CARLSON SCHOOL OF MANAGEMENT
UNIVERSITY OF MINNESOTA

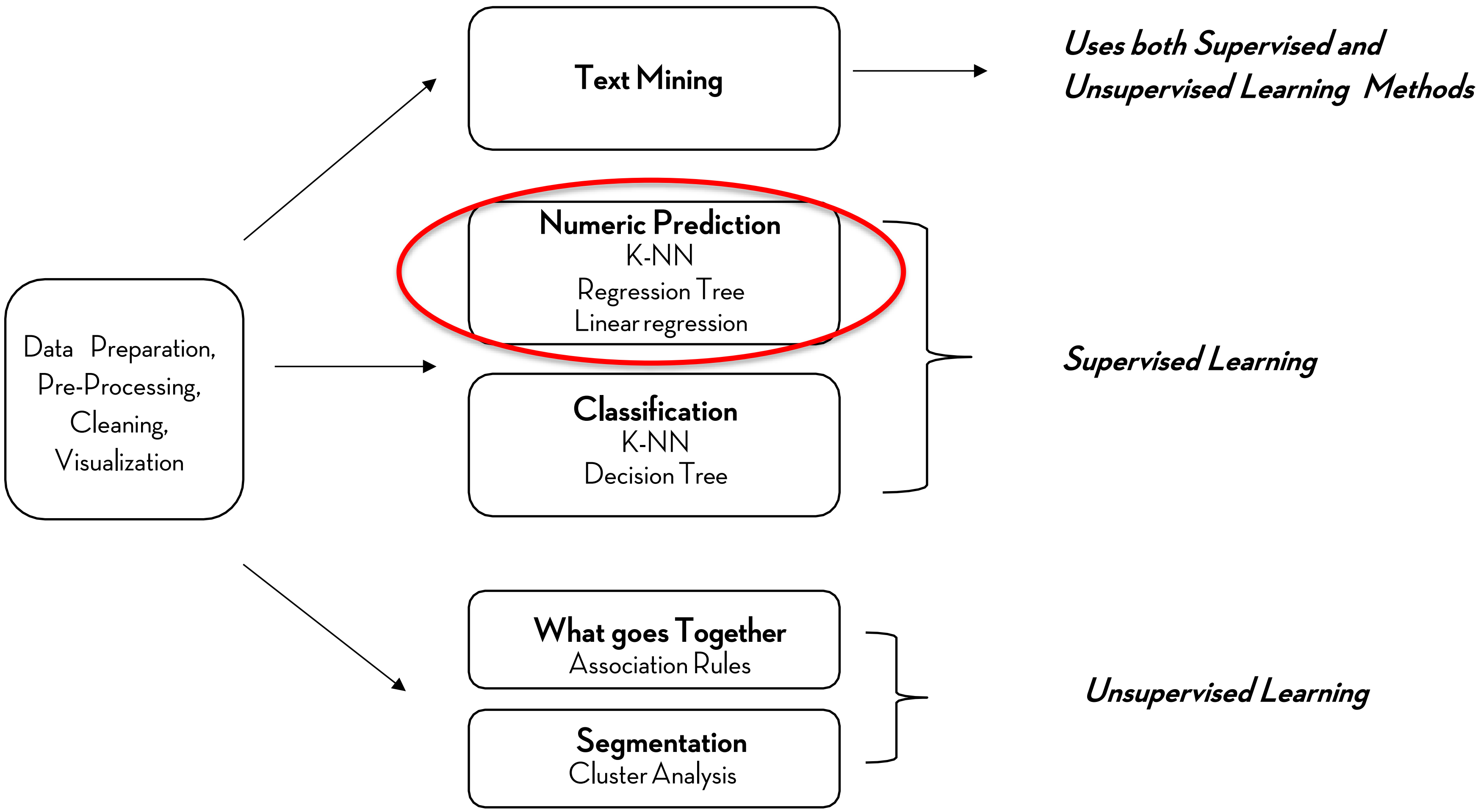
IDSC 4444 (004)
Predictive Analytics:
Numeric Prediction

