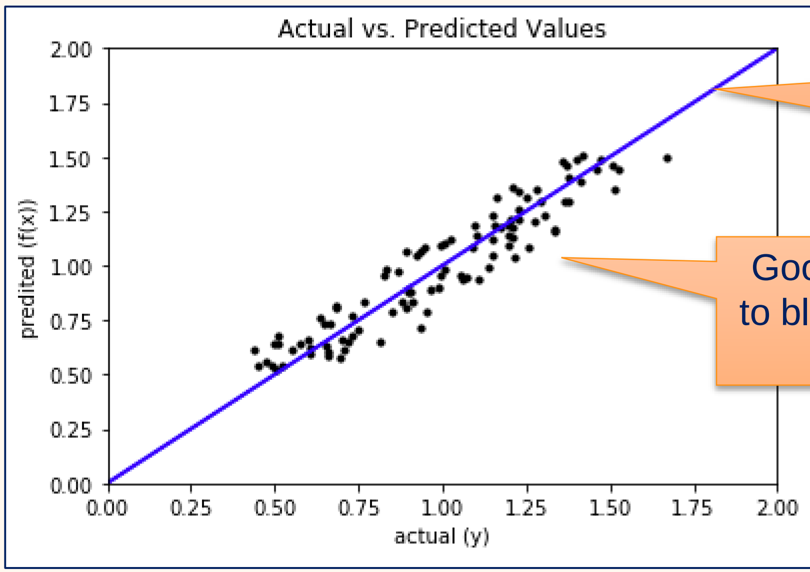
How do you evaluate
$f^*(o) \approx f(\phi(o))$

- Goal: Approximation of the dependent variable
  - $f^*(o) \approx h(\phi(o))$

→ Use Test Data
  - Structure is the same as training data
  - Apply approximated regression function

$\phi(o)$					$f^*(o)$	$f(\phi(o))$
Top Speed	Engine Size	Horse Power	Weigth	Year	value	prediction
250	1.4	130	1254	2003	**7.8**	**7.5**
280	1.8	185	1430	2010	**6.3**	**6.9**
…	…	…	…	…	…	

Visual Comparison



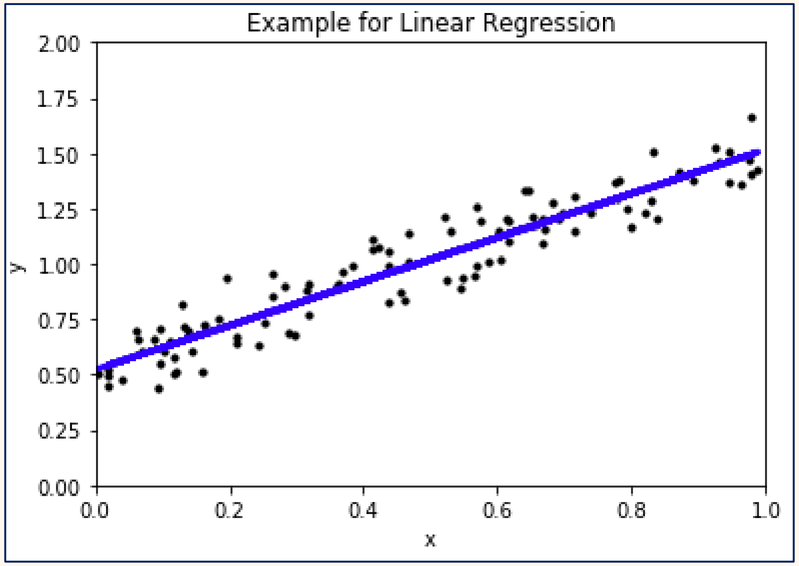
Perfect Prediction
→ Deviation from blue line as visual indicator for prediction quality

