Linear Regression

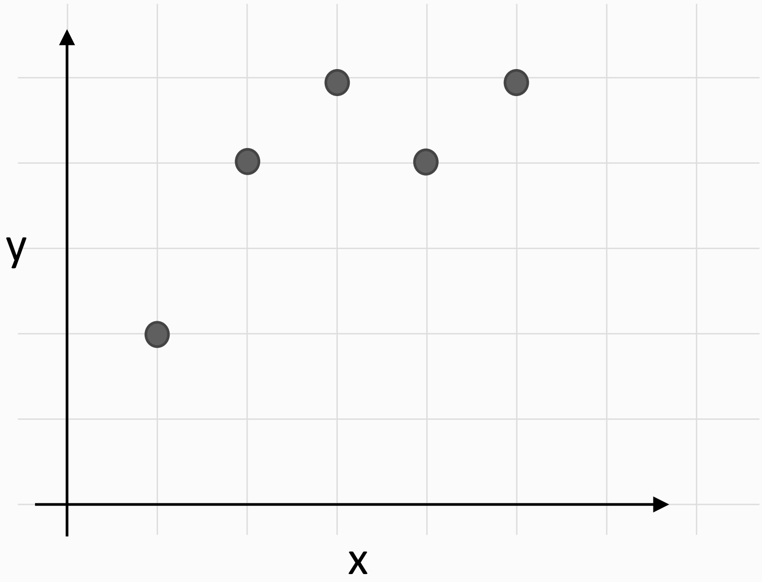
Areej Alasiry

Lecture Outline

- Linear Regression
- Errors in linear regression
- Finding Optimal Fit
- How to improve the regression models