Zihong Huang
Information & Decision Sciences
Carlson School of Management
huan0707@umn.edu

Agenda

- **Methods**
  - k-NN for numeric prediction
  - Regression Trees
  - Linear Regression
- **Evaluating metrics**
  - Average Error
  - MAE
  - MAPE
  - RMSE
  - Total SSE

An Overview


Text Mining
Uses both Supervised and Unsupervised Learning Methods
Data Preparation, Pre-Processing, Cleaning, Visualization
Numeric Prediction
K-NN
Regression Tree
Linear regression
Classification
K-NN
Decision Tree
Supervised Learning
What goes Together
Association Rules
Segmentation
Cluster Analysis
Unsupervised Learning

Numeric Prediction

- Predicting a **numeric outcome variable** (instead of a categorical outcome variable)
- Some classification techniques can be naturally extended to numeric prediction:
  - k-NN (Nearest-Neighbors)
  - Regression Trees (In Classification, called Decision Trees)
- Others are more tailored for numeric prediction:
  - E.g., Linear Regression

Classification or Numeric?

- Sample Prediction Tasks:
  - **Classification:**
    - ✓ Will a customer buy a given product?
    - ✓ Will this team win the game?
  - **Numeric Prediction:**
    - ✓ How much will a customer buy/spend?
    - ✓ How much will each team score?

k-NN for Numeric Prediction

- Process almost equivalent to what we have seen last time
- For a given observation, identify the k nearest-neighbors : k is chosen to minimize RMS errors (more details later)
- Then, we use **average outcome values of the nearest neighbors as the prediction**
- Can also be a weighted average, weight decreasing with distance:
  - More importance to the closet points

	(Age)	(Income)	(Gender)	Y Amount spent
A	25	55000	M	$ 5
B	32	120000	M	$ 25
C	43	150000	F	$ 50

Labeled Data:
We know the Y

New observation to predict

D	40	130000	F	???