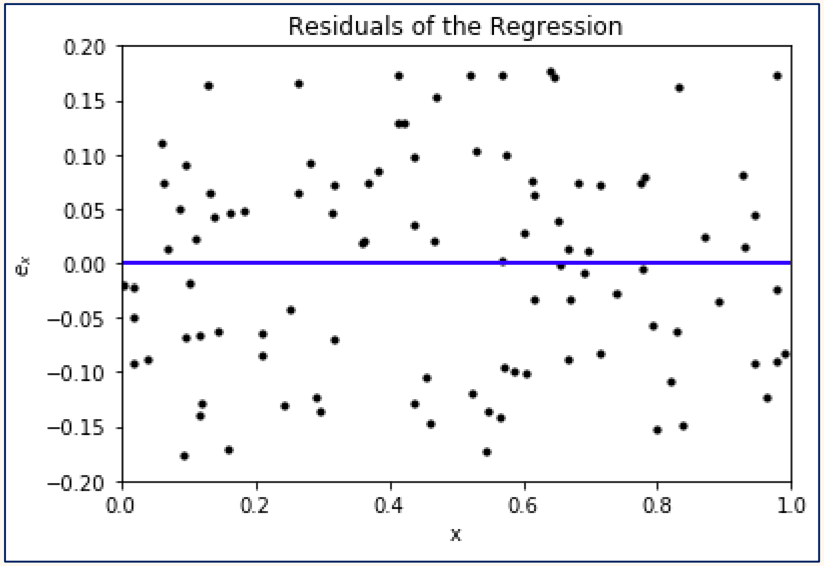
Good prediction. Data close to blue line, regular pattern of deviations

