German Research Center for Artificial Intelligence

Machine Learning (DLMDSML01)

Nghia Duong-Trung[1]
[1] **German Research Center for Artificial Intelligence (DFKI GmbH)**
Internal Use @ IU International University of Applied Science, Berlin Campus

SPONSORED BY THE

Federal Ministry
of Education
and Research

Section 3: Regression

German Research Center for Artificial Intelligence

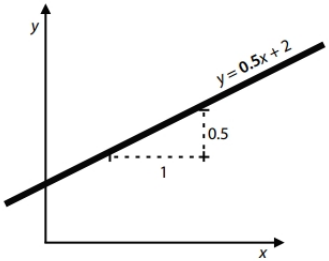
- Textbook: check it in MyCampus
- Basic concepts
- Linear & Nonlinear & Logistic & Quantile regression
- Regularization
- Transfer task

Basic concepts

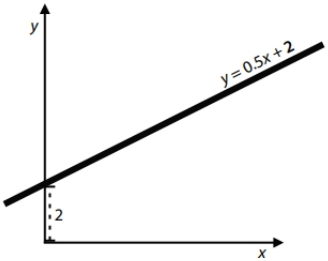
The equation of a line

German Research Center for Artificial Intelligence

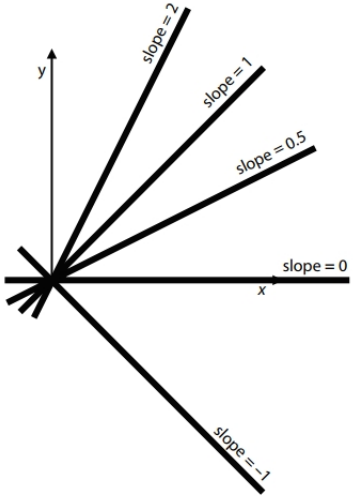
- Two components:
  - The slope
  - The y-intercept
- $y = 0.5x + 2$

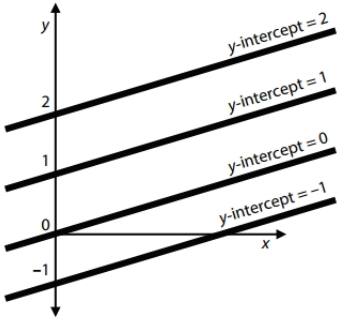


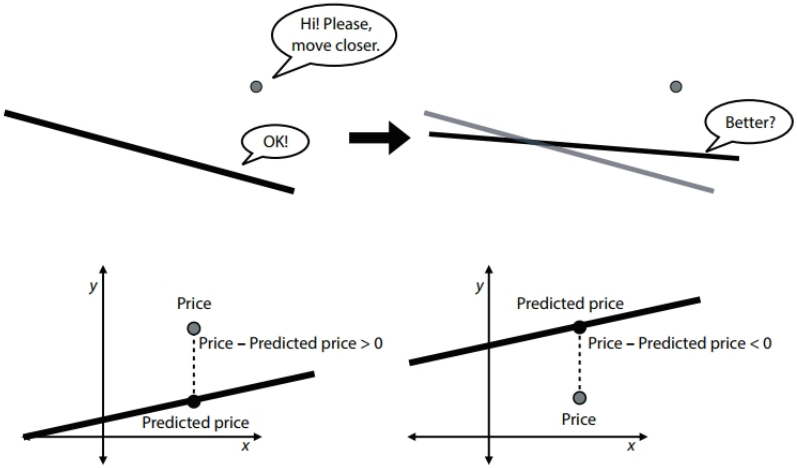
Slope = 0.5



y-intercept = 2


y
slope = 2
slope = 1
slope = 0.5
slope = 0
x
slope = –1

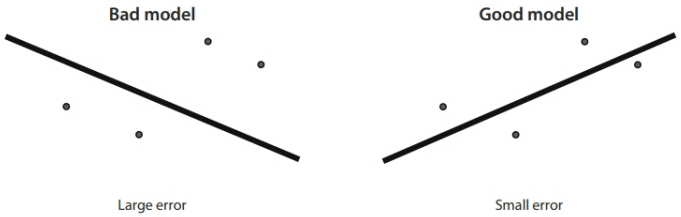

y
2
1
0
–1
y-intercept = 2
y-intercept = 1
y-intercept = 0
y-intercept = –1
x

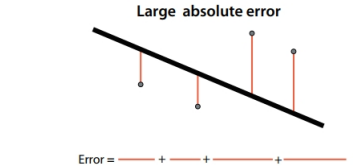

Hi! Please, move closer.
OK!
Better?
y
Price
Price – Predicted price > 0
Predicted price
x
y
Predicted price
Price – Predicted price < 0
Price
x

