INFO 6105
Data Science Engineering Methods and Tools

Lecture 1
Introduction to ML and Linear Regression

Ebrahim Nasrabadi
nasrabadi@northeastern.edu

College of Engineering
Northeastern University

Fall 2019

Course Information

This course will introduce the students to many concepts, techniques, and algorithms in data mining and machine learning including

- data cleaning and preprocessing
- linear regression and logistic regression
- accuracy metrics
- feature engineering & model selection
- decision trees, random forests and boosting

and ML tools including

- R (statistical tool)
- scikit-learn: machine learning in Python

Textbook

- *An introduction to Statistical Learning with Applications in R*, G. James, D. Witten, T. Hastie and R. Tibshirani

Grading

Assignments	30%
Quiz and Class participation	20%
Midterm Project	15%
Midterm	10%
Final Project	15%
Final	10%

Problem Sets

There are 3 problem sets. They are an important part of the learning experience, which is why they are required.

Quizzes

There are 4 quizzes.

Tools

We will use R and Python to illustrate concepts, analyze data sets, build predictive models, and evaluate the fit of the models.

Roughly Schedule

Week	Lecture
1	Introduction, Applications, and Concepts
2	Linear Regression
3	Polynomial Regression and Regularization
4	Classification and Logistic Regression
5	Model Evaluation and Cross-validation
6	Model Selection and Regularization
7	Introduction to scikit-learn
8	In class Lab: Predict Future Sales
9	Midterm & Project
10	Regression Trees
11	Classification Trees
12	Random Forests and Boosting
13	Gradient boosting
14	Neural Networks
15	In Class Lab: Click Prediction
16	Final Exam & Project

Instructor

Education

- B.S. in Applied Math,
  - Kerman University, 2002
- M.S. in Industrial Engineering
  - Sharif University, 2003
- PhD in CS and Applied Math
  - TU of Berlin, 2009

Work Experience

- Postdoc
  - MIT, 2010-2013
- Assistant Professor
  - Clemson Uni., 2013-2014
- Research Scientist
  - Amazon.com, 2014- 2018

Currently

Principal Applied Researcher, Expedia Group

Research Interests

- Broadly interested in optimization, machine learning, data science.
- Particularly interested in optimization under uncertainty and network optimization

Statistical and ML Problems

- What is the impact of a new feature or change to user experience?
- Is there any relationship between GPA and starting salary after graduation?
- What is the effect of package designs on sales?
- How to interpret polls. How many individuals you need to sample for your inferences to be acceptable? What is meant by the margin of error?
- What is the effect of market strategy on market share?
- How accurately can we predict demand for a product or service?
- What is the default risk (i.e., the chance that companies or individuals will be unable to make the required payments on their debt obligations)?
- What is the chance that a house goes under contract in its first two weeks on the market?

Statistical and ML Problems

- Detect spam emails
- Identify the numbers in a handwritten zip code.
- Establish the relationship between salary and demographic variables in population survey data
- Identify hot homes (e.g., homes that are likely to sell within two weeks)
- Estimate a home's value
- Predict whether someone will have a heart attack on the basis of demographic, diet and clinical measurements.
- Determine whether a loan will default
- Image Recognition
- Predict whether a user will
  - click on an ad
  - like a photo
  - buy a product
  - watch a movie

ML Applications

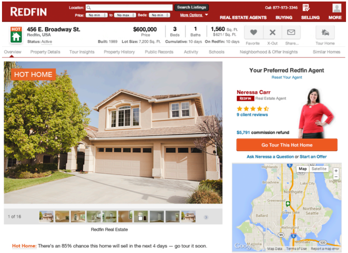
Hot Homes

Predict how quickly a home will sell.

Redfin takes into account more than 500 factors including

- the home's price
- property type
- neighborhood
- sale history
- how quickly similar homes in that neighborhood tend to sell

to predict how quickly a home will sell.



456 E. Broadway St.
Redfin, USA
Status: Active

$600,000 Price | 3 Beds | 1 Baths | 1,560 Sq. Ft. $621 / Sq. Ft.

Built: 1989 Lot Size: 7,200 Sq. Ft. Cumulative: 10 days On Redfin: 10 days

Redfin Real Estate

Hot Home: There's an 85% chance this home will sell in the next 4 days — go tour it soon.

Your Preferred Redfin Agent
Reset Your Agent

Neressa Carr
REDFIN Real Estate Agent
9 client reviews

$5,791 commission refund

Go Tour This Hot Home

Ask Neressa a Question or Start an Offer

ML Application

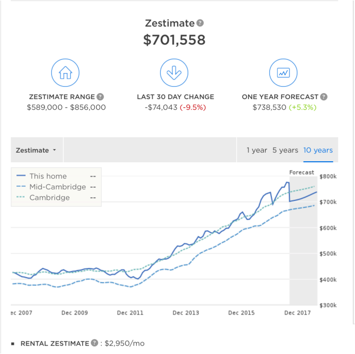
Zestimates

Estimate market value for an individual home

Zillow takes into account factors including

- property type
- location
- market conditions
- # of Bedrooms, # of Baths

to estimate a home's value and forecast the value one year from now, based on current home and market information.