Euclidian Distance

d(D, A)	2.8489
d(D, B)	1.9545
d(D, C)	0.5285

K = 1, ? Prediction: $ 50

K = 3, ? Prediction: (5+25+50)/3 = $ 26.67

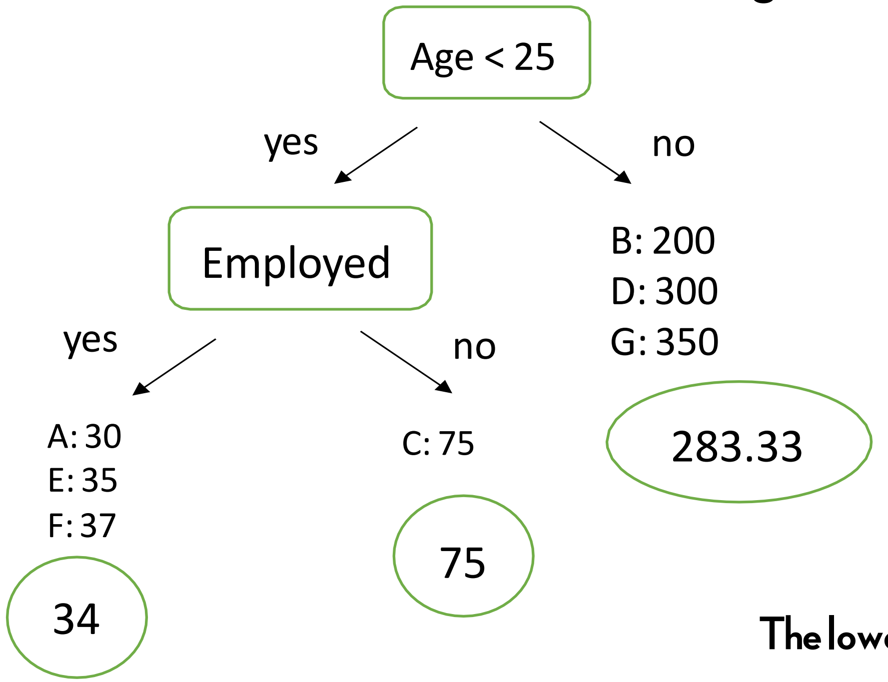
k-NN: Pros and Cons

- Both for categorical and numerical outcomes
- **Pros:**
  - Simple method, effective at capturing complex relationship without really building a model
  - Makes no assumption about data distribution
- **Cons:**
  - If you have a lot of attributes (lots of columns), you need a lot of observations (rows) to make reasonable predictions -> Curse of dimensionality
  - Slower learner, not good for “real-time” or timely predictions.

Regression Trees

- Trees for numeric (continuous) outcome variables: Regression Trees
- Follows the same Recursive Partitioning (RPART) procedure
- At each leaf (final) node, we use **average of outcome values** of data points in that leaf node
- "Purity" is measured using the **sum of squared deviations (SSD)** from the average outcome value at that node

Ind.	Age	Employed	Credit-rating	Loan Amount
A	23	Yes	Fair	$ 30
B	28	No	Excellent	$ 200
C	22	No	Fair	$ 75
D	35	No	Fair	$ 300
E	21	Yes	Fair	$ 35
F	22	Yes	Fair	$ 37
G	33	No	Excellent	$ 350



Purity of "Employed-Yes":

A: $(30-34)^2 = 16$
E: $(35-34)^2 = 1$
F: $(37-34)^2 = 9$

SSD = 26

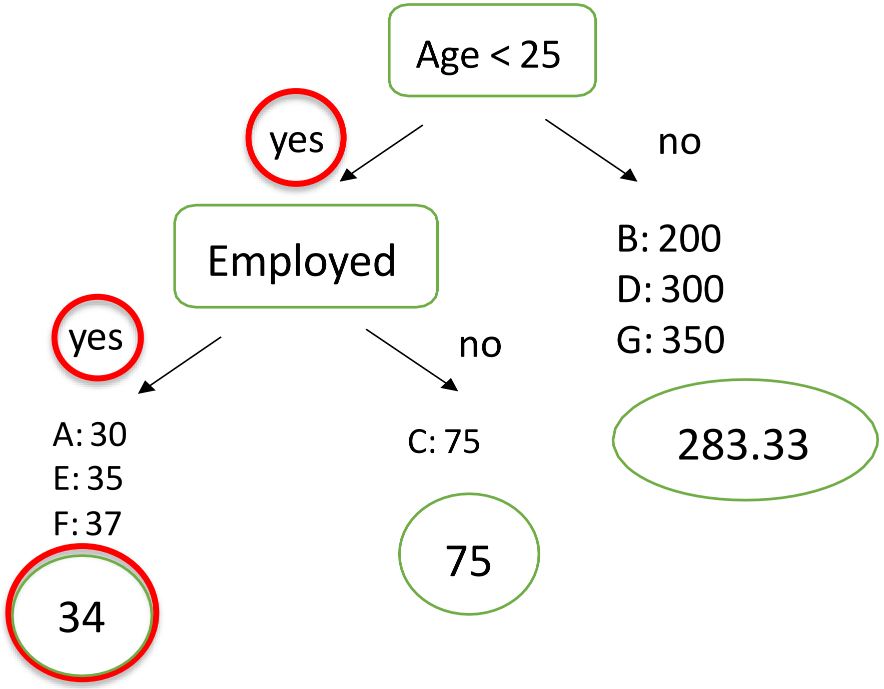
The lower the SSD, the higher the purity

Regression Trees

- To predict the outcome of a new record, the attributes are tested against the regression tree
- Example:

