Machine Learning - Regression

Professor Widom's Instructional Odyssey

www.professorwidom.org

Stanford University · Google · Instabase · Google Cloud Platform · Association for Computing Machinery · Very Large Data Bases Endowment Inc. · amazon web services

Data Tools and Techniques

- Basic Data Manipulation and Analysis
  Performing well-defined computations or asking well-defined questions ("queries")
- Data Mining
  Looking for patterns in data
- Machine Learning
  Using data to build models and make predictions
- Data Visualization
  Graphical depiction of data
- Data Collection and Preparation

Machine Learning

Using data to build models and make predictions

Supervised machine learning

- Set of labeled examples to learn from: training data
- Develop model from training data
- Use model to make predictions about new data

Unsupervised machine learning

- Unlabeled data, look for patterns or structure (similar to data mining)

Machine Learning

Using data to build models and make predictions

Supervised machine learning

- Set of labeled examples to learn from: training data
- Develop model from training data
- Use model to make predictions about new data

Unsupervised

- Unlabeled
  (similar to

Also…

- Semi-supervised learning
  Labeled + unlabeled
- Active learning
  Semi-supervised, ask user for labels
- Reinforcement learning
  Develop & refine model as data arrives

Regression

Using data to build models and make predictions

- Supervised
- Training data, each example:
  - Set of predictor values - “independent variables”
  - Numeric output value - “dependent variable”
- Model is function from predictors to output
  - Use model to predict output value for new predictor values
- Example
  - Predictors: mother height, father height, current age
  - Output: height

Other Types of Machine Learning

Using data to build models and make predictions

- Classification
  - Like regression except output values are labels or categories
  - Example
    - Predictor values: age, gender, income, profession
    - Output value: buyer, non-buyer
- Clustering
  - Unsupervised
  - Group data into sets of items similar to each other
  - Example - group customers based on spending patterns

Back to Regression

- Set of predictor values - "independent variables"
- Numeric output value - "dependent variable"
- Model is function from predictors to output

Training data

$w_1, x_1, y_1, z_1 \rightarrow o_1$

$w_2, x_2, y_2, z_2 \rightarrow o_2$

$w_3, x_3, y_3, z_3 \rightarrow o_3$

……

Model

$f(w, x, y, z) = o$

Back to Regression

Goal: Function f applied to training data should produce values as close as possible in aggregate to actual outputs

Training data

$w_1, x_1, y_1, z_1 \rightarrow o_1$
$w_2, x_2, y_2, z_2 \rightarrow o_2$
$w_3, x_3, y_3, z_3 \rightarrow o_3$
......

Model

$f(w, x, y, z) = o$

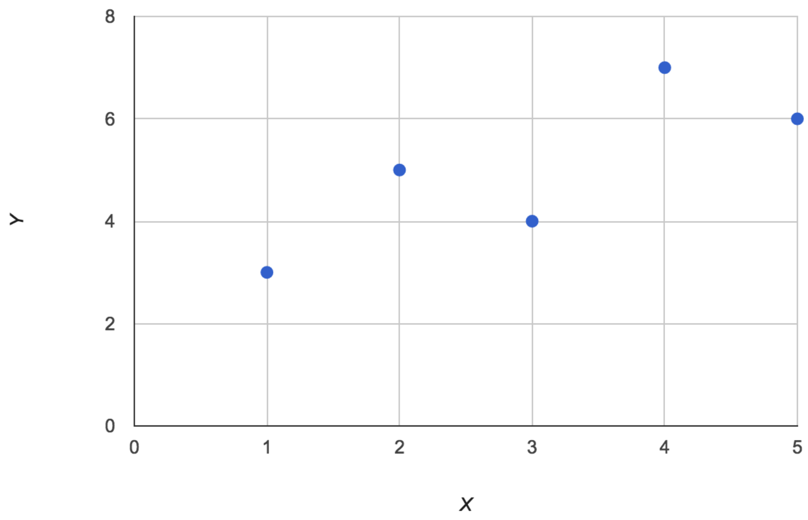
$f(w_1, x_1, y_1, z_1) = o_1'$
$f(w_2, x_2, y_2, z_2) = o_2'$
$f(w_3, x_3, y_3, z_3) = o_3'$

Simple Linear Regression

We will focus on:

- One numeric predictor value, call it *x*
- One numeric output value, call it *y*

