Gilt ML Course

- Pádraig Cunningham
  - UCD Computer Science
  - Insight Data Analytics Centre
- Content
  - Regression
    - Linear Regression
    - Multivariate Regression
    - Logistic Regression
  - Predictive Analytics
    - Evaluation
    - Other Classification Algorithms
  - Recommender Systems
- Data files
    - http://www.csi.ucd.ie/content/gilt-ml-course
    - http://goo.gl/bm6cd6

Regression

- Linear Regression
  - In Excel, In Weka
- Multiple Regression
- Logistic Regression
  - Odds, Log Odds
  - In Weka

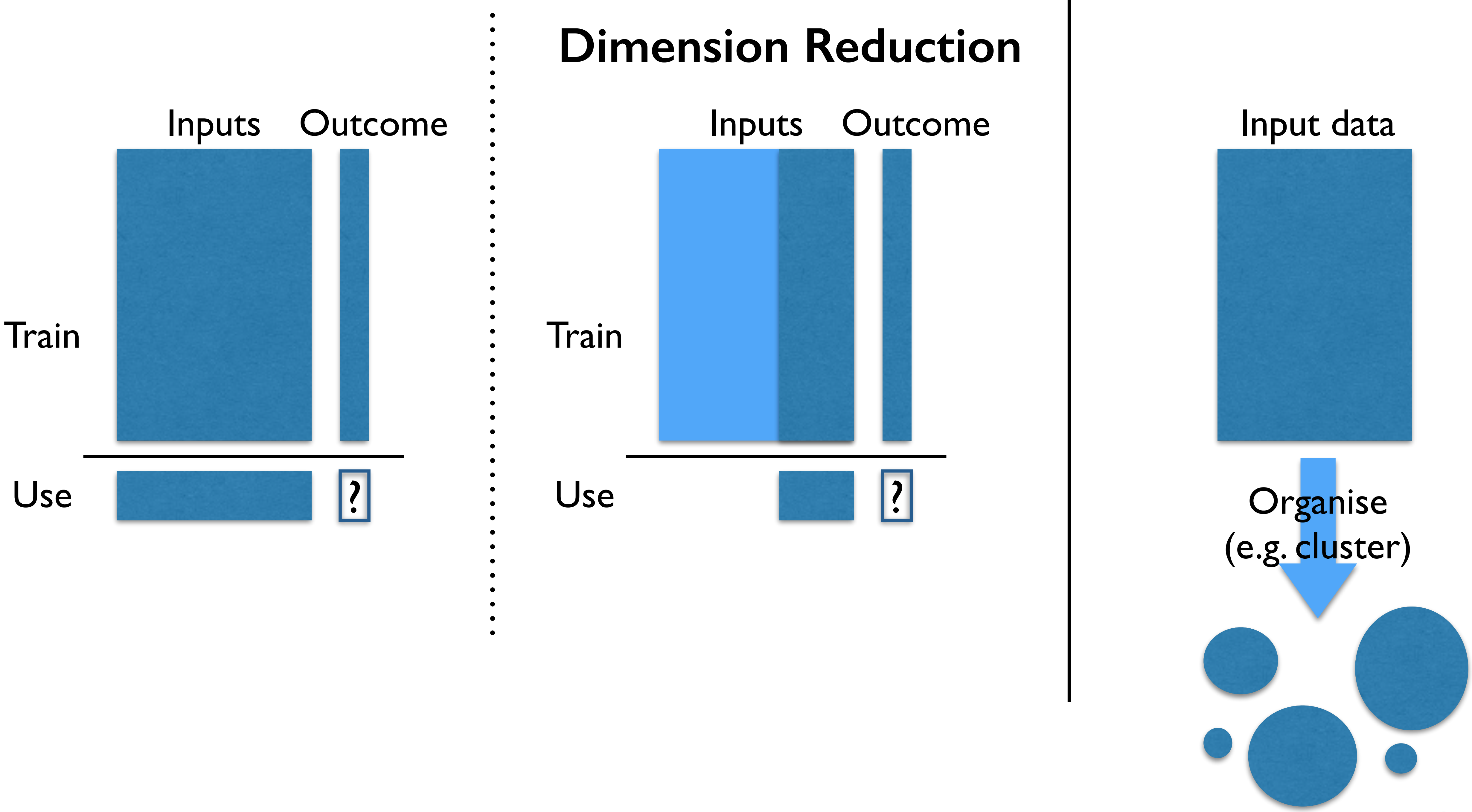
Supervised Machine Learning

- aka: Predictive Analytics
- Regression or Classification?
- Regression
- Logistic Regression (actually classification)
- Other Classification Models
  - Naive Bayes
  - Neural Networks
  - Support Vector Machines