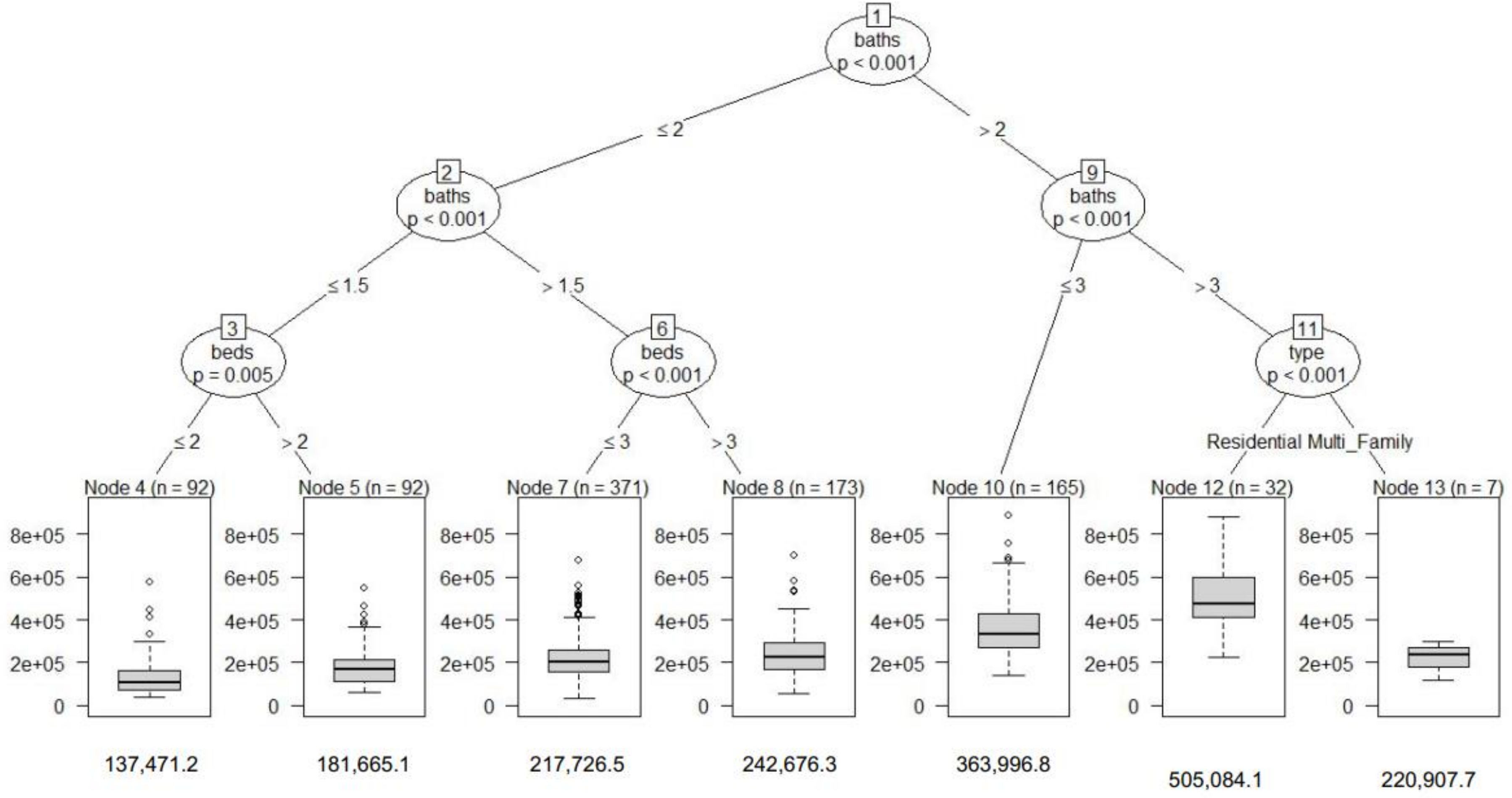
Ind.	Age	Employed	Credit-rating	Loan Amount
ZZ	23	Yes	Fair	??