Zestimate

$701,558

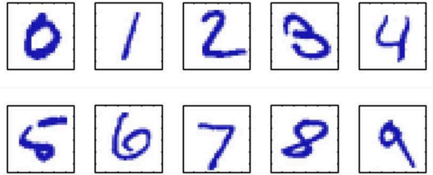
ZESTIMATE RANGE
$589,000 - $856,000

LAST 30 DAY CHANGE
-$74,043 (-9.5%)

ONE YEAR FORECAST
$738,530 (+5.3%)

RENTAL ZESTIMATE : $2,950/mo

Hand-written Digit Recognition



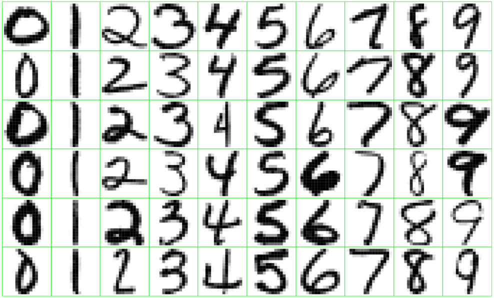
- Assume that Images are 28 × 28 pixels
- Represent input image as a vector $x \in R^{784}$
- Learn a classifier $f(x)$ such that,

$$f : x \rightarrow \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$$

Hand-written Digit Recognition

This is classification problem.

- Collect data:



- Model Building
- Cross Validation and Model Evaluation
- Model Deployment and Improvement:

Data Analysis

Before making inferences from data it is essential to examine all your variables.

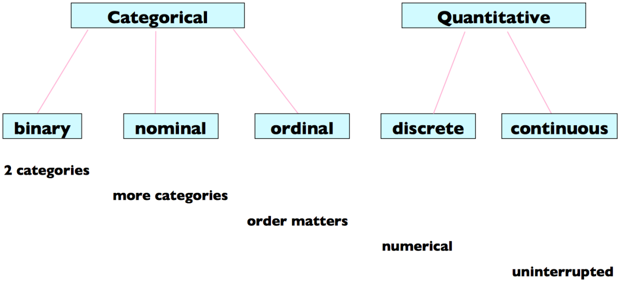
Why?

To listen to the data:

- to catch mistakes and missing values
- to see patterns in the data
- to find violations of statistical assumptions
- to generate hypotheses
- and because if you don't, you will have trouble later

Types of Data

- Quantitative data consist of values representing counts or measurements
  - Discrete: Units in a shipment
  - Continuos: Shipment weight
- Categorical data consist of values that can be placed into nonnumeric categories.
  - Binary: Hot Home or Not
  - Nominal: Property Type (House, Condo, Townhouse)
  - Ordinal: Medal Type (Gold, Silver, Bronze)



Categorical

Quantitative

binary

nominal

ordinal

discrete

continuous

2 categories

more categories

order matters

numerical

uninterrupted

Dimensionality of Data Sets

- Univariate: Measurement made on one feature per subject
    - User Activity: number of likes per user
    - Purchase History: number of units per customer
    - Shipment: weight per package
- Bivariate: Measurement made on two features per subject
    - User Activity: umber of likes and comments per user
    - Purchase History: dollar amount and number of units per customer
    - Shipment: weight and volume per package
- Multivariate: Measurement made on many features per subject
    - User Activity: number of likes, comments, views, ... per user
    - Purchase History: dollar amount, number of units, number of unique products, ... per customer
    - Shipment: weight, volume, and number of units, ... per package

Numerical Summaries of Data

- Central Tendency measures: They are computed to give a “center” around which the measurements in the data are distributed.
- Variation or Variability measures They describe “data spread” or how far away the measurements are from the center.

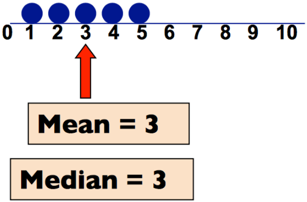
Central Tendency measures

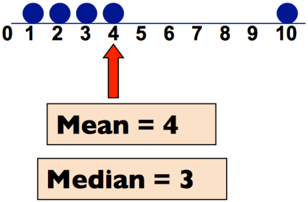
The most two common measures are

- Mean (average value): best for symmetric distributions without outliers

$$\bar{x} = \sum_{i=1}^{n} \frac{x_i}{n}$$

- Median (the exact middle value): Median is useful for skewed distributions or data with outliers





Variability measures

- Variance: Average of squared deviations of values from the mean

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

- Standard Deviation: Standard deviations are simply the square root of the variance

$$\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$$

Why Squared Deviations?

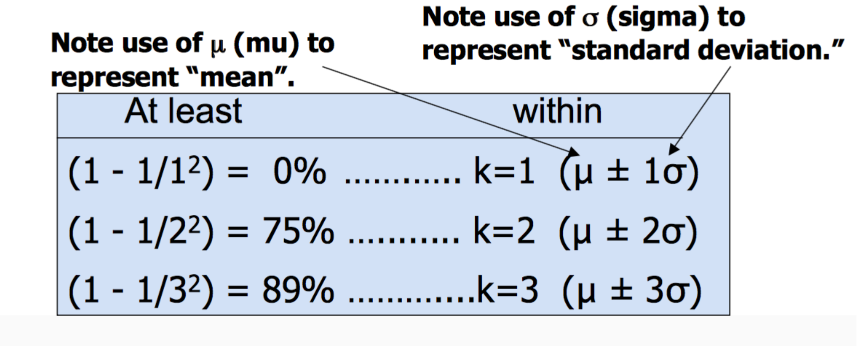
- Adding deviations will yield a sum of 0
- Absolute values do not have nice mathematical properties
- Squares eliminate the negatives

Result

Increasing contribution to the variance as you go farther from the mean

Chebyshev’s inequality

Regardless of how the data are distributed, a certain percentage of values ($1\text{-}1/k^2$) are within k standard deviations from the mean:



Note use of μ (mu) to represent “mean”.