Linear Regression

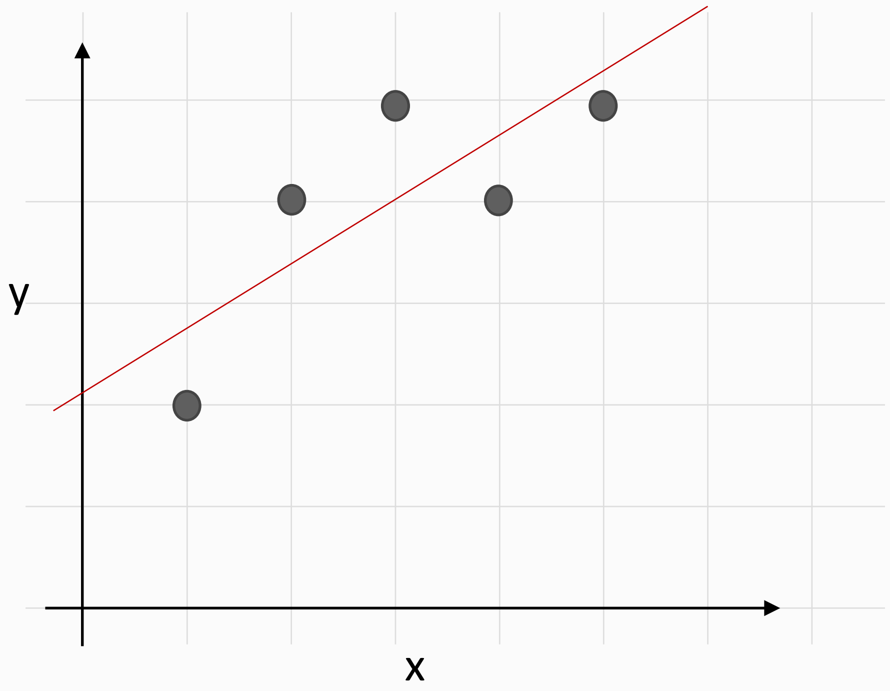
General Concept



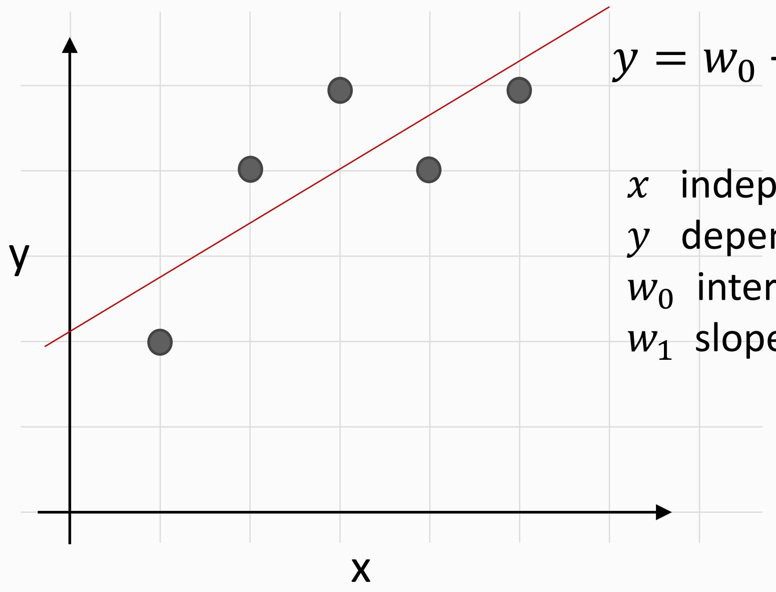
$$y = w_0 + w_1 x$$

x independent variable
y dependent variable
w_0 intercept
w_1 slope

Error in Linear Regression



Finding the Optimal Fit



$y = w_0 + w_1 x$

x independent variable
y dependent variable
w_0 intercept
w_1 slope

x	y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})^2$	$(x-\bar{x})\times(y-\bar{y})$
1	2	-2	-2	4	4
2	4	-1	0	1	0
3	5	0	1	0	0
4	4	1	0	1	0
5	5	2	1	4	2

$$b_1 = \frac{\sum(x-\overline{x})\,(y-\overline{y})}{\sum(x-\overline{x})^2} = \frac{6}{10} = 0.6$$

$$y = w_0 + w_1 x$$
$$4 = w_0 + (0.6 * 3)$$
$$4 = w_0 + (1.8)$$
$$b_0 = 4 - 1.8 = 2.2$$

$$\boxed{y = 2.2 + 1.8\,x}$$

Ref: *“https://bit.ly/2wwqaUW*

Finding the Optimal Fit

- Linear regression find the line y=f(x) that minimizes the sum of squared errors over all the training dataset.

$$\sum_{i=1}^{n}\left(y_i - f(x_i)\right), where\ f(x_i) = w_0 + \sum_{i=1}^{m-1} w_i x_i$$

$$w = (A^T\ A)^{-1} A^T b \ \ , \text{where b vector of target values}$$

Nice example: https://www.youtube.com/watch?v=Qa_FI92_qo8

Better Regression Models

Removing Outlier

From: The Data Science Design MANUAL. Steven S. Skiena, ISBN: 978-3-319-55444-0 ©2017.[Chapter 9]