- Test on Age: Age < 25: Yes
- Test on Employed: Yes
- Reach Leaf node:
  - Customer ZZ prediction: **34**

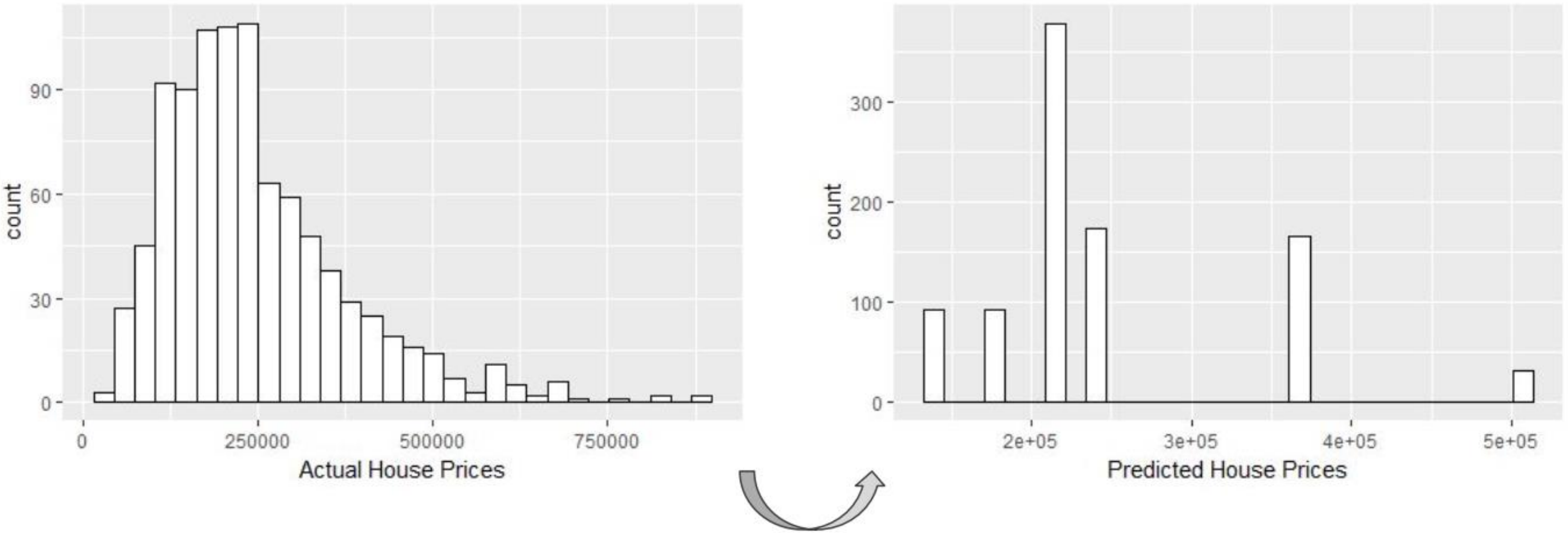


Example: Regression Trees

- **Predicting Minneapolis Housing Prices:** Price ~ Beds + Baths + Property Type



Regression Trees Discretize Outcomes



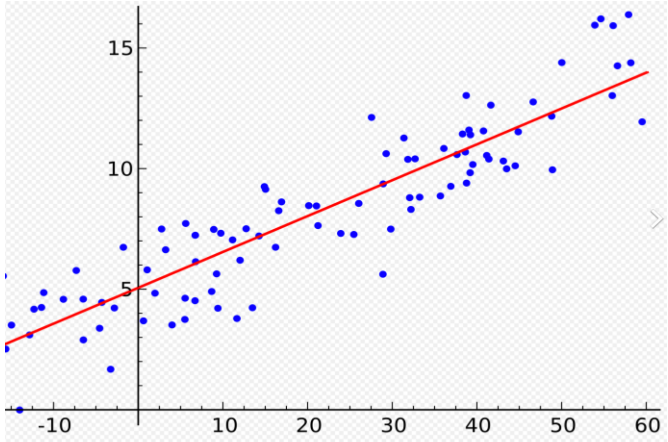
Note how the scales of the two histograms differ!