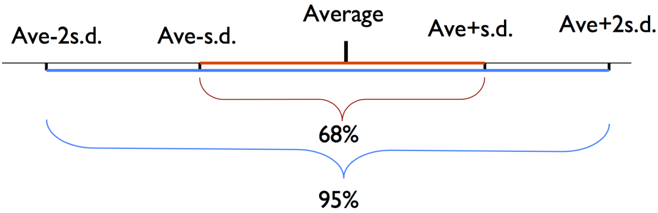
Note use of σ (sigma) to represent “standard deviation.”

At least	within
$(1 - 1/1^2) =$ 0%	k=1 $(\mu \pm 1\sigma)$
$(1 - 1/2^2) =$ 75%	k=2 $(\mu \pm 2\sigma)$
$(1 - 1/3^2) =$ 89%	k=3 $(\mu \pm 3\sigma)$

Often We Can Do Better

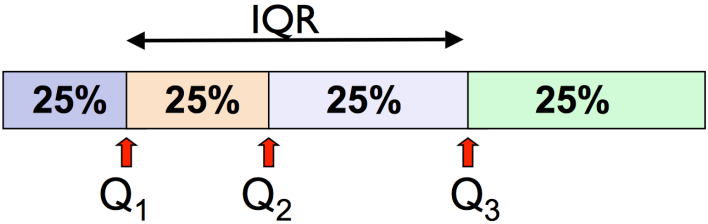
For many lists of observations, especially if their histogram is bell-shaped

- Roughly 68% of the observations in the list lie within 1 standard deviation of the average
- 95% of the observations lie within 2 standard deviations of the average



Variability measures: Quartiles and IQR

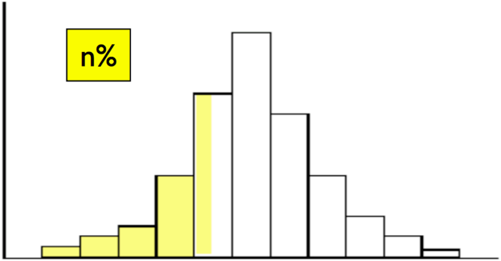
- The first quartile,Q1, is the value for which 25% of the observations are smaller and 75% are larger
- Only 25% of the observations are greater than the third quartile
- IQR= $Q_3 - Q_1$: Interquartile range, also called middle 50%,



Variability measures: Percentiles

In general the n^{th} percentile is a value such that $n\%$ of the observations fall at or below or it

- $Q_1 = 25^{th}$ percentile
- Median $= 50^{th}$ percentile
- $Q_2 = 75^{th}$ percentile



Graphical Summaries of Data

Bar plot

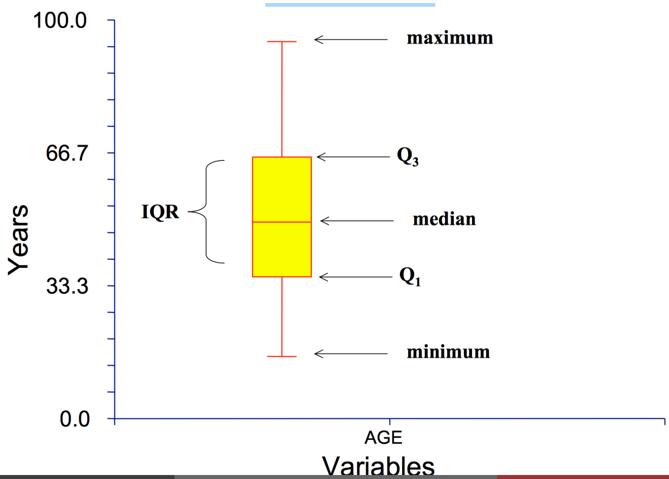
- Used for categorical variables to show frequency or proportion in each category.
- Translate the data from frequency tables into a pictorial representation...

Histogram

- Used to visualize distribution (shape, center, range, variation) of continuous variables
- “Bin size” important

Box Plots

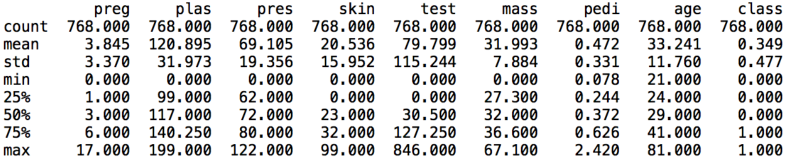
The box plot is a standardized way of displaying the distribution of data based on the five number summary: minimum, first quartile, median, third quartile, and maximum.