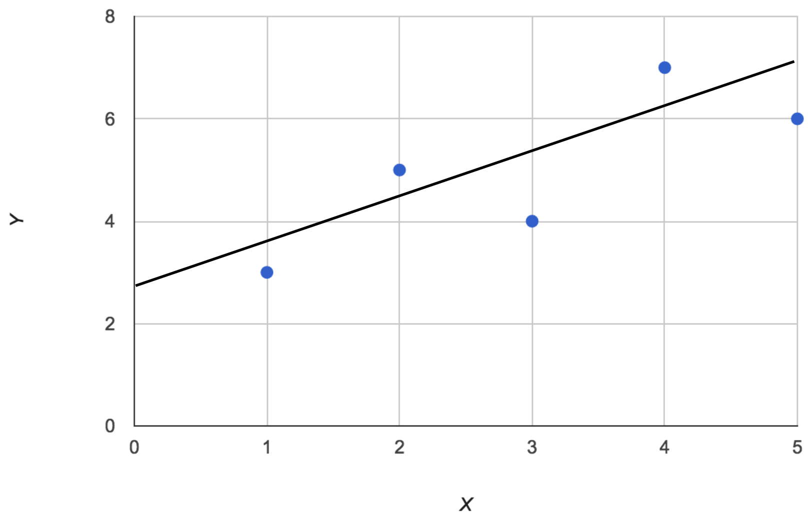
➢ Data items are points in two-dimensional space



Simple Linear Regression

We will focus on:

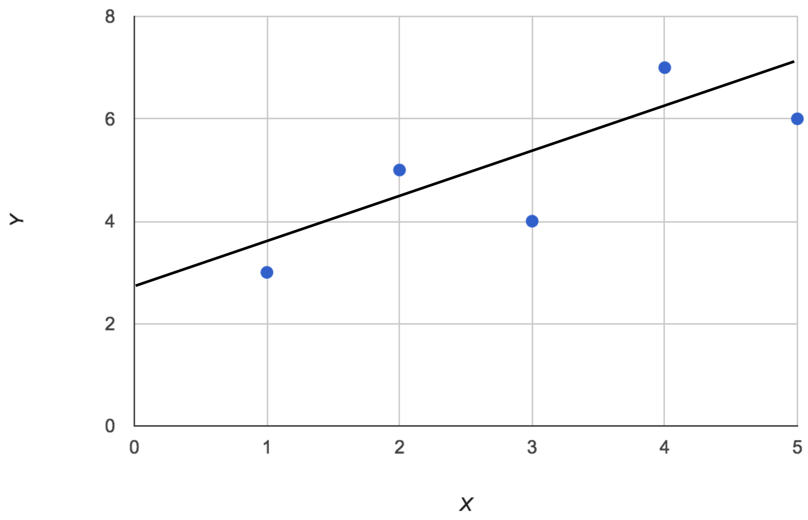
- One numeric predictor value, call it *x*
- One numeric output value, call it *y*
- Functions *f(x)=y* that are lines (for now)