First: Task Categories

Supervised Learning

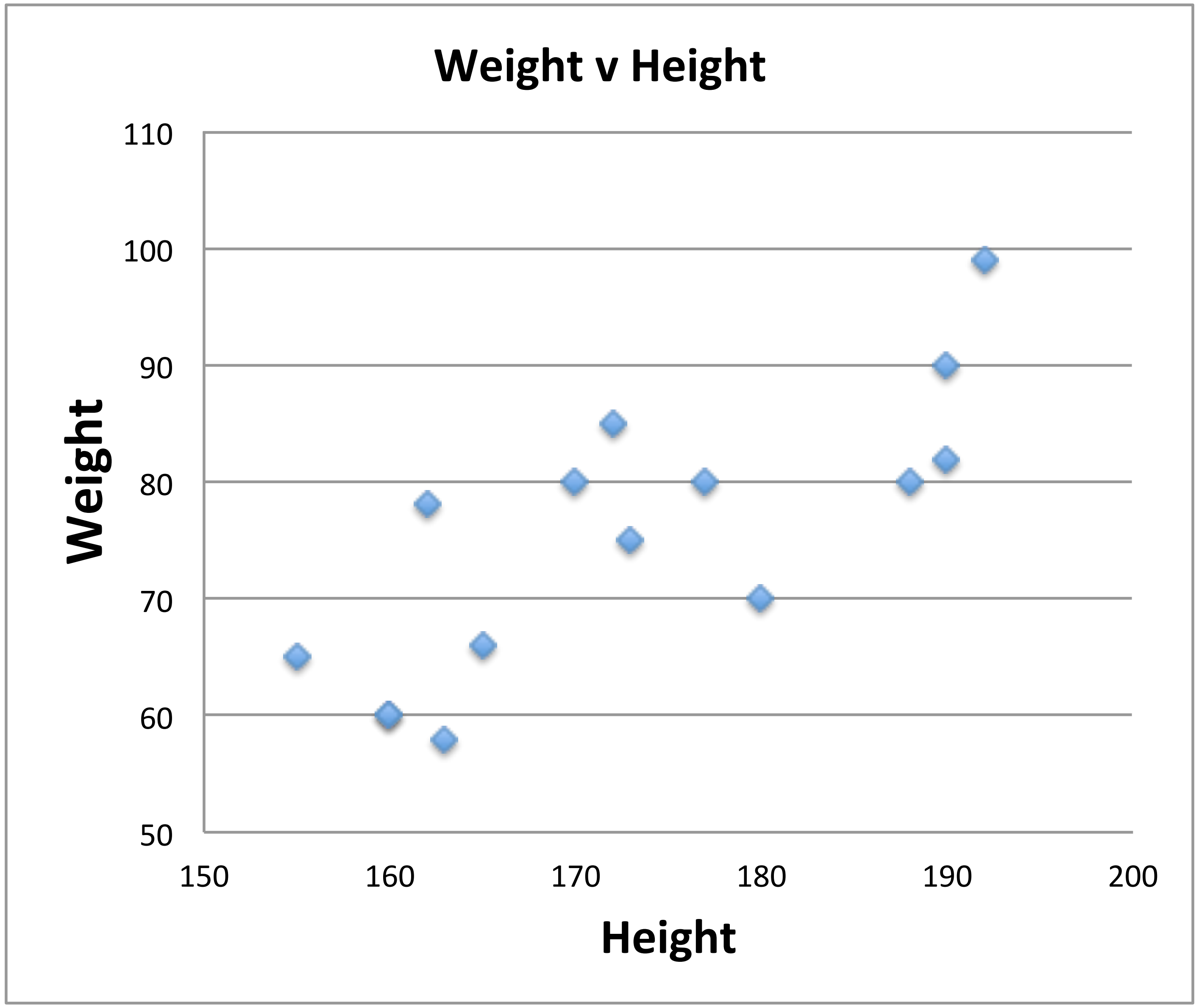
Unsupervised



Insight

Simple Regression

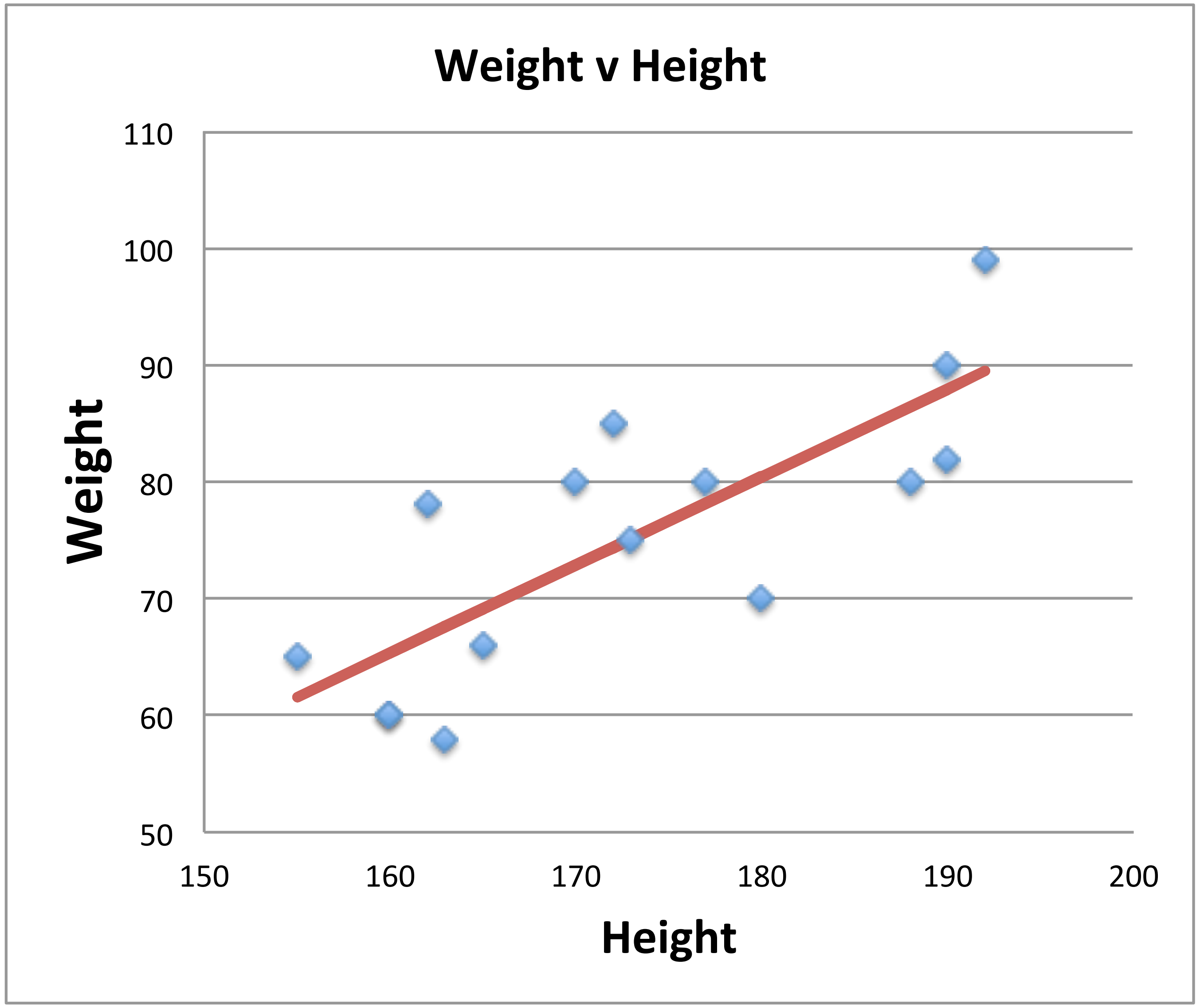
Height	Weight
173	75
160	60
190	82
192	99
162	78
165	66
155	65
163	58
170	80
172	85
180	70
160	60
188	80
190	90
177	80



Predict Weight, given only Height

Simple Regression

Height	Weight	Pred-Weight
173	75	75.1
160	60	65.32
190	82	87.88
192	99	89.39
162	78	66.83
165	66	69.08
155	65	61.56
163	58	67.58
170	80	72.84
172	85	74.35
180	70	80.36
160	60	65.32
188	80	86.38
190	90	87.88
177	80	78.11