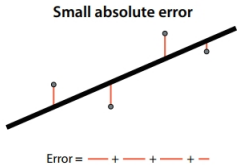
The error function

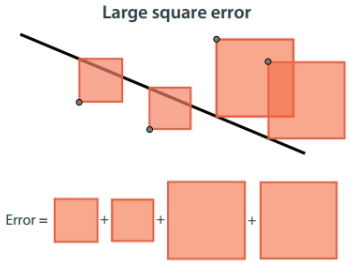
German Research Center for Artificial Intelligence

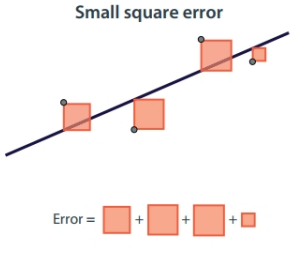
- The error function is a metric that tells how our model is doing
- Error function is also sometimes called loss function or cost function in the literature
- how do we define a good error function for linear regression models?
  - Absolute error is the sum of vertical distances from the line to the points in the dataset, and the
  - Square error is the sum of the squares of these distances




Large absolute error
Error =


Small absolute error
Error =

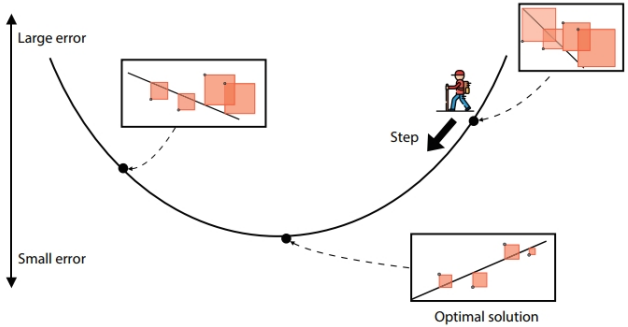

Large square error
Error =


Small square error
Error =

Gradient descent

German Research Center for Artificial Intelligence

- How to decrease an error function by slowly descending from a mountain
- This process uses derivatives to minimize the error function



Linear & Nonlinear & Logistic & Quantile regression

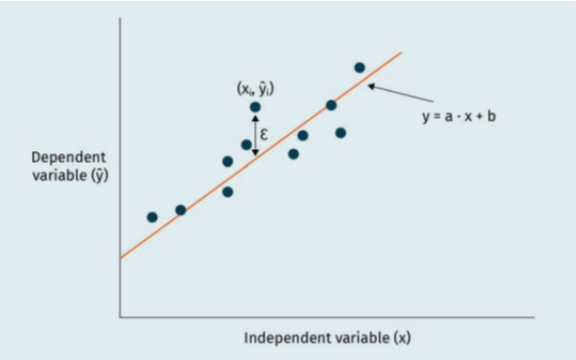
Linear & Nonlinear regression

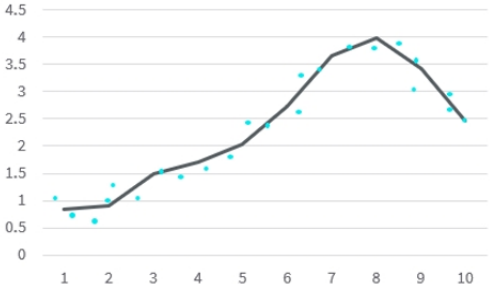
German Research Center for Artificial Intelligence

- Regression: This statistical method attempts to determine the strength and character of the relationship between dependent variable(s) and independent variable(s)
- In linear regression, the relationship is a straight-line equation: $y = a_1 x_1 + a_2 x_2 + \cdots + a_n x_n + b$
  - where $\{x_1, x_2, \ldots, x_n\}$ are the independent variables of the dataset, and $\{a_1, a_2, \ldots, a_n\}$ are the coefficients of these variables, implying the weight that each independent variable shares in the resultant target variable y. The b value is the constant or bias term
- in nonlinear regression, the relationship is a nonlinear equation (e.g., polynomial, exponential): $y = a_1 x_1^2 + a_2 x_2^3 + \cdots + a_n x_n^{n+1} + b$