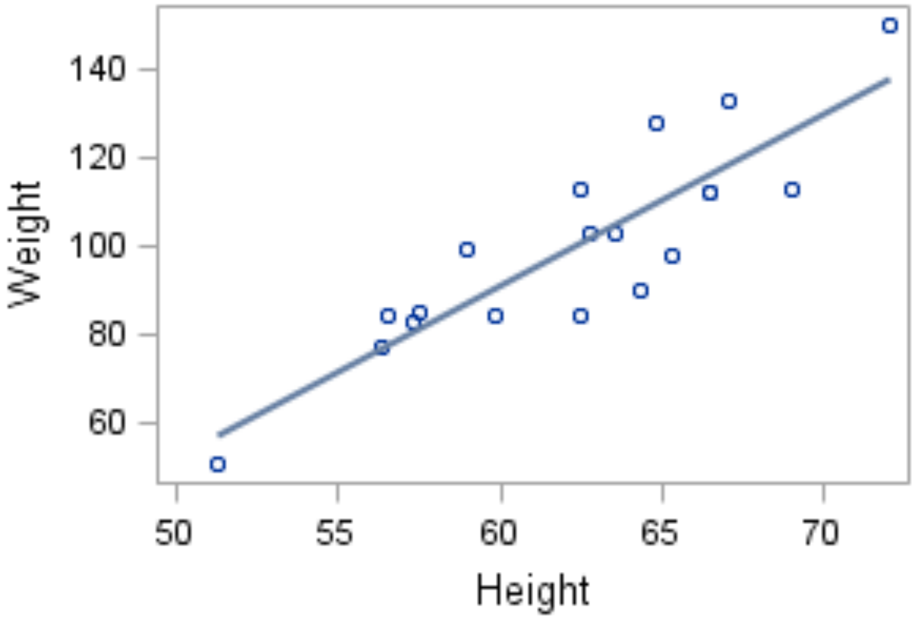
Simple Linear Regression

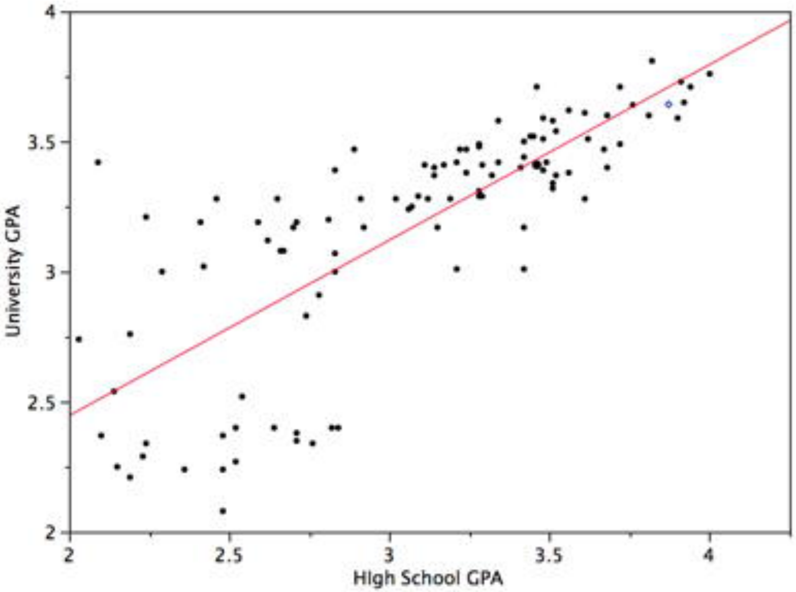
Functions $f(x)$=y that are lines: $y = ax + b$

$y = 0.8x + 2.6$



"Real" Examples (from Overview)

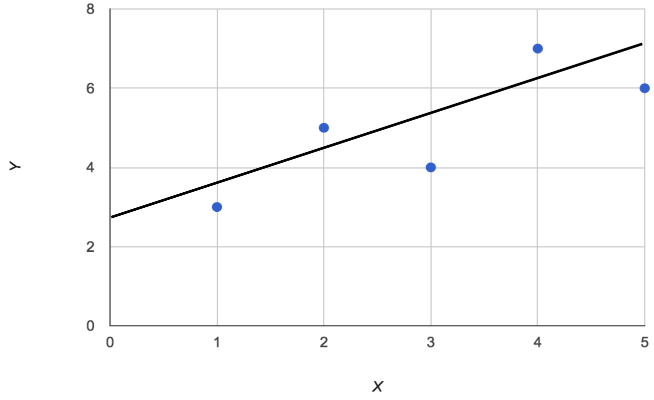




Summary So Far

- Given: Set of known (*x*,*y*) points
- Find: function $f(x)=ax+b$ that "best fits" the known points, i.e., $f(x)$ is close to *y*
- Use function to predict *y* values for new *x*'s

➢ Also can be used to test correlation



Correlation and Causation (from Overview)

Correlation – Values track each other

- Height and Shoe Size
- Grades and Entrance Exam Scores

Causation – One value directly influences another

- Education Level → Starting Salary
- Temperature → Cold Drink Sales

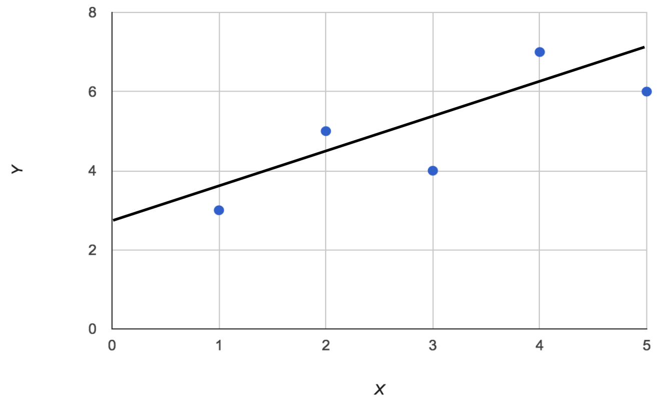
Correlation and Causation (from Overview)