Pima Indians Diabetes Data Set

- Sources: National Institute of Diabetes and Digestive and Kidney Diseases
- Objective: Predict based on diagnostic measurements whether a patient has diabetes
- Number of Instances: 768 (all females and at least 21 years old)
- Number of Features: 8 plus class
  1. Number of times pregnant
  2. Plasma glucose concentration a 2 hours in an oral glucose tolerance test
  3. Diastolic blood pressure (mm Hg)
  4. Triceps skin fold thickness (mm)
  5. 2-Hour serum insulin (mu U/ml)
  6. Body mass index (weight in kg/(height in m)^2)
  7. Diabetes pedigree function
  8. Age (years)
  9. Class variable (0 or 1)
- Download dataset: Machine Learning Repository

Descriptive Statistics



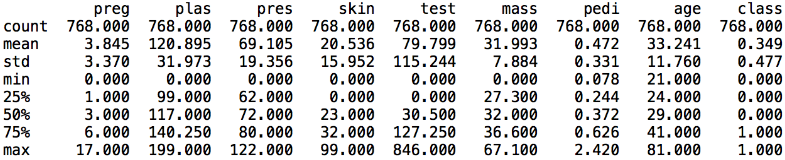
```
         preg     plas     pres     skin     test     mass     pedi      age    class
count 768.000  768.000  768.000  768.000  768.000  768.000  768.000  768.000  768.000
mean    3.845  120.895   69.105   20.536   79.799   31.993    0.472   33.241    0.349
std     3.370   31.973   19.356   15.952  115.244    7.884    0.331   11.760    0.477
min     0.000    0.000    0.000    0.000    0.000    0.000    0.078   21.000    0.000
25%     1.000   99.000   62.000    0.000    0.000   27.300    0.244   24.000    0.000
50%     3.000  117.000   72.000   23.000   30.500   32.000    0.372   29.000    0.000
75%     6.000  140.250   80.000   32.000  127.250   36.600    0.626   41.000    1.000
max    17.000  199.000  122.000   99.000  846.000   67.100    2.420   81.000    1.000
```

Observations

There are columns that have a minimum value of zero. On some columns, a value of zero does not make sense and indicates an invalid or missing value.

- Plasma glucose concentration: 5
- Diastolic blood pressure: 35
- Triceps skinfold thickness: 227
- 2-Hour serum insulin: 374
- Body mass index: 11

Descriptive Statistics



```
         preg     plas     pres     skin     test     mass     pedi      age    class
count 768.000  768.000  768.000  768.000  768.000  768.000  768.000  768.000  768.000
mean    3.845  120.895   69.105   20.536   79.799   31.993    0.472   33.241    0.349
std     3.370   31.973   19.356   15.952  115.244    7.884    0.331   11.760    0.477
min     0.000    0.000    0.000    0.000    0.000    0.000    0.078   21.000    0.000
25%     1.000   99.000   62.000    0.000    0.000   27.300    0.244   24.000    0.000
50%     3.000  117.000   72.000   23.000   30.500   32.000    0.372   29.000    0.000
75%     6.000  140.250   80.000   32.000  127.250   36.600    0.626   41.000    1.000
max    17.000  199.000  122.000   99.000  846.000   67.100    2.420   81.000    1.000
```

Observations

There are columns that have a minimum value of zero. On some columns, a value of zero does not make sense and indicates an invalid or missing value.

- Plasma glucose concentration: 5
- Diastolic blood pressure: 35
- Triceps skinfold thickness: 227
- 2-Hour serum insulin: 374
- Body mass index: 11

Descriptive Statistics

Observations (Cont.)

There are columns that have a minimum value of zero. On some columns, a value of zero does not make sense and indicates an invalid or missing value.

- We can see that columns 1,2 and 5 have just a few zero values, whereas columns 3 and 4 show a lot more, nearly half of the rows.
- This highlights that different “missing value” strategies may be needed for different columns, e.g. to ensure that there are still a sufficient number of records left to train a predictive model.

Missing Values

Why is it important?

- Wrong conclusions
- Some statistical models fail if there is missing values
- Leading to poor performance

How to handle missing values?

- Remove Rows With Missing Values
- Impute Missing Values (e.g., Use a model to replace missing values.)
  - A constant value that has meaning within the domain, such as 0, distinct from all other values.
  - A value from another randomly selected record.
  - A mean, median or mode value for the column.
  - A value estimated by another predictive model.

Missing Values

Why is it important?

- Wrong conclusions
- Some statistical models fail if there is missing values
- Leading to poor performance

How to handle missing values?

- Remove Rows With Missing Values
- Impute Missing Values (e.g., Use a model to replace missing values.)
  - A constant value that has meaning within the domain, such as 0, distinct from all other values.
  - A value from another randomly selected record.
  - A mean, median or mode value for the column.
  - A value estimated by another predictive model.

Missing Values

Notes

- Any imputing performed on the training dataset will have to be performed on new data in the future when predictions are needed from the finalized model. This needs to be taken into consideration when choosing how to impute the missing values.
- For example, if you choose to impute with mean column values, these mean column values will need to be stored to file for later use on new data that has missing values.

Unsupervised vs Supervised Learning

Supervised Machine Learning

We have input variables X (also called *features*) and an output variable y (also called *response* or *target*) and you use an algorithm to learn a mapping function from the input to the output.

$$y = h(X)$$

Goal

Approximate the mapping function so well that when you have new input data X that you can predict the output variable y.

Supervised learning problems are grouped into regression and classification problems.

- Classification: A classification problem is when the output variable is a category, such as "click" or"non-click"
- Regression: A regression problem is when the output variable is a real value, such as "sales".

Supervised Machine Learning

We have input variables X (also called *features*) and an output variable y (also called *response* or *target*) and you use an algorithm to learn a mapping function from the input to the output.

$$y = h(X)$$

Goal

Approximate the mapping function so well that when you have new input data X that you can predict the output variable y.