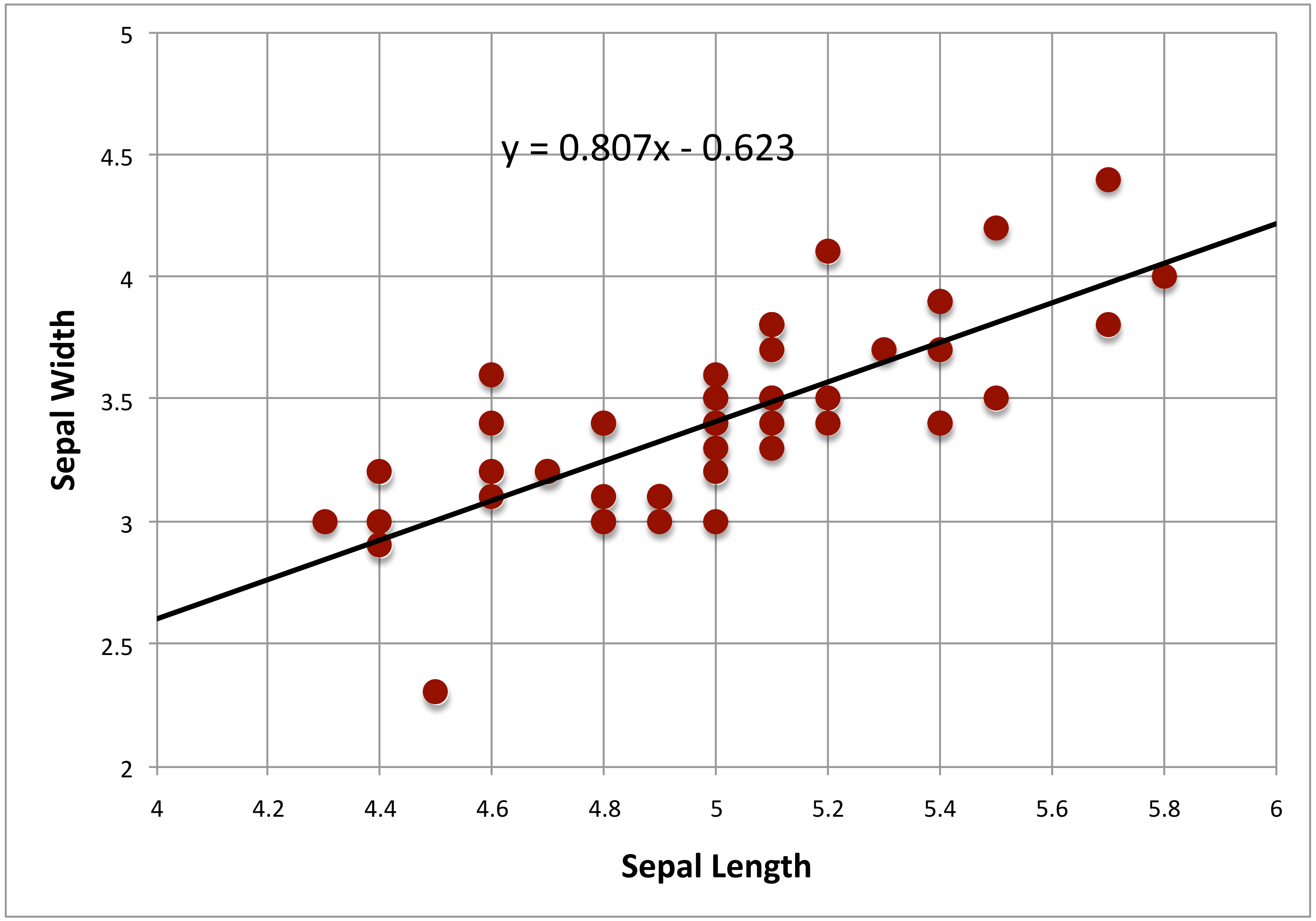


Weight = 0.75 x Height - 54.99

Linear Regression

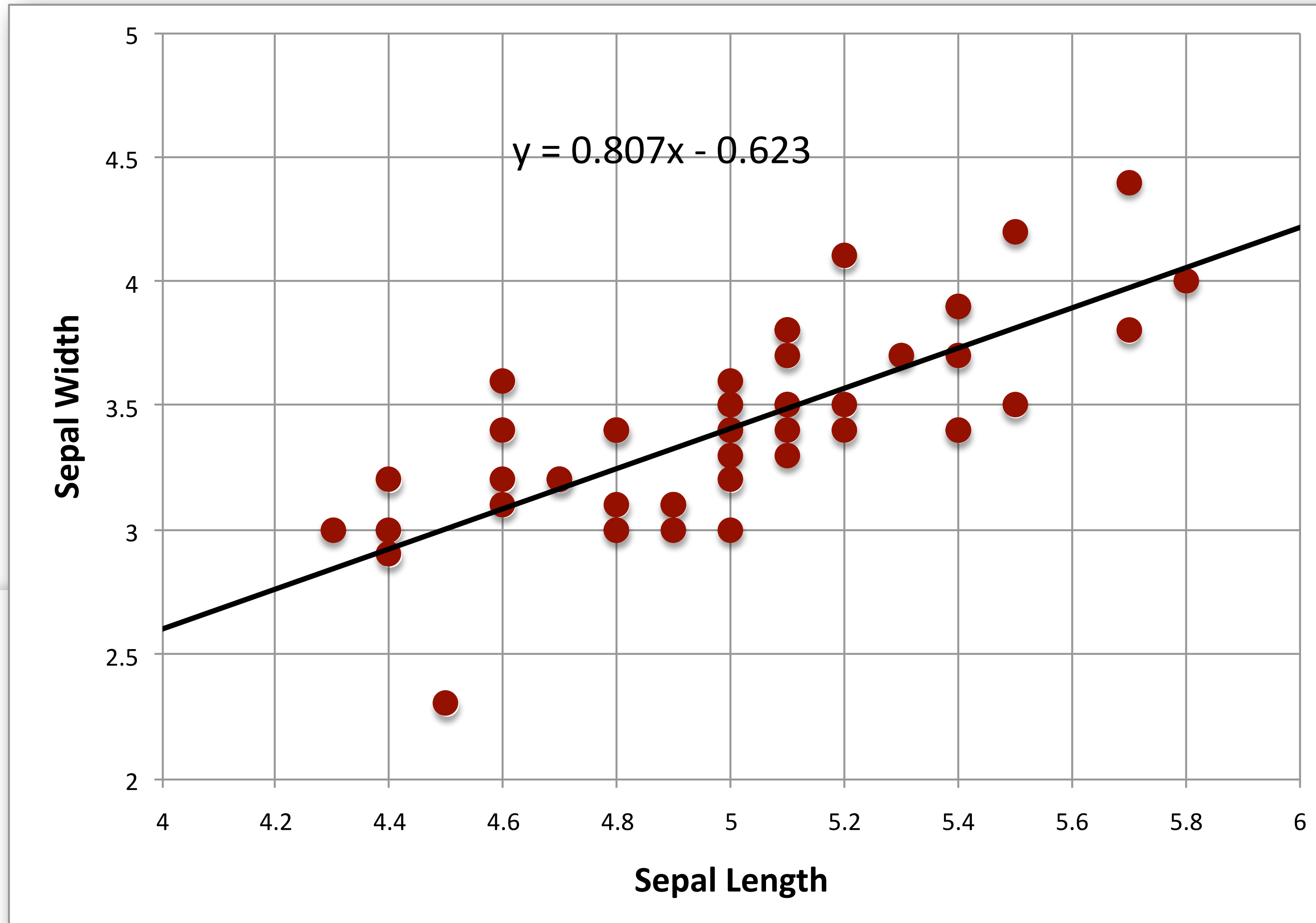
- Iris Setosa from Iris data set
  - Sepal Width
  - Sepal Length
- Model

$$y = \beta_0 + \beta_1 x$$



Linear Regression

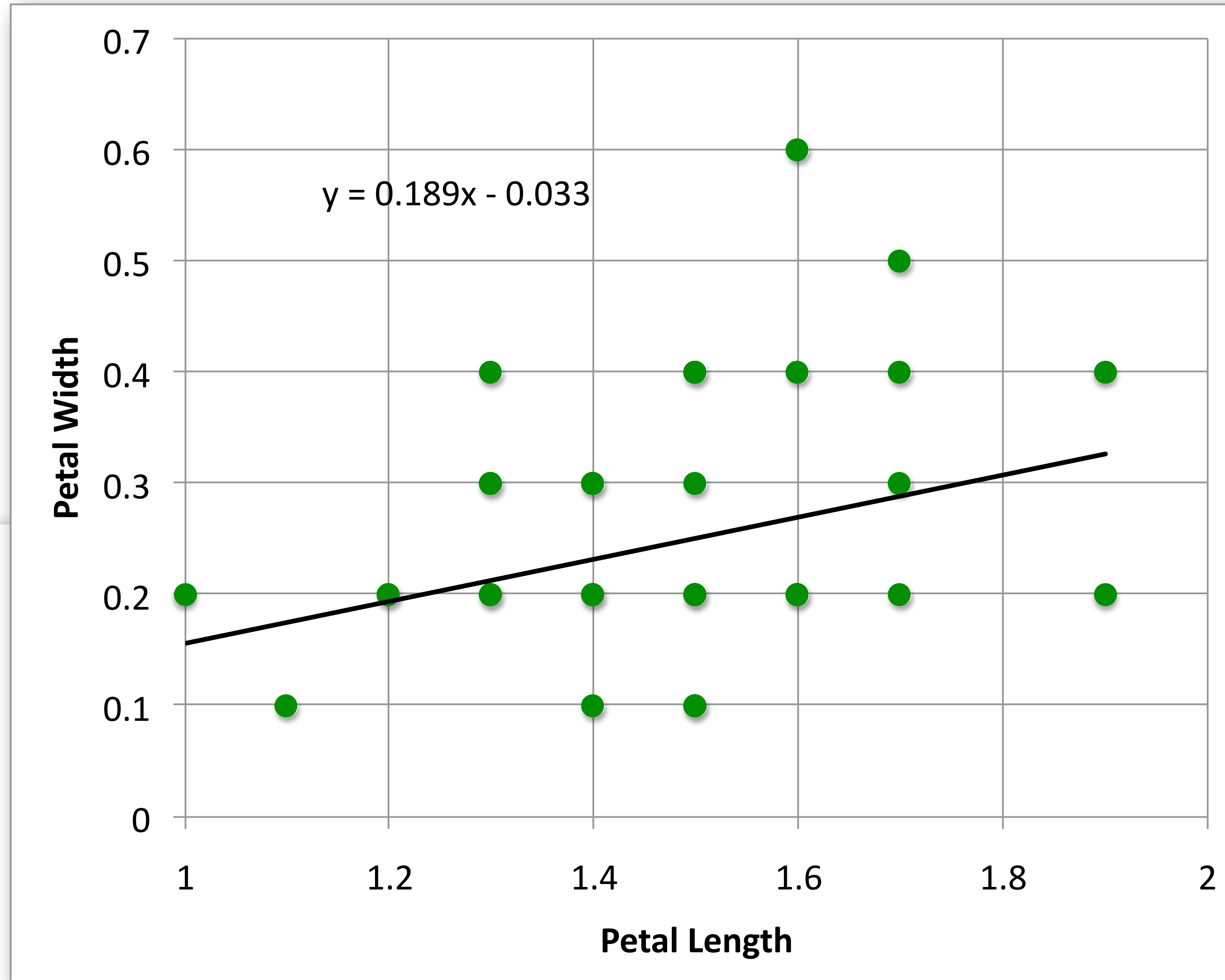
- In Weka
  - Load iris-setosa.csv
  - Remove last 3 atts.




```
Linear Regression Model

sepalwidth =

      0.8072 * sepallength +
     -0.623

Time taken to build model: 0.05 seconds

=== Evaluation on training set ===
=== Summary ===

Correlation coefficient                  0.7468
Mean absolute error                      0.1995
Root mean squared error                  0.2509
Relative absolute error                 69.0675 %
Root relative squared error             66.5071 %
Total Number of Instances               50
```