Correlation – Values track each other

- Height and Shoe Size
- Grades and Entrance Exam Scores

Find: function *f(x)=ax+b* that "best fits" the known points, i.e., *f(x)* is close to *y*

The better the function fits the points, the more correlated *x* and *y* are



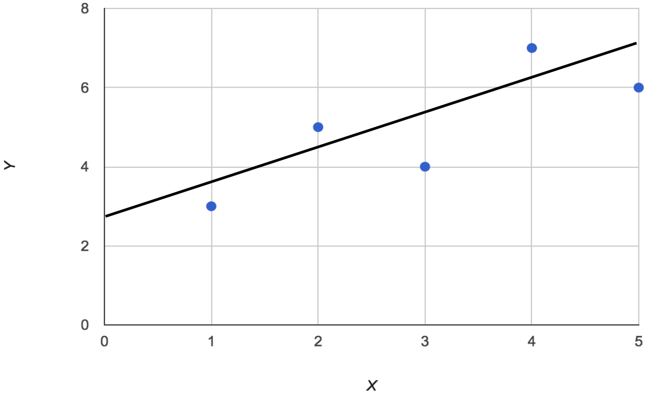
Regression and Correlation

The better the function fits the points,
the more correlated *x* and *y* are

- Linear functions only
- Correlation – Values track each other

  Positively – when one goes up the other goes up
- Also negative correlation

  When one goes up the other goes down
  - Latitude versus temperature
  - Car weight versus gas mileage
  - Class absences versus final grade



Next

- Calculating simple linear regression
- Measuring correlation
- Hands-on with datasets
- Shortcomings and dangers
- Polynomial regression
- Done!

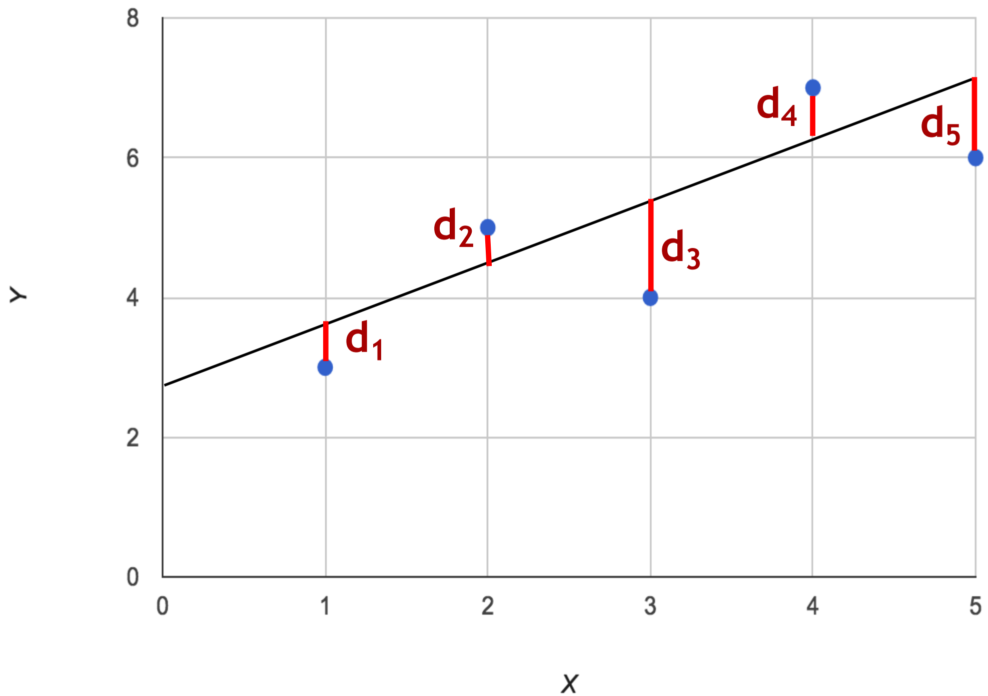
Calculating Simple Linear Regression

Method of least squares

- Given a point and a line, the error for the point is its vertical distance *d* from the line, and the squared error is d^2
- Given a set of points and a line, the sum of squared error (SSE) is the sum of the squared errors for all the points
- Goal: Given a set of points, find the line that minimizes the SSE