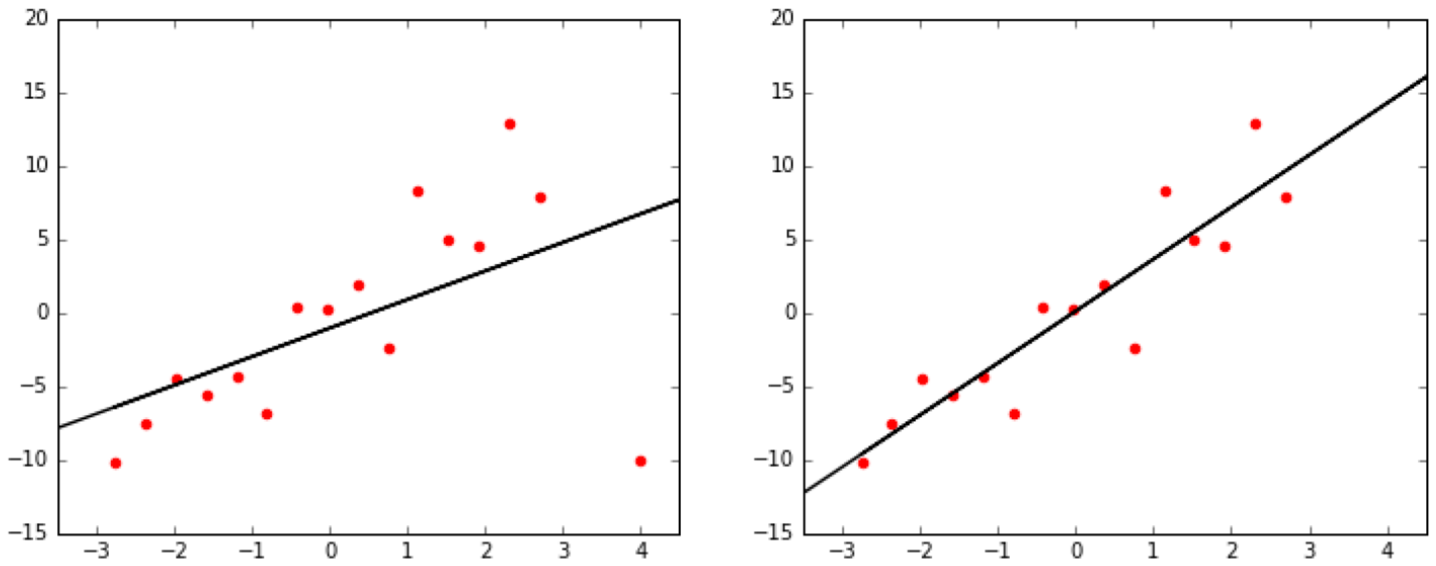


Figure 9.4: Removing outlier points (left) can result in much more meaningful fits (right).

Fitting non-linear functions

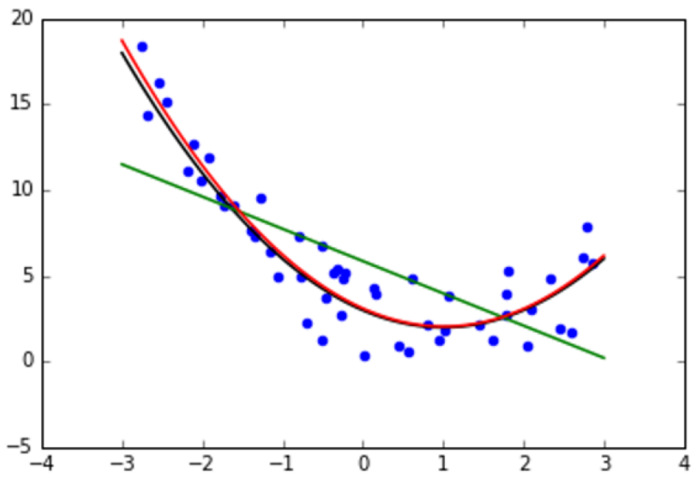


Figure 9.5: Higher-order models (red) can lead to better fits than linear models (green).

From: The Data Science Design MANUAL. Steven S. Skiena, ISBN: 978-3-319-55444-0 ©2017.[Chapter 9]

Feature and Target Scaling

- Example:
  - $GDP \ = \ \$10{,}000\, x_1 + \$10{,}000{,}000{,}000{,}000\, x_2$
  - $x_1 = Population\ Size$
  - $x_2 = Literacy\ Rate$
- Problems:
  - Unreadable coefficients
  - Numerical Optimization
  - Inappropriate formulations

Feature and Target Scaling

- Feature Scaling: Z-Scores

$$Z(x) = \frac{(x - \mu)}{\sigma}$$

- Sublinear Feature Scaling

$$\log(x) \quad and \, \sqrt{x}$$

- Sublinear Target Scaling

$$\log(y)$$

Dealing with Highly Correlated Variables

- In order to build a highly-predictive model, where we have a feature that is highly correlated with the target.

- However, there is the issue when the features are highly correlated with each other.

How to Build a Linear Regression Model

The steps:

- Split the dataset into training and test dataset.
- Use the training the dataset to fit the model
- Performs the prediction over the test dataset
- Evaluate the performance

Evaluating Score-based Model

- R-Squared

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(p - o)^2}{\sum_{i=1}^{n}(\mu - o)^2}$$

References

- *The Data Science Design MANUAL. Steven S. Skiena, ISBN: 978-3-319-55444-0 ©2017.[Chapter 9]*