Allows insights into where predictions are good/bad

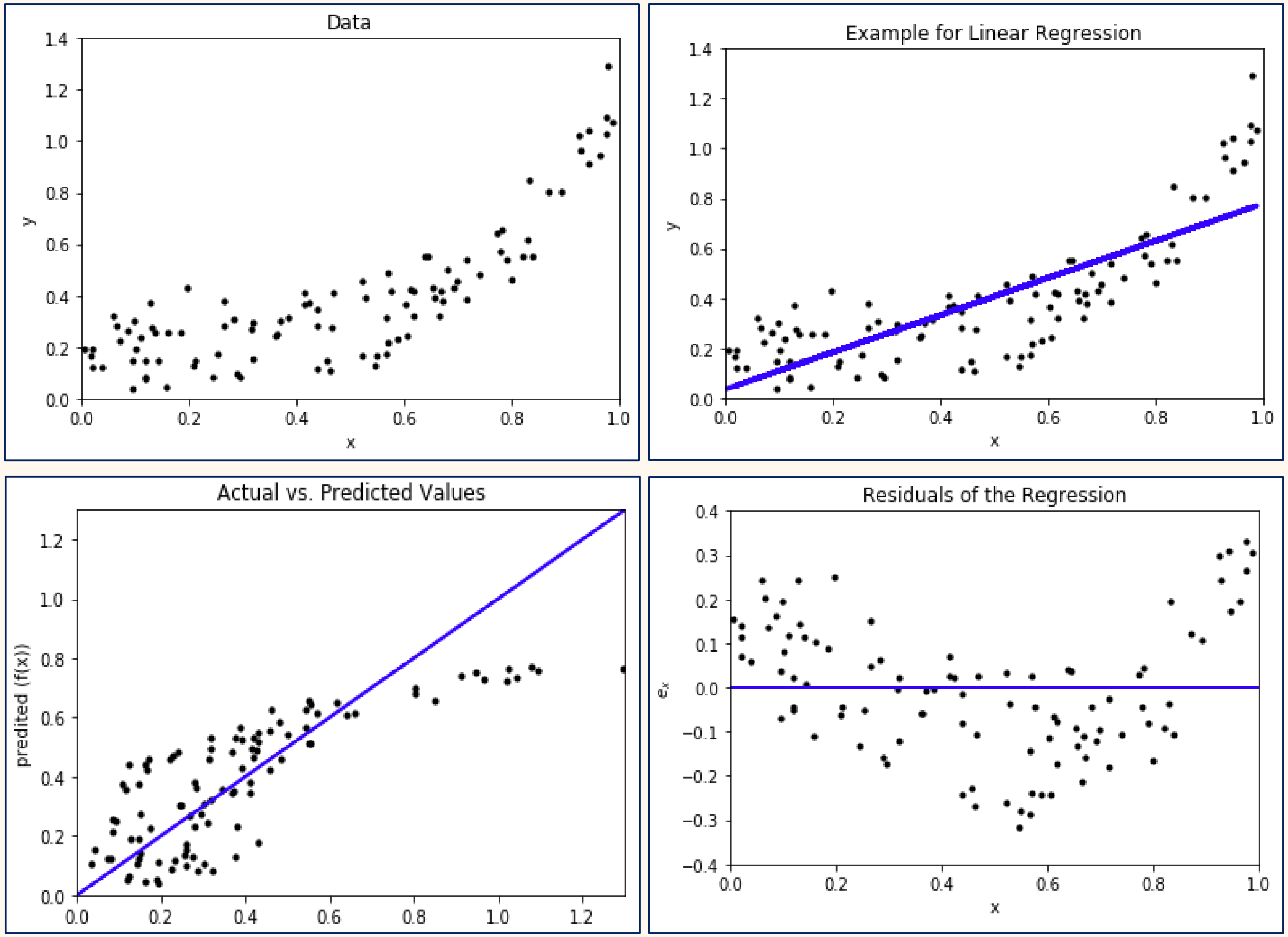
Residuals

- Differences between predictions and actual values
  - $e_x = y - f(x)$





Visual Comparison of a Bad Fit



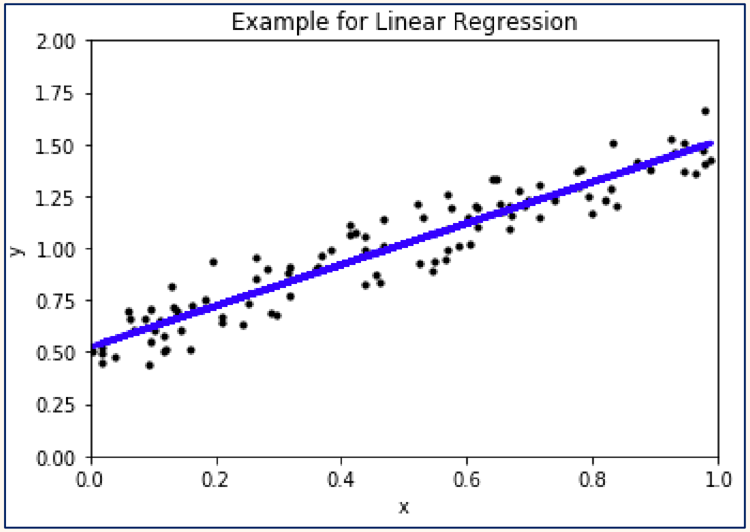
Measures for Regression Quality

- Mean Absolute Error (MAE)
  - $MAE = \frac{1}{|X|} \sum_{x \in X} |e_x|$
- Mean Squared Error (MSE)
  - $MSE = \frac{1}{|X|} \sum_{x \in X} (e_x)^2$
- R squared (R^2)
  - Fraction of the variance that is explained by the regression
  - $R^2 = 1 - \frac{\sum_{x \in X} (y - f(x))^2}{\sum_{x \in X} \left(y - mean(y)\right)^2}$
- Adjusted R squared ($\bar{R}^2$)
  - Takes number of features into account
  - $\bar{R}^2 = 1 - (1 - R^2) \frac{|X| - 1}{|X| - m - 1}$

Outline

- Overview
- **Linear Regression Models**
- Comparison of Regression Models
- Summary

Linear Regression



- Regression as a linear function
  - $y = b_0 + b_1 x_1 + \cdots b_m x_m$
  - b_0 is the interception with the axis
  - $b_1, \ldots, b_m$ are the linear coefficients

- Calculated with Ordinary Least Squares
  - Optimizes MSE!
  - $\min \left|\left| b_0 + Xb - y \right|\right|_2^2$ Square of euclidean distance
  - $X = \begin{pmatrix} x_{1,1} & \cdots & x_{1,m} \\ \vdots & \ddots & \vdots \\ x_{n,1} & \cdots & x_{n,m} \end{pmatrix}$ n is the number of instances in the training data
  - $b = (b_1, \ldots, b_m)$
  - $y = (y_1, \ldots, y_n)$