Linear Regression

- Estimate Petal Width
  - from Petal Length




```
Linear Regression Model

petalwidth =

      0.1893 * petallength +
     -0.0331

Time taken to build model: 0 seconds

=== Evaluation on training set ===
=== Summary ===

Correlation coefficient                  0.3063
Mean absolute error                      0.0794
Root mean squared error                  0.101
Relative absolute error                 94.7497 %
Root relative squared error             95.1932 %
Total Number of Instances               50
```

LR: Parameter Estimation

- The parameters can be calculated directly from the data.

$$y = \beta_0 + \beta_1 x$$

$$\beta_1 = \frac{n\Sigma(xy) - \Sigma x \Sigma y}{n\Sigma(x^2) - (\Sigma x)^2}$$

$$\beta_0 = \frac{\Sigma y - m\Sigma x}{n}$$

- see: http://www.csi.ucd.ie/files/iris-setosa.xls

Multiple Regression

- Estimate Petal Width
  - from Petal Length, Sepal Length & Sepal Width

```
Linear Regression Model

petalwidth =

      0.0251 * sepallength +
      0.0488 * sepalwidth +
      0.1569 * petallength +
     -0.2782

Time taken to build model: 0 seconds

=== Evaluation on training set ===
=== Summary ===

Correlation coefficient                  0.3865
Mean absolute error                      0.0762
Root mean squared error                  0.0979
Relative absolute error                 90.8572 %
Root relative squared error             92.2296 %
Total Number of Instances               50
```

```
Linear Regression Model

petalwidth =

      0.1893 * petallength +
     -0.0331

Time taken to build model: 0 seconds

=== Evaluation on training set ===
=== Summary ===

Correlation coefficient                  0.3063
Mean absolute error                      0.0794
Root mean squared error                  0.101
Relative absolute error                 94.7497 %
Root relative squared error             95.1932 %
Total Number of Instances               50
```

Multiple Regression

	A	B	C	D	E
72	Sepal-len	Sepal-width	Petal-len	Petal-width	
73	5.1	3.5	1.4	0.2	Iris-setosa
74	4.9	3	1.4	0.2	Iris-setosa
75	4.7	3.2	1.3	0.2	Iris-setosa
76	4.6	3.1	1.5	0.2	Iris-setosa
77	5	3.6	1.		
78	5.4	3.9	1.		
79	4.6	3.4	1.		
80	5	3.4	1.		
81	4.4	2.9	1.		
82	4.9	3.1	1.		
83	5.4	3.7	1.		

```
Linear Regression Model

petalwidth =

      0.0251 * sepallength +
      0.0488 * sepalwidth +
      0.1569 * petallength +
     -0.2782

Time taken to build model: 0 seconds

=== Evaluation on training set ===

icient                          0.3865
or                              0.0762
 error                          0.0979
 error                         90.8572 %
ared error                     92.2296 %
nstances                       50
```