Supervised learning problems are grouped into regression and classification problems.

- Classification: A classification problem is when the output variable is a category, such as “click” or“non-click”
- Regression: A regression problem is when the output variable is a real value, such as “sales”.

Supervised Machine Learning

We have input variables X (also called *features*) and an output variable y (also called *response* or *target*) and you use an algorithm to learn a mapping function from the input to the output.

$$y = h(X)$$

Goal

Approximate the mapping function so well that when you have new input data X that you can predict the output variable y.

Supervised learning problems are grouped into regression and classification problems.

- Classification: A classification problem is when the output variable is a category, such as "click" or"non-click"
- Regression: A regression problem is when the output variable is a real value, such as "sales".

Machine Learning Algorithms

Supervised Algorithms

- Linear regression for regression problems.
- Logistic regression for classification problems.
- Tree-based models for classification and regression problems.
- Support vector machines for classification problems.

Linear Regression

- simple approach for supervised learning
- useful tool for prediction a quantitative response and understand the relationship between data

Example

Suppose that we have training examples.

How can we use learn from this data to

- predict y for a new x?
- measure the percentage change in y in response to a percent change in x?

x	y
\$10.99	5.89
\$11.99	5.29
\$12.99	4.76
\$13.99	4.61
\$14.99	3.82
\$15.99	3.34
\$16.99	3.35
\$17.99	3.21
\$18.99	3.08
\$19.99	3.01
\$21.99	2.93

Linear Regression

- simple approach for supervised learning
- useful tool for prediction a quantitative response and understand the relationship between data

Example

Suppose that we have training examples.
How can we use learn from this data to

- predict y for a new x?
- measure the percentage change in y in response to a percent change in x?

x	y
\$10.99	5.89
\$11.99	5.29
\$12.99	4.76
\$13.99	4.61
\$14.99	3.82
\$15.99	3.34
\$16.99	3.35
\$17.99	3.21
\$18.99	3.08
\$19.99	3.01
\$21.99	2.93

Sales vs Price

Linear regression

We can represent data as

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

where $x_i =$ price and $y_i =$ average sales for $i^{th} observation$.

We seek a function/hypothesis $h : \mathbb{R}_{\geq 0} \longrightarrow \mathbb{R}_{\geq 0}$ such that $h(x)$ is a good predictor for corresponding values of y.

Modeling Choice: We assume that dependency of y on x is linear and approximate y as a linear function of x:

$$h_\beta(x) = \beta_0 + \beta_1 x = \beta^T X$$

Here, β_0 and β_1 are called *model parameters* or *weights*.

Linear regression

We can represent data as

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

where $x_i =$ price and $y_i =$ average sales for $i^{th} observation$.

We seek a function/hypothesis $h : \mathbb{R}_{\geq 0} \longrightarrow \mathbb{R}_{\geq 0}$ such that $h(x)$ is a good predictor for corresponding values of y.

Modeling Choice: We assume that dependency of y on x is linear and approximate y as a linear function of x:

$$h_\beta(x) = \beta_0 + \beta_1 x = \beta^T X$$

Here, β_0 and β_1 are called *model parameters* or *weights*.

How to determine β?

Find β_0 and β_1 such that the linear model fits the training data well

input	actual	predicted	residual/error
x_1	y_1	$\bar{y}_1 = \beta_0 + \beta_1 x_1$	$e_1 = y_1 - \bar{y}_1$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_i	y_i	$\bar{y}_i = \beta_0 + \beta_1 x_i$	$e_i = y_i - \bar{y}_i$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_n	y_n	$\bar{y}_n = \beta_0 + \beta_n x_n$	$e_n = y_n - \bar{y}_n$

How to determine β?

We define Residual Sum of Squares (RSS) and Mean Square Error (MSE) as

$$\begin{aligned}\text{RSS} :=& e_1^2 + e_2^2 + \ldots + e_n^2 = \sum_{i=1}^{n} e_i^2 \\ =& \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \\ \text{MSE} :=& \text{RSS}/n\end{aligned}$$

This is sometimes called the loss function and denoted by $L(\beta)$.

Choose β so as to minimize RSS.

How to determine β?

We define Residual Sum of Squares (RSS) and Mean Square Error (MSE) as

$$\begin{aligned} \text{RSS} :=& e_1^2 + e_2^2 + \ldots + e_n^2 = \sum_{i=1}^{n} e_i^2 \\ =& \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \\ \text{MSE} :=& \text{RSS}/n \end{aligned}$$

This is sometimes called the loss function and denoted by $L(\beta)$.

Choose β so as to minimize RSS.

Ordinary Least Squares (OLS)

For simplicity, assume one single variable. Then,

$$L(\beta_0, \beta_1) = \frac{1}{n} \sum_{i=1}^{n} \Big(y_i - (\beta_0 + \beta_1 x_i) \Big)^2$$

To find the β, we need to take the derivative of L with respect to β_0, β_1.

$$\frac{\partial L}{\partial \beta_0} = \frac{1}{n} \sum_{i=1}^{n} 2 \Big(y_i - (\beta_0 + \beta_1 x_i) \Big)(-1)$$

$$\frac{\partial L}{\partial \beta_1} = \frac{1}{n} \sum_{i=1}^{n} 2 \Big(y_i - (\beta_0 + \beta_1 x_i) \Big)(-x_i)$$