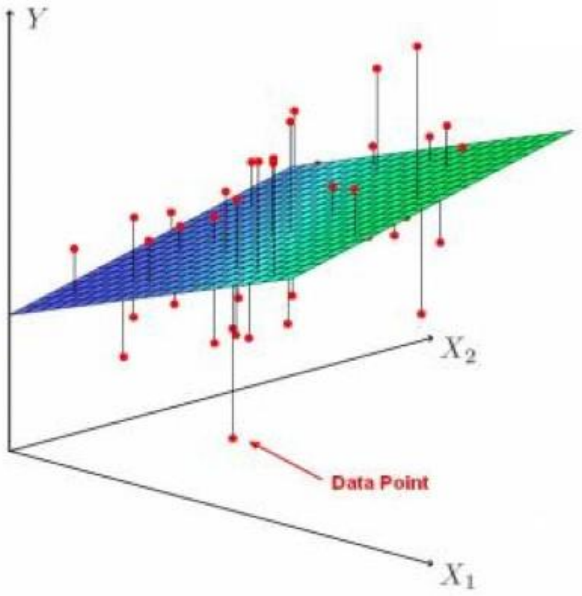
Regression Trees: Pros and Cons

- **Pros:**
  - No parametric assumptions, <u>no need to normalize the data before</u>
  - Good for variable selection: the tree selects what are supposed to be the most relevant attributes
  - Robust to outliers
- **Cons:**
  - Unstable: slightly change in the data can lead to very different splits
  - Since the splits are done with one attribute one time, it can miss interesting relationship between the predictors

Linear Regression

- **Simple Linear Regression**
  - $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$
    - ✓ β: coefficients
    - ✓ β_0: intercept
    - ✓ β_1: slope of X
    - ✓ ε_i: random noise
  - Equivalent to fitting a line
- **Multiple Linear Regression**
  - $Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + \varepsilon_i$
  - Equivalent to fitting a plane
- Can be used for either **explanatory** or **predictive** tasks
  - Explanatory: Explaining the average effect of attributes on an outcome
  - **Predictive**: Predicting the outcome value for new records.





Linear Regression

- <u>Objective</u>: predict Y by finding (estimating) the values of the parameters β

- **Ordinary Least Squares (OLS):**
  - Find the values of the coefficients β that minimize **the sum of squared residuals**: differences between actual and predicted values of the dependent variable Y
  - The "true" linear relationship between Y and X is

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + \varepsilon_i$$

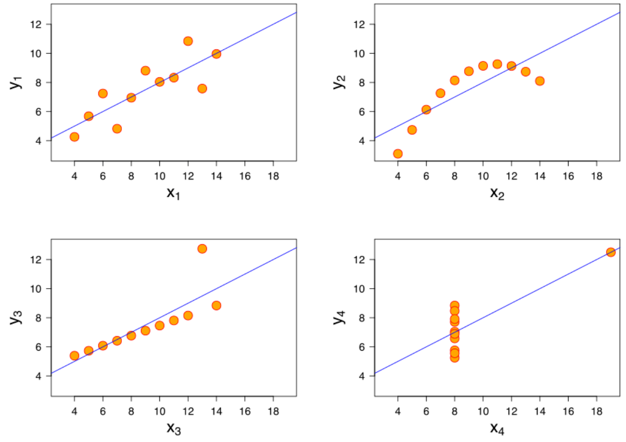
  - The predictions will be based on the "estimated" relationship

$$\widehat{Y_i} = \widehat{\beta_0} + \widehat{\beta_i} X_{i1} + \cdots + \widehat{\beta_i} X_{ip}$$

  - Residual: $Y_i - \widehat{Y_i}$

Assumption Checking before Linear Regression

- OLS produces "good" predictions, if the following assumptions hold:
  - The relationship between Y and Xs is linear
    - ✓ Use scatterplots to check this
  - The observations (records) are independent from each other
  - The noise ε_i, (or, equivalently, Y) follows a normal distribution
    - ✓ Use a histogram to check this
  - The Xs should not be too "collinear"
    - ✓ We should NOT be able to linearly predict one attribute from the other attributes
    - ✓ Use a correlation matrix to check this