Ridge Regression

- Still a linear function

- OLS allows multiple solutions for $n > m$
- Ridge regression penalizes solutions with large coefficients

- Calculated with *Tikhonov regularization*
  - $\min ||b_0 + Xb - y||_2^2 + ||\Gamma b||_2^2$
  - We use $\Gamma = \alpha I$

Regularization Term

Identity matrix

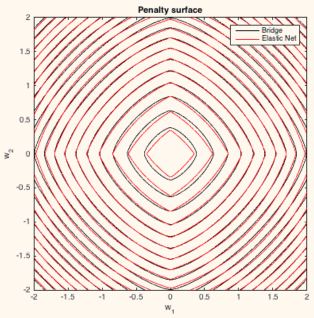
- Use α to regulate regularization strength
  - $\min ||b_0 + Xb - y||_2^2 + \alpha ||b||_2^2$

Lasso Regression

- Still a linear function
- Penalizing large coefficient does not remove redundencies
  - Extreme example: identical features that predict perfectly
    - $\mathrm{y} = x_1 = x_2$,
  - Ridge
    - $b_1 = b_2 = 0.5$
  - One coefficient zero would be better
    - $b_1 = 1, b_2 = 0$

- Lasso: Ridge with Manhatten norm
  - $\min \left|\left|b_0 + Xb - y\right|\right|_2^2 + \alpha \left|\left|b\right|\right|_1$
- Increases the likelihood of coefficients being exactly zero
  - Selects relevant features

Elastic Net Regression



- Still a linear function

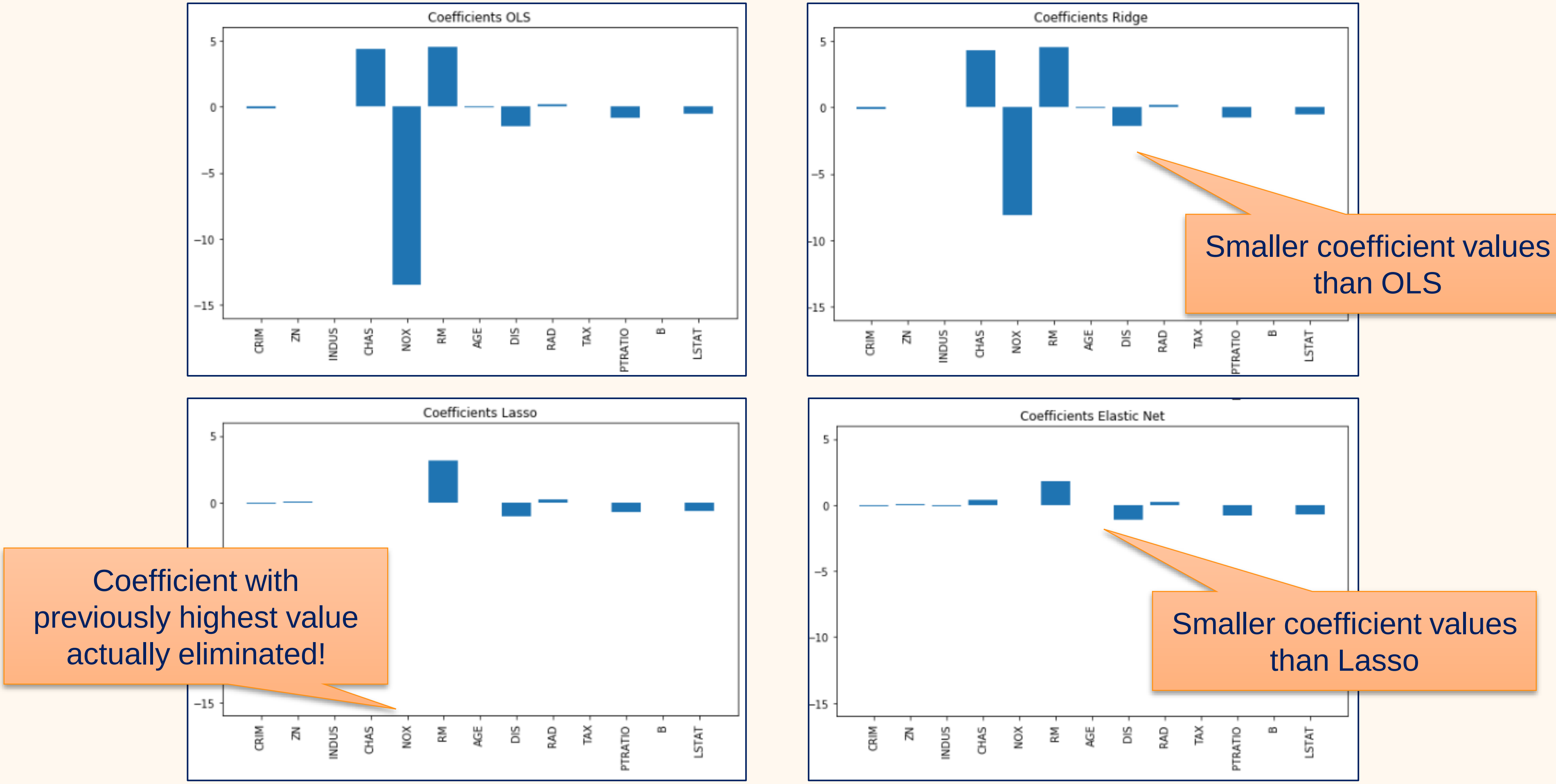
- Lasso tends to select one of multiple correlated features at random
  - Potential loss of information
- Elastic Net combines Ridge and Lasso
  - Keeps only relevant correlated features and minimizes coefficients