Calculating Simple Linear Regression

Method of least squares

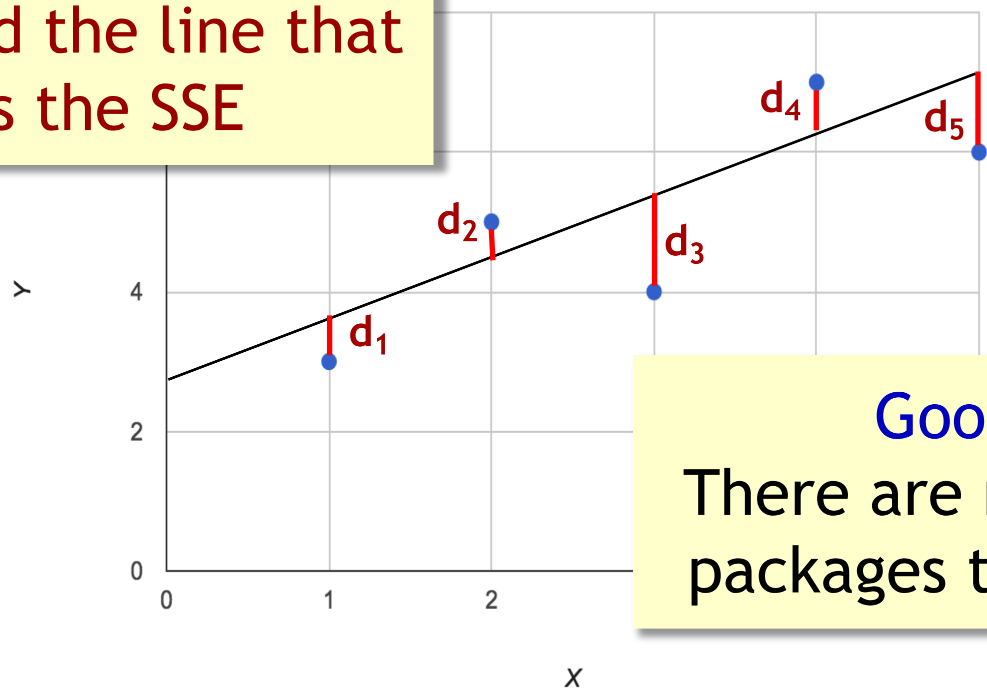


$$SSE = d_1^2 + d_2^2 + d_3^2 + d_4^2 + d_5^2$$

Calculating Simple Linear Regression

Method of least squares

Goal: Find the line that minimizes the SSE



Good News!
There are many software packages to do it for you

$$SSE = d_1^2 + d_2^2 + d_3^2 + d_4^2 + d_5^2$$

Measuring Correlation

More help from software packages…

Pearson's Product Moment Correlation (PPMC)

- "Pearson coefficient", "correlation coefficient"
- Value **r** between 1 and -1

  1 maximum positive correlation

  0 no correlation

  -1 maximum negative correlation

Coefficient of determination

"The better the function fits the points, the more correlated *x* and *y* are"

- r^2, R^2, "R squared"
- Measures fit of any line/curve to set of points
- Usually between 0 and 1
- For simple linear regression R^2 = Pearson^2

Measuring Correlation

More he ... ckages…

Swapping x and y axes yields same values

Pearson's Proc ... ion (PPMC)

- "Pearson coefficient", "correlation coefficient"
- Value **r** between 1 and -1

  1 maximum positive correlation
  0 no correlation
  -1 maximum negative correlation

Coefficient of determination

"The better the function fits the points, the more correlated *x* and *y* are"

- r^2, R^2, "R squared"
- Measures fit of any line/curve to set of points
- Usually between 0 and 1
- For simple linear regression R^2 = Pearson^2

Correlation Game

http://aionet.eu/corguess (*)

Try to get:

Right answers ≥ 10, Guesses ≤ Right answers × 2

Anti-cheating: Pictures = Right answers + 1

(*) Improved version of "Wilderdom correlation guessing game" thanks to Poland participant Marcin Piotrowski