Linear & Nonlinear regression

German Research Center for Artificial Intelligence





Simple linear regression

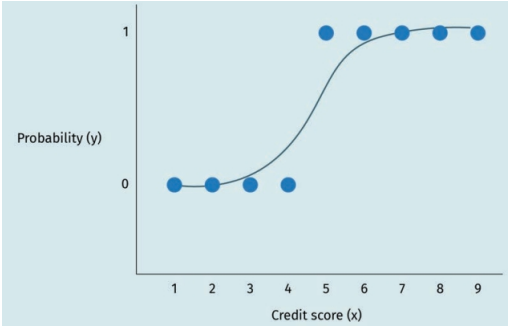
German Research Center for Artificial Intelligence

- $y = ax + b$
- $\hat{y} = y + \epsilon$
- For each data point, the difference between the predicted value y and the actual observation $\hat{y}$ is the residue ϵ
- The evaluation metric for the simple regression model is the sum of the squared residues (i.e., sum of the squared errors E), which we aim to minimize
  - $E = \sum(\hat{y} - y)^2 = \sum(\hat{y} - (ax + b))^2$

Logistic regression

German Research Center for Artificial Intelligence

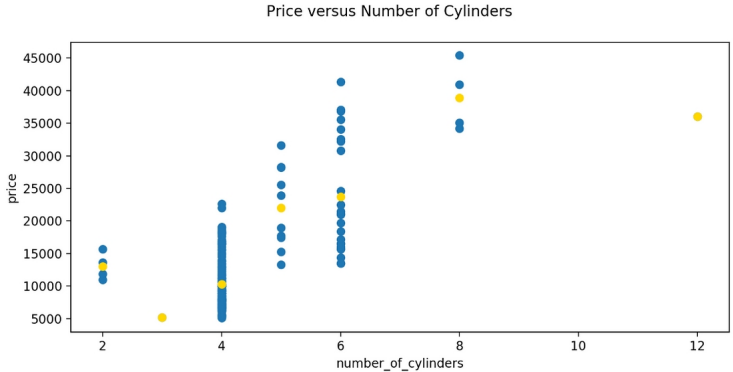
- Logistic regression is considered an extension of the linear regression analysis, and its corresponding model can be used to classify input data records into a set of given categories or discrete values that form the dependent variable
- This statistical model uses a logistic function to model a binary dependent variable



Quantile regression

German Research Center for Artificial Intelligence

- In a regression model, one is normally interested in estimating the conditional mean of the response variable

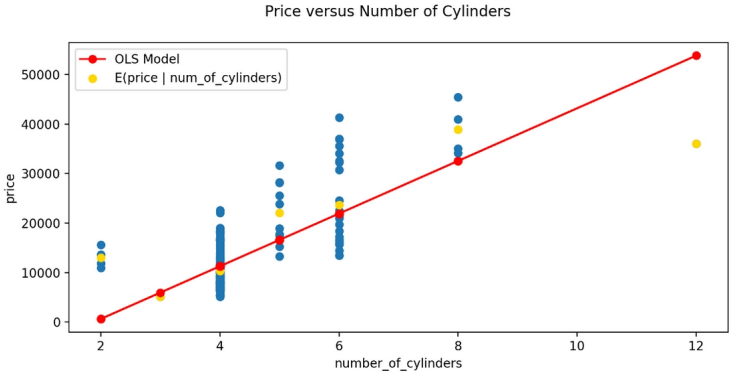


The gold dots represent the observed mean price conditioned upon the number of cylinders

Quantile regression

German Research Center for Artificial Intelligence

- If we wanted to estimate this conditional mean price using a regression model, we could employ the following linear model:
  $E(\text{price}|\text{num-of-cylinders}_i) = b + a_i * \text{num-of-cylinders}_i$



Quantile regression

German Research Center for Artificial Intelligence

- In some data sets, the mean is not a good exemplar of the data.
  - The data are highly skewed to the left or to the right. Such is often the case in insurance claims or healthcare costs data where most claims are small valued but the data has a long tail of claims of increasing value
  - The data is multi-modal i.e. it has more than one highly frequently occurring value
  - The data contains influential outliers
- The mean does not adequately represent the nature of the data. One may be better served by estimating the conditional median