We look for β_0, β_1 that satisfy $\frac{\partial L}{\partial \beta_0} = 0$ and $\frac{\partial L}{\partial \beta_1} = 0$

Ordinary Least Squares (OLS)

$$
\begin{aligned}
\frac{\partial L}{\partial \beta_0} = 0 &\implies -\frac{1}{n}\sum_{i=1}^{n} y_i + \beta_1 \frac{1}{n}\sum_{i=1}^{n} x_i + \beta_0 = 0 \\
&\implies \beta_0 = \frac{1}{n}\sum_{i=1}^{n} y_i - \beta_1 \frac{1}{n}\sum_{i=1}^{n} x_i \\
&\implies \beta_0 = \bar{y} - \beta_1 \bar{x}
\end{aligned}
$$

Ordinary Least Squares (OLS)

$$
\begin{aligned}
\frac{\partial L}{\partial \beta_1} = 0 \implies & -\frac{1}{n}\sum_{i=1}^{n} x_i y_i + \beta_1 \frac{1}{n}\sum_{i=1}^{n} x_i^2 + \beta_0 \sum_{i=1}^{n} x_i = 0 \\
\implies & -\frac{1}{n}\sum_{i=1}^{n} x_i y_i + \beta_1 \frac{1}{n}\sum_{i=1}^{n} x_i^2 + \\
& \qquad \Big(\frac{1}{n}\sum_{i=1}^{n} y_i - \beta_1 \frac{1}{n}\sum_{i=1}^{n} x_i\Big) \sum_{i=1}^{n} x_i = 0 \\
\implies & \beta_1 \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \beta_1 \big(\frac{1}{n}\sum_{i=1}^{n} x_i\big)^2 = \\
& \qquad \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \big(\frac{1}{n}\sum_{i=1}^{n} x_i\big)\big(\frac{1}{n}\sum_{i=1}^{n} y_i\big) \\
\implies & \beta_1 = \frac{\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)\left(\frac{1}{n}\sum_{i=1}^{n} y_i\right)}{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2}
\end{aligned}
$$

Ordinary Least Squares (OLS)

Optimal Choices:

$$\beta_0 = \frac{1}{n}\sum_{i=1}^{n} y_i - \beta_1 \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\beta_1 = \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

Interpretation of β_1

- $\beta_1 > 0$: Positive correlation – There is a relationship between two variables in which both variables move in tandem
- $\beta_1 < 0$: Negative correlation – There is a relationship between two variables in which one variable increases as the other decreases
- $\beta_1 = 0$: No correlation – There is a no relationship between two variables

Ordinary Least Squares (OLS)

Optimal Choices:

$$
\begin{aligned}
\beta_0 &= \frac{1}{n}\sum_{i=1}^{n} y_i - \beta_1 \frac{1}{n}\sum_{i=1}^{n} x_i \\
\beta_1 &= \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\big(\sum_{i=1}^{n} x_i\big)\big(\sum_{i=1}^{n} y_i\big)}{\sum_{i=1}^{n} x_i^2 - \big(\sum_{i=1}^{n} x_i\big)^2} \\
&= \frac{\operatorname{Cov}(X, Y)}{\operatorname{Var}(X)}
\end{aligned}
$$

Multiple Linear regression

Question

How do we extend if we have many variables?

Training Examples: $(\vec{x}_1, y_1), (\vec{x}_2, y_2), \ldots, (\vec{x}_n, y_n)$ where
$\vec{x}_i = (x_{i1}, \ldots, x_{ij}, \ldots, x_{im})^T$
Linear Model:

$$h_\beta(\vec{x}) = \vec{\beta}^T \vec{x}$$

Loss Function:

$$\begin{aligned} L(\vec{\beta}) &= \frac{1}{n} \sum_{i=1}^{n} \Big(y_i - (\beta_0 + \beta_1 x_{i1} + \ldots + \beta_j x_{ij} + \ldots \beta_m x_{im}) \Big)^2 \\ &= \frac{1}{n} ||\vec{y} - X\vec{\beta}||^2 \end{aligned}$$

where X is a $n \times (m+1)$ matrix with 1's in the first column.

Multiple Linear regression

Note that

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1m} \\ 1 & x_{21} & x_{22} & \dots & x_{2m} \\ 1 & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nm} \end{bmatrix} \qquad \vec{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix} \qquad \vec{\beta} = \begin{bmatrix} 1 \\ \beta_1 \\ \vdots \\ \beta_m \end{bmatrix}$$

where x_{ij} is the value of feature j and y_i is the response for $i$the observation. Then, we can write

$$X\vec{\beta} = \begin{bmatrix} \beta_0 + \beta_1 x_{11} + \ldots + \beta_m x_{1m} \\ \beta_0 + \beta_1 x_{21} + \ldots + \beta_m x_{2m} \\ \vdots \\ \beta_0 + \beta_1 x_{1n} + \ldots + \beta_m x_{nm} \end{bmatrix} \quad \vec{y} - X\vec{\beta} = \begin{bmatrix} y_1 - (\beta_0 + \beta_1 x_{11} + \ldots + \beta_m x_{1m}) \\ y_2 - (\beta_0 + \beta_1 x_{21} + \ldots + \beta_m x_{2m}) \\ \vdots \\ y_n - (\beta_0 + \beta_1 x_{1n} + \ldots + \beta_m x_{nm}) \end{bmatrix}$$

How to determine $\vec{\beta}$?