Other correlation games:

http://guessthecorrelation.com/

http://www.rossmanchance.com/applets/GuessCorrelation.html

http://www.istics.net/Correlations/

Regression Through Spreadsheets

City temperatures (using Cities.csv)

1. temperature (y) versus latitude (x)
2. temperature (y) versus longitude (x)
3. longitude (y) versus temperature (x)

Your Turn

Correlations in the World Cup data

- Use Teams.csv and/or Players.csv *(unmodified)*
- Linear trendlines only
- What is the strongest positive correlation you can find? (highest R^2 value)
- What is the strongest negative correlation you can find? (highest R^2 value)

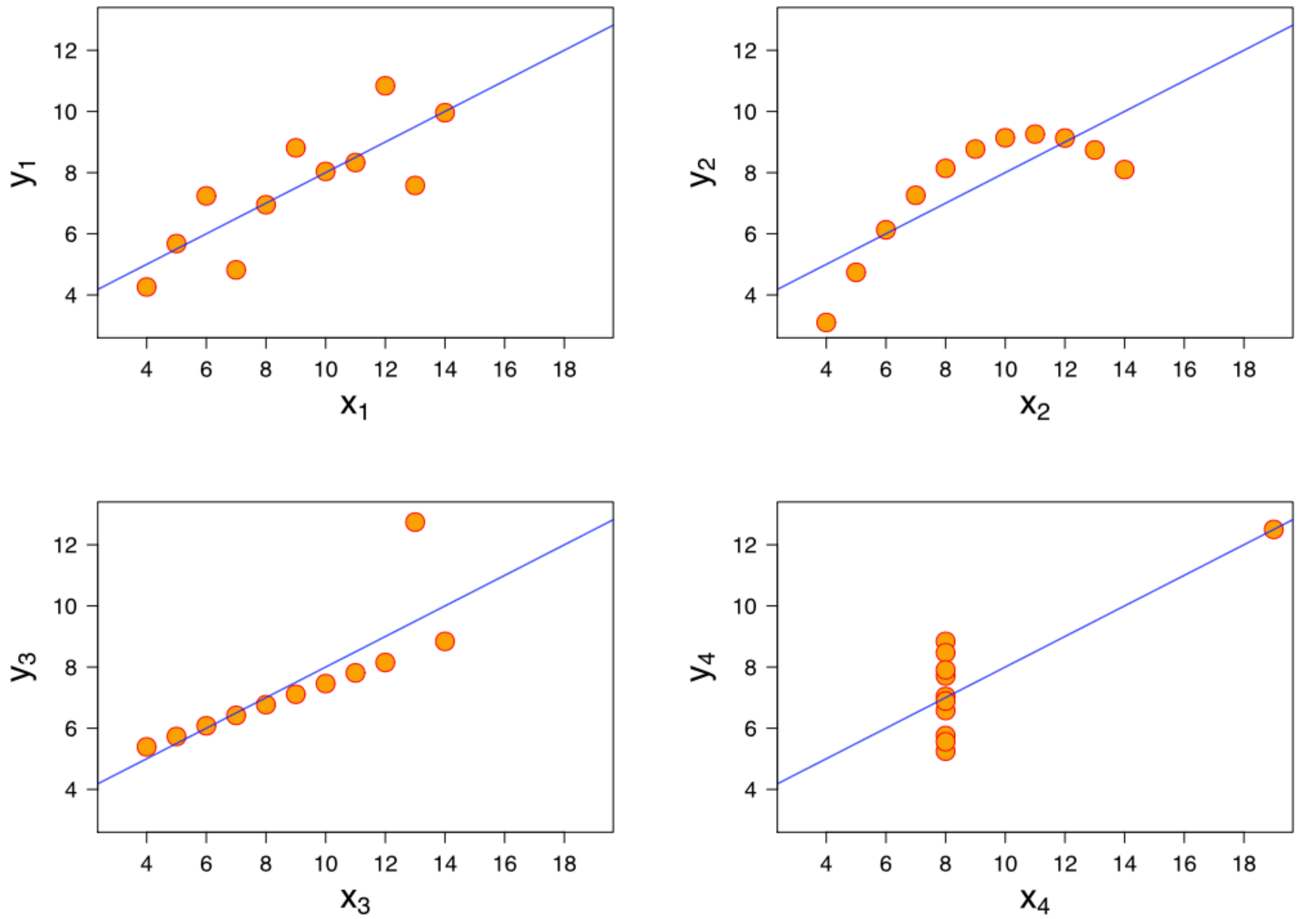
Regression Through Spreadsheets (2)

Spreadsheet “correl()” function

Shortcomings of Simple Linear Regression

Anscombe's Quartet (From Overview)

Also identical R^2 values!



