CST8390 BUSINESS INTELLIGENCE & DATA ANALYTICS

Week 6

Regression

Agenda

- Linear regression
  - Simple linear regression
  - Multiple linear regression
- Multivariate regression
- Logistic regression

Types of Relationships

- Deterministic (or functional) relationship
- Statistical relationship

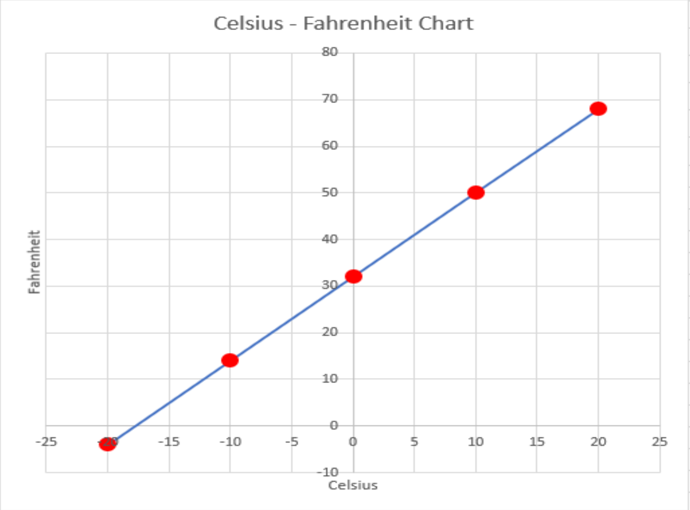
Deterministic (or functional) Relationship

- Ex. Relationship between Celsius and Fahrenheit

  - $F = \frac{9}{5} * C + 32$

The observed (*x*, *y*) data points fall directly on the line.

m	C	b	F
1.8	-20	32	-4
1.8	-10	32	14
1.8	0	32	32
1.8	10	32	50
1.8	20	32	68



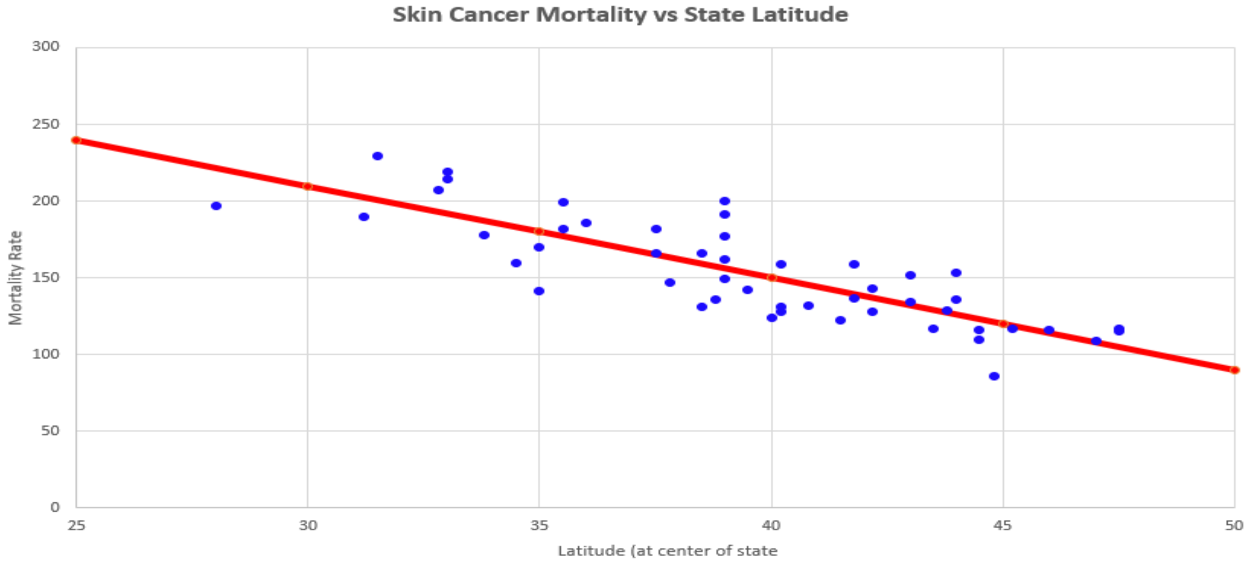
For deterministic relationship, the equation exactly describes the relationship between the two variables.

Statistical Relationship

- Examples
  - Height and weight — as height increases, you'd expect weight to increase, but not perfectly.
  - Alcohol consumed and blood alcohol content — as alcohol consumption increases, you'd expect one's blood alcohol content to increase, but not perfectly.
  - Driving speed and gas mileage — as driving speed increases, you'd expect gas mileage to decrease, but not perfectly.