A few facts from Matrix Calculus

$$\frac{\partial[X\vec{\beta}]}{\partial\vec{\beta}} = X^T \qquad \frac{\partial[\vec{\beta}^T A\vec{\beta}]}{\partial\vec{\beta}} = 2A^\top\vec{\beta}$$

Minimizing the Loss:

$$\begin{aligned}
||\vec{y} - X\vec{\beta}||^2 &= (\vec{y} - X\vec{\beta})^T(\vec{y} - X\vec{\beta}) \\
&= \vec{y}^T\vec{y} - (X\vec{\beta})^T\vec{y} - \vec{y}X\vec{\beta} + \vec{\beta}X^T X\vec{\beta} \\
&= \vec{y}^T\vec{y} - 2\vec{y}X\vec{\beta} + \vec{\beta}X^T X\vec{\beta}
\end{aligned}$$

Note that $\vec{y}^T\vec{y}$ is a constant. Hence

$$\frac{\partial L}{\partial\vec{\beta}} = -2\vec{X}^T\vec{y} + \vec{2}X^T X\vec{\beta}$$

Ordinary Least Squares (OLS)

Optimal Choices:

$$\begin{aligned}
\frac{\partial L}{\partial \vec{\beta}} = 0 &\implies -\vec{X}^T\vec{y} + \vec{X}^T X\vec{\beta} = 0 \\
&\implies \vec{X}^T X\vec{\beta} = \vec{X}^T\vec{y} \\
&\implies \vec{\beta} = (\vec{X}^T X)^{-1}\vec{X}^T\vec{y}
\end{aligned}$$

assuming $(\vec{X}^T X)^{-1}$ exists.

Ordinary Least Squares (OLS)

Simple Linear regression

$$
\begin{aligned}
h(x) &= \beta_0 + \beta_1 x \\
\frac{\partial L}{\partial \beta_0} &= 0 \quad \frac{\partial L}{\partial \beta_1} = 0 \\
\hat{\beta}_1 &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\
\hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x}
\end{aligned}
$$

Multiple Linear regression

$$
\begin{aligned}
h(X) &= X\vec{\beta} \\
L(\vec{\beta}) &= ||\vec{y} - X\vec{\beta}||^2 \\
&= (\vec{y} - X\vec{\beta})^T(\vec{y} - X\vec{\beta}) \\
&= \vec{y}^T\vec{y} - 2\vec{y}X\vec{\beta} + \vec{\beta}X^TX\vec{\beta} \\
\frac{\partial L}{\partial \vec{\beta}} &= 0 - 2X^T\vec{y} + 2X^TX\vec{\beta} \\
\frac{\partial L}{\partial \vec{\beta}} &= 0 \implies \vec{\beta} = (X^TX)^{-1}X^T\vec{y}
\end{aligned}
$$

Ordinary Least Squares (OLS)

Simple Linear regression

$$\begin{aligned}
h(x) &= \beta_0 + \beta_1 x \\
\frac{\partial L}{\partial \beta_0} &= 0 \quad \frac{\partial L}{\partial \beta_1} = 0 \\
\hat{\beta}_1 &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\
\hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x}
\end{aligned}$$

Multiple Linear regression

$$\begin{aligned}
h(X) &= X\vec{\beta} \\
L(\vec{\beta}) &= ||\vec{y} - X\vec{\beta}||^2 \\
&= (\vec{y} - X\vec{\beta})^T(\vec{y} - X\vec{\beta}) \\
&= \vec{y}^T\vec{y} - 2\vec{y}X\vec{\beta} + \vec{\beta}X^T X\vec{\beta} \\
\frac{\partial L}{\partial \vec{\beta}} &= 0 - 2X^T\vec{y} + 2X^T X\vec{\beta} \\
\frac{\partial L}{\partial \vec{\beta}} &= 0 \implies \vec{\beta} = (X^T X)^{-1} X^T \vec{y}
\end{aligned}$$

OLS Pros and Cons

Pros

- Efficient computation
- Unique minimum
- Stable under perturbation of data
- Easy to interpret

Cons

- Influenced by outliers
- $(X^\top X)^{-1}$ need not exist.

Hypothesis Testing

Question

Is there a relationship between x and y?

We perform a hypothesis test on the parameters. We test the *null hypothesis*

$$H_0 : \text{There is no relationship between } x \text{ and } y$$

versus the *alternative hypothesis*

$$H_A : \text{There is some relationship between } x \text{ and } y$$

Mathematically, this corresponds to testing

$$H_0 : \beta_1 = 0$$
$$H_A : \beta_1 \neq 0$$

Hypothesis Testing

Question

Is there a relationship between x and y?

We perform a hypothesis test on the parameters. We test the *null hypothesis*

$$H_0 : \text{There is no relationship between } x \text{ and } y$$

versus the *alternative hypothesis*

$$H_A : \text{There is some relationship between } x \text{ and } y$$

Mathematically, this corresponds to testing

$$H_0 : \beta_1 = 0$$
$$H_A : \beta_1 \neq 0$$

Hypothesis Testing

Question

Is there a relationship between x and y?

We perform a hypothesis test on the parameters. We test the *null hypothesis*

$$H_0 : \text{There is no relationship between } x \text{ and } y$$

versus the *alternative hypothesis*

$$H_A : \text{There is some relationship between } x \text{ and } y$$

Mathematically, this corresponds to testing

$$\begin{aligned} H_0 &: \beta_1 = 0 \\ H_A &: \beta_1 \neq 0 \end{aligned}$$

Standard Error

We need to determine whether $\hat{\beta}_1$ is sufficiently far from zero so we can be confident that $\beta_1 \neq 0$.