Reminder

Goal: Function f applied to training data should produce values as close as possible in aggregate to actual outputs

Training data

$w_1, x_1, y_1, z_1 \rightarrow o_1$
$w_2, x_2, y_2, z_2 \rightarrow o_2$
$w_3, x_3, y_3, z_3 \rightarrow o_3$
……

Model

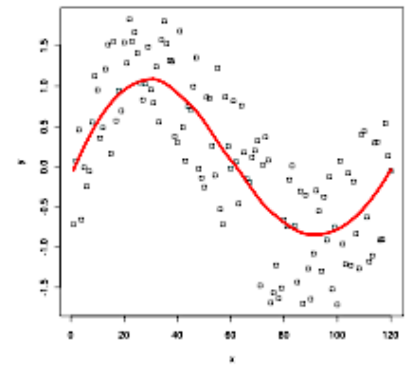
$f(w, x, y, z) = o$

$f(w_1, x_1, y_1, z_1) = o_1'$
$f(w_2, x_2, y_2, z_2) = o_2'$
$f(w_3, x_3, y_3, z_3) = o_3'$

Polynomial Regression

Given: Set of known (x,y) points

Find: function f that "best fits" the known points, i.e., $f(x)$ is close to y



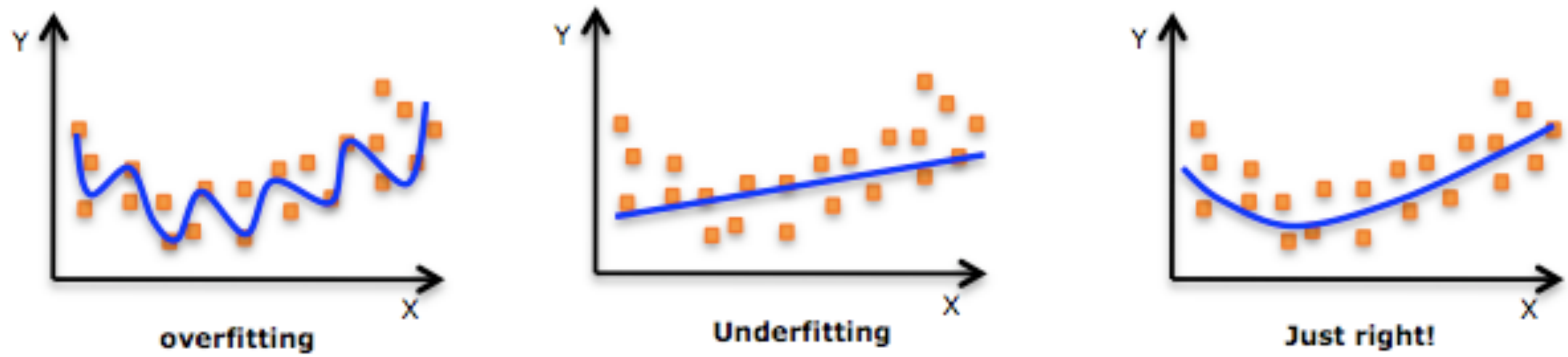
$$f(x) = a_0 + a_1 x + a_2 x^2 + \ldots + a_n x^n$$

- "Best fit" is still method of least squares
- Still have coefficient of determination R^2 (no r)
- Pick smallest degree ***n*** that fits the points reasonably well

Also exponential regression: $f(x) = a\, b^x$

Dangers of (Polynomial) Regression

Overfitting and Underfitting (From Overview)



Anscombe's Quartet in Action

Regression Summary

- Supervised machine learning
- Training data:

  Set of input values with numeric output value
- Model is function from inputs to output

  Use function to predict output value for inputs
- Balance complexity of function against "best fit"
- Also useful for quantifying correlation

  For linear functions, the closer the function fits the points, the more correlated the measures are

Machine Learning - Regression

Stanford University

Google

Instabase

Google Cloud Platform

Association for Computing Machinery

Very Large Data Bases Endowment Inc.

amazon web services