Evaluating Performance

- This is a key difference between numeric prediction and classification
- For numeric prediction: the **prediction error** is defined as the difference between predicted outcome and actual outcome
- Several performance metrics, defined based on prediction error:
  - **AE:** Average Error
  - **MAE:** Mean Absolute Error
  - **MAPE:** Mean Absolute Percentage Error
  - **RMSE:** Root Mean Squared Error
  - **Total SSE:** Total Sum of Squared Error

Prediction Error

Actual (a_i)	Predicted (p_i)	Error ($e_i = p_i - a_i$)
125	140	15 (overprediction)
300	225	-75 (underprediction)
75	75	0
450	250	-200 (underprediction)

- For each observation, a_i is the actual (correct) value; p_i is the predicted value
- **Prediction Error = predicted value – actual value**
- If predicted value >actual value: overprediction
- If predicted value <actual value: underprediction

Average Error (AE)

- Average Error (Mean Error): $\sum_{i=1}^{n} \frac{e_i}{n}$

Actual (a_i)	Predicted (p_i)	Error ($e_i = p_i - a_i$)
125	140	15 (overprediction)
300	225	-75 (underprediction)
75	75	0
450	250	-200 (underprediction)
TOTAL		**-260**
AE		**-260/4 = -65**

- It gives an indication of whether we are under-predicting or over-predicting

Mean Absolute Error (MAE)

- Mean Absolute Error (MAE): $\sum_{i=1}^{n} \frac{|e_i|}{n}$

Actual (a_i)	Predicted (p_i)	Error ($e_i = p_i - a_i$)	|Error|
125	140	15	15
300	225	-75	75
75	75	0	0
450	250	-200	200
TOTAL			**290**
MAE			**290/4 = 72.5**

- MAE gives the magnitude of the average absolute error (in any direction)
  - The direction of the errors (that is, over-prediction or under-prediction) is lost

Mean Absolute Percentage Error (MAPE)

- Mean Absolute Percentage Error (MAPE): $100 * \sum_{i=1}^{n} \frac{\left|\frac{e_i}{a_i}\right|}{n}$

Actual (a_i)	Predicted (p_i)	Error ($e_i = p_i - a_i$)	|Error|	|Error / a_i|
125	140	15	15	15/125 = 0.12
300	225	-75	75	75/300 = 0.25
75	75	0	0	0/75 = 0
450	250	-200	200	200/450 = 0.44
TOTAL				**0.81**
MAPE				**0.81/4*100 = 20%**

- MAPE is relative to the actual values: gives a percentage score of how much predictions deviate (on average) from the actual values
  - E.g., a result of 20% indicates that, on average, the predictions deviate from the actual values by about 20% (in any direction)

Root Mean Square Error (RMSE)

- Root Mean Square Errors (RMSE): $\sqrt{\sum_{i=1}^{n} \frac{(e_i)^2}{n}}$

Actual (a_i)	Predicted (p_i)	Error ($e_i = p_i - a_i$)	$(\text{Error})^2$
125	140	15	225
300	225	-75	5625
75	75	0	0
450	250	-200	40000
TOTAL			**45850**
Avg			**11462.5**
RMSE			**107.1**

- RMSE penalizes larger errors. RMSE should be more useful when large errors are particularly undesirable.
- It has the same units of the outcome variable, e.g., If your Y is in $, interpret the RMSE in terms of $ amounts

Total Sum of Squares (TSS)

- Total Sum of Squares (TSS): $\sum_{i=1}^{n}(e_i)^2$

Actual (a_i)	Predicted (p_i)	Error ($e_i = p_i - a_i$)	$(\text{Error})^2$
125	140	15 (overprediction)	225
300	225	-75 (underprediction)	5625
75	75	0	0
450	250	-200 (underprediction)	40000
TOTAL $(\text{Error})^2$			**45850**

- Useful in measuring variability of predictions

Questions?