- In Excel
  - =LINEST(D73:D122,A73:C122,TRUE,TRUE)

0.157	0.049	0.025	-0.278

Insight

Multiple Regression

- Allow attribute selection

```
Linear Regression Model

petalwidth =

      0.0251 * sepallength +
      0.0488 * sepalwidth +
      0.1569 * petallength +
     -0.2782

Time taken to build model: 0 seconds

=== Evaluation on training set ===
=== Summary ===

Correlation coefficient                  0.3865
Mean absolute error                      0.0762
Root mean squared error                  0.0979
Relative absolute error                 90.8572 %
Root relative squared error             92.2296 %
Total Number of Instances               50
```

```
Linear Regression Model

petalwidth =

      0.0656 * sepalwidth +
      0.1638 * petallength +
     -0.22

Time taken to build model: 0 seconds

=== Evaluation on training set ===
=== Summary ===

Correlation coefficient                  0.3827
Mean absolute error                      0.0752
Root mean squared error                  0.0981
Relative absolute error                 89.6547 %
Root relative squared error             92.3863 %
Total Number of Instances               50
```

BikeShare Data

- season : season (1:spring, 2:summer, 3:fall, 4:winter)
- yr : year (0: 2011, 1:2012)
- mnth : month (1 to 12)
- holiday : whether day is holiday or not
- weekday : day of the week
- workingday : if day is neither weekend nor holiday is 1, otherwise is 0.

+ weathersit :
    - 1: Clear
    - 2: Mist
    - 3: Rain

- temp : Normalized temperature
- atemp: Normalized feeling temperature
- hum: Normalized humidity
- windspeed: Normalized wind speed

Outcomes

- casual: count of casual users
- registered: count of registered users
- cnt: both casual and registered

BikeShare Model

Linear Regression Model

registered =

447	* season +
1754	* yr +
-23	* mnth +
-244	* holiday +
42	* weekday +
948	* workingday +
-497	* weathersit +
834	* temp +
2678	* atemp +
-625	* hum +
-1695	* windspeed +
768	

casual =

61	* season +
286	* yr +
-15	* mnth +
-274	* holiday +
26	* weekday +
-828	* workingday +
-113	* weathersit +
1194	* temp +
894	* atemp +
-393	* hum +
-862	* windspeed +
700	

For other examples see: http://logisticregressionanalysis.com/

Odds and Probabilities

- What is the probability of throwing a 6?
  - 1/6 = 0.16667
- What are the odds of throwing a 6?
  - 1:5
  - 1 to 5
  - 1/5 = 0.2
  - 0.2 is the monetary stake required to win 1 unit in a wager with fair odds.

$$odds = \frac{prob}{1 - prob}$$

$$prob = \frac{odds}{1 + odds}$$

Logistic Regression

- In linear regression the dependent variable is a numeric value

$$y = \beta_0 + \beta_1 x$$

- In logistic regression the dependent variable is the log odds that an outcome variable is 1.