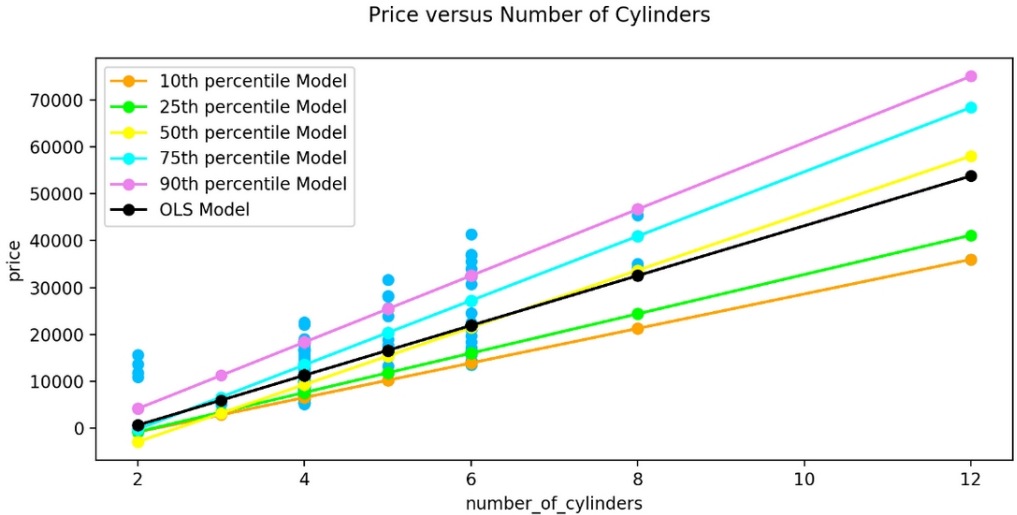
Quantile regression

German Research Center for Artificial Intelligence

- A model for estimating the median price of automobiles
  Median(price|num-of-cylinders$_i$) $= b + a_i *$ num-of-cylinders$_i$, where
  Probability(price $\leq b + a_i *$ num-of-cylinders$_i$) $= 0.5$
- We may even go a step further. In a data set, the median is the 0.5 quantile (or 50th percentile) point meaning that 50% of the data points are less than the value of the median
- Similarly, there are other quantile points that can be defined
  - The 0.1 quantile point (10th percentile) is the value such that only 10% of the data set is smaller than this value

Quantile regression

German Research Center for Artificial Intelligence



Regularization

Regularization

German Research Center for Artificial Intelligence

- Regression analysis tries to develop the best fit between the independent variables and the dependent variable such that the loss function is minimized
- the developed model can **overfit** the training dataset and lose the value of generalization when applied to a different testing dataset
- To avoid the issues of overfitting and outliers and to have a more robust model, we penalize the loss function by adding a penalty term to the regression model

Regularization

German Research Center for Artificial Intelligence

- Ridge Regression or L_2 L2regularization

$$E = \sum(\hat{y} - y)^2 + \lambda \sum W^2 \tag{1}$$

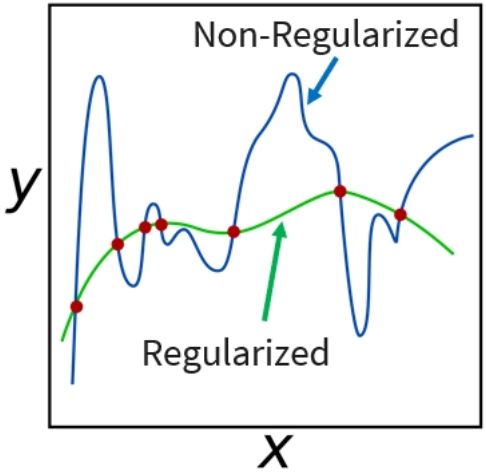
- Lasso Regression or L_1 L2regularization

$$E = \sum(\hat{y} - y)^2 + \lambda \sum |W| \tag{2}$$

- Selecting the optimum value of the regularization parameter λ is a trade-off

Regularization

German Research Center for Artificial Intelligence



Transfer task

German Research Center for Artificial Intelligence

- https://www.anaconda.com/products/distribution
- https://scikit-learn.org/stable/
- https://www.tutorialspoint.com/scikit_learn/index.htm

Questions?