Question: How far is far enough?

It depends on the standard error that indicates how the estimated parameter varies under repeated sample.

Standard Error

$$\mathrm{SE}(\hat{\beta}_1) = \sqrt{\frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

$$\mathrm{SE}(\hat{\beta}_0) = \sqrt{\sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right]}$$

where $\sigma^2 = (\epsilon)$

Standard Error

We need to determine whether $\hat{\beta}_1$ is sufficiently far from zero so we can be confident that $\beta_1 \neq 0$.

Question: How far is far enough?

It depends on the standard error that indicates how the estimated parameter varies under repeated sample.

Standard Error

$$\mathrm{SE}(\hat{\beta}_1) = \sqrt{\frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

$$\mathrm{SE}(\hat{\beta}_0) = \sqrt{\sigma^2\left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]}$$

where $\sigma^2 = (\epsilon)$

Standard Error

We need to determine whether $\hat{\beta}_1$ is sufficiently far from zero so we can be confident that $\beta_1 \neq 0$.

Question: How far is far enough?

It depends on the standard error that indicates how the estimated parameter varies under repeated sample.

Standard Error

$$\mathrm{SE}(\hat{\beta}_1) = \sqrt{\frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

$$\mathrm{SE}(\hat{\beta}_0) = \sqrt{\sigma^2\left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]}$$

where $\sigma^2 = (\epsilon)$

Standard Error

Standard error can be used to compute the confidence intervals and perform hypothesis testing:

- 68% confidence interval for β_1:

$$\hat{\beta}_1 \pm \mathrm{SE}(\hat{\beta}_1)$$

- 95% confidence interval for β_1:

$$\hat{\beta}_1 \pm 2 \cdot \mathrm{SE}(\hat{\beta}_1)$$

Note: A 95% confidence interval is defined as a range of values such that with 95% probability, the range will contain the true unknown value of the parameter.

Standard Error

Standard error can be used to compute the confidence intervals and perform hypothesis testing:

- 68% confidence interval for β_1:

$$\hat{\beta}_1 \pm \mathrm{SE}(\hat{\beta}_1)$$

- 95% confidence interval for β_1:

$$\hat{\beta}_1 \pm 2 \cdot \mathrm{SE}(\hat{\beta}_1)$$

Note: A 95% confidence interval is defined as a range of values such that with 95% probability, the range will contain the true unknown value of the parameter.

Hypothesis Testing

To test

$$H_0 : \beta_1 = 0$$
$$H_A : \beta_1 \neq 0$$

we compute a *t-statistic*, given by

$$t = \frac{\hat{\beta}_1 - 0}{\mathrm{SE}(\hat{\beta}_1)}$$

The t-statistic measures the number of standard deviations that the parameter is away from zero.

The larger t-statistic is, the more likely we reject H_0.

p-value: the probability of observing any value equal to $|t|$ or larger assuming there is no relationship between x and y.

Hypothesis Testing

To test

$$H_0 : \beta_1 = 0$$
$$H_A : \beta_1 \neq 0$$

we compute a *t-statistic*, given by

$$t = \frac{\hat{\beta}_1 - 0}{\mathrm{SE}(\hat{\beta}_1)}$$

The t-statistic measures the number of standard deviations that the parameter is away from zero.

The larger t-statistic is, the more likely we reject H_0.

p-value: the probability of observing any value equal to $|t|$ or larger assuming there is no relationship between x and y.

Hypothesis Testing

To test

$$H_0 : \beta_1 = 0$$
$$H_A : \beta_1 \neq 0$$

we compute a *t-statistic*, given by

$$t = \frac{\hat{\beta}_1 - 0}{\mathrm{SE}(\hat{\beta}_1)}$$

The t-statistic measures the number of standard deviations that the parameter is away from zero.

The larger t-statistic is, the more likely we reject H_0.

p-value: the probability of observing any value equal to $|t|$ or larger assuming there is no relationship between x and y.

Overall accuracy of the model

R-squared

R-squared or fraction of variance explained by the linear model is

$$R^2 = \frac{\text{TSS} - \text{RSS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}}$$

where

$$\text{RSS} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$\text{TSS} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

Note that

$$0 \leq R^2 \leq 1.$$

In general, the higher the R-squared, the better the model fits the training data.

Overall accuracy of the model

R-squared

R-squared or fraction of variance explained by the linear model is

$$R^2 = \frac{\text{TSS} - \text{RSS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}}$$

where

$$\text{RSS} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$\text{TSS} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

Note that

$$0 \leq R^2 \leq 1.$$

In general, the higher the R-squared, the better the model fits the training data.

Overall accuracy of the model

R-squared will always increase if we add more features. To penalize R^2 as we add more variables, we use

$$\text{Adjusted } R^2 = 1 - \frac{\text{RSS}/(n-m-1)}{\text{TSS}/(n-1)}$$

Adjusted R^2 pays a price for inclusion of unnecessary features in the model.

Results for the example

	Coefficient	**Std. Error**	**t-statistic**	**p-value**
Intercept	8.20001	0.61754	13.279	3.24e-07
Price	-0.26363	0.03735	-7.058	5.94e-05

- R^2: 0.847 and Adjusted R^2: 0.83