$$ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x$$

Logistic Regression

- log odds is the dependent variable

$$ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x$$

more generally

$$ln\frac{p}{1-p} = \beta_0 + \sum_{j=1}^{d} \beta_j x_j$$

$$\frac{p}{1-p} = odds = e^{(\beta_0 + \beta_1 x)}$$

$$p = \frac{e^{(\beta_0+\beta_1 x)}}{1+e^{(\beta_0+\beta_1 x)}}$$

given $prob = \frac{odds}{1+odds}$

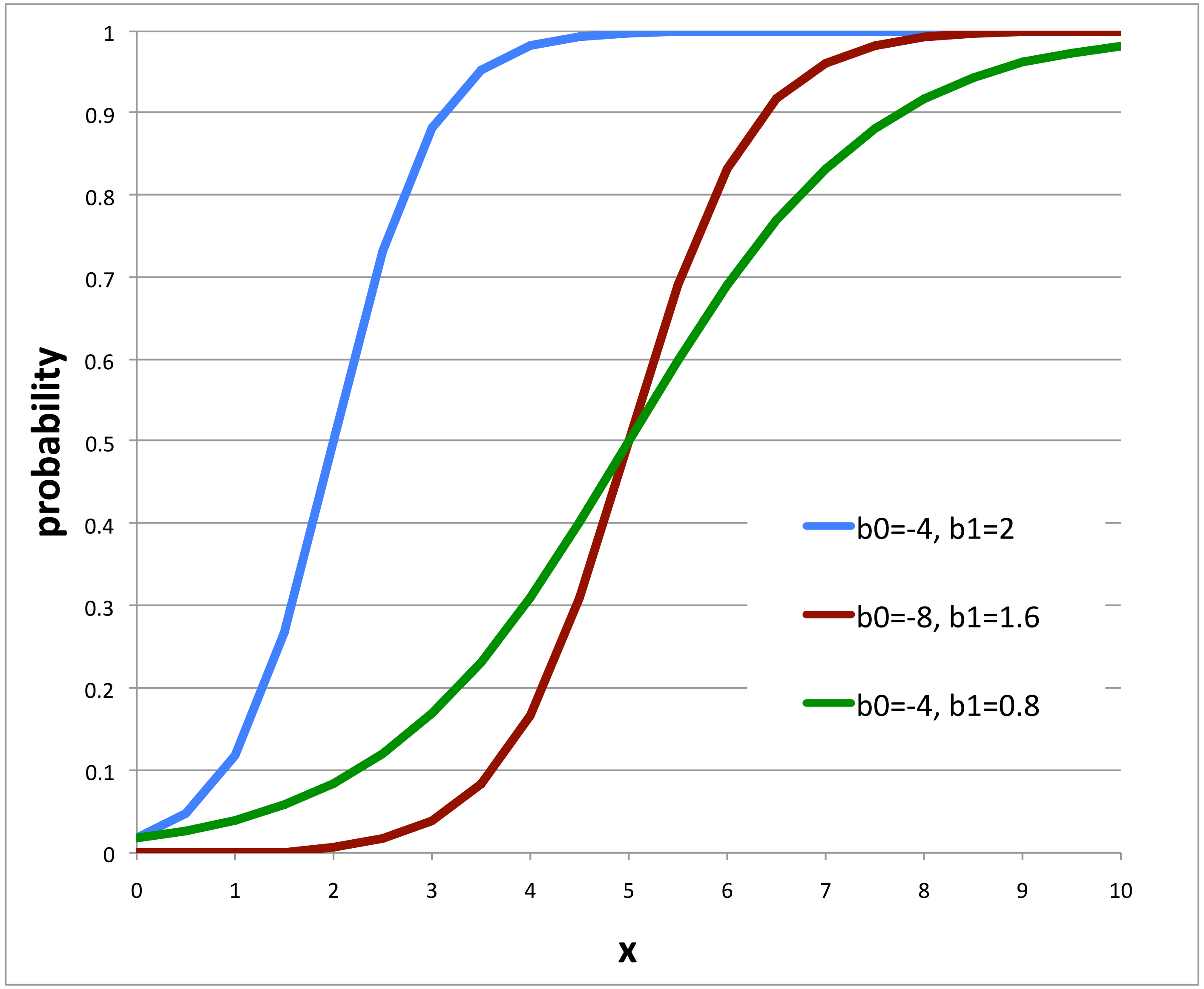
$$p = \frac{1}{1+e^{-(\beta_0+\beta_1 x)}}$$

Insight

Logistic Regression

- Equivalent to a single layer neural network with a sigmoid transfer function

$$p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$



Binge Drinking Example

- 17,096 students, 3,314 binge drinkers
- $p = 0.194, odds = 0.24$
- predictive variable is gender
  - $x = 1$ if man, $x = 0$ if woman
- $p_{men} = 0.227, odds_{men} = 0.294, ln(odds_{men}) = -1.23$
- $p_{women} = 0.17\ odds_{women} = 0.205, ln(odds_{women}) = -1.59$

Taken from text book by Moore and McCabe
http://bcs.whfreeman.com/ips5e/content/cat_080/pdf/moore16.pdf

Binge Drinking Example

- p_{men} = 0.227, $odds_{men}$ = 0.294, $ln(odds_{men})$ = -1.22
- p_{women} = 0.17 $odds_{women}$ = 0.205, $ln(odds_{women})$ = -1.58

$$ln\left(\frac{p_{men}}{1-p_{men}}\right) = \beta_0 + \beta_1 x \qquad ln\left(\frac{p_{women}}{1-p_{women}}\right) = \beta_0$$

$$\beta_0 = -1.58; \beta_1 = 1.58 - 1.22 = 0.36$$

Interpretability: being male adds 0.36 to the log odds

Building (Training) LR Models

- Models can have many *input* variables
- Training is done in an iterative fashion
  - see Iteratively Re-weighted Least Squares
    - http://en.wikipedia.org/wiki/Iteratively_reweighted_least_squares
  - see Charles Elkan tutorial at UCSD
    - http://cseweb.ucsd.edu/~elkan/250B/logreg.pdf

Insight

Regression

- Linear Regression
  - In Excel, In Weka
- Multiple Regression
- Logistic Regression
  - Odds, Log Odds
  - In Weka