- Use ratio ρ between alphas for assigning more weight to Ridge/Lasso
  - $\min||b_0 + Xb - y||_2^2 + \rho \cdot \alpha||b||_1 + \frac{(1-\rho)}{2}\alpha||b||_2^2$

Outline

- Overview
- Linear Regression Models
- **Comparison of Regression Models**
- Summary

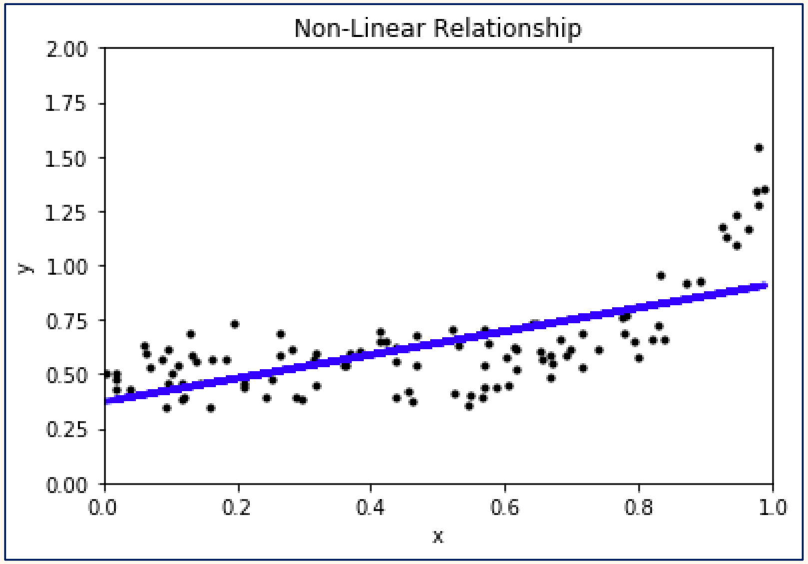
Comparison of Regression Models



All models trained with the same data and almost same performance

Non-linear Regression

- Many relationships are not linear



- Polynomial Regression
- Support Vector Regression
- Neural Networks

Outline

- Overview
- Linear Regression Models
- Comparison of Regression Models
- **Summary**

Summary

- Regression finds relationships between independent and dependent variables

- Linear regression as simple model often effective

- Regularization can improve solutions
  - Lasso, Ridge, Elastic, …

- Many non-linear approaches
  - Require care with the application
  - Overfitting can be very easy