Statistical Relationship

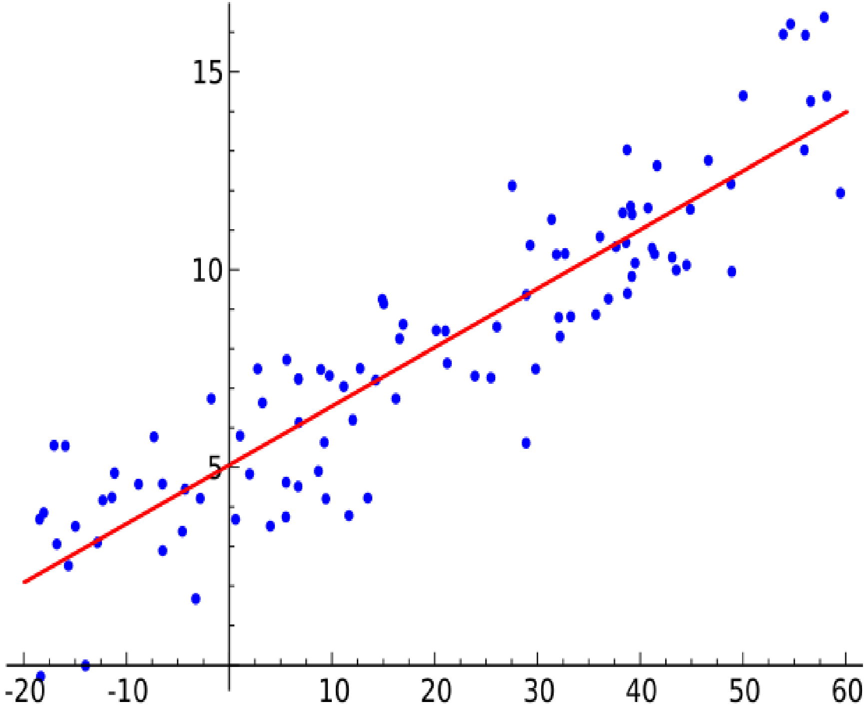
Example: The response variable y is the mortality due to skin cancer (number of deaths per 10 million people) and the predictor variable x is the latitude (degrees North) at the center of 49 states in the U.S.



State	Lat	Mort	Ocean	Long
Alabama	33	219	1	87
Arizona	34.5	160	0	112
Arkansas	35	170	0	92.5
California	37.5	182	1	119.5
Colorado	39	149	0	105.5
Connectic	41.8	159	1	72.8
Delaware	39	200	1	75.5
Wash,D.C	39	177	0	77
Florida	28	197	1	82
Georgia	33	214	1	83.5
Idaho	44.5	116	0	114
Illinois	40	124	0	89.5
Indiana	40.2	128	0	86.2
Iowa	42.2	128	0	93.8
Kansas	38.5	166	0	98.5
Kentucky	37.8	147	0	85
Louisiana	31.2	190	1	91.8
Maine	45.2	117	1	69
Maryland	39	162	1	76.5
Massachu	42.2	143	1	71.8
Michigan	43.5	117	0	84.5
Minnesota	46	116	0	94.5
Mississipp	32.8	207	1	90
Missouri	38.5	131	0	92
Montana	47	109	0	110.5
Nebraska	41.5	122	0	99.5
Nevada	39	191	0	117
NewHamp	43.8	129	1	71.5
NewJerse	40.2	159	1	74.5
NewMexic	35	141	0	106
MewYork	43	152	1	75.5
NorthCaro	35.5	199	1	79.5
NorthDak	47.5	115	0	100.5
Ohio	40.2	131	0	82.8
Oklahoma	35.5	182	0	97.2
Oregon	44	136	1	120.5
Pennsylva	40.8	132	0	77.8
RhodeIsla	41.8	137	1	71.5
SouthCar	33.8	178	1	81
SouthDak	44.8	86	0	100
Tennesse	36	186	0	86.2
Texas	31.5	229	1	98
Utah	39.5	142	0	111.5
Vermont	44	153	1	72.5
Virginia	37.5	166	1	78.5
Washingt	47.5	117	1	121
WestVirgi	38.8	136	0	80.8
Wisconsir	44.5	110	0	90.2
Wyoming	43	134	0	107.5

Regression

- When you have a series of continuous data that follow some sort of pattern.
- determines the strength of the relationship between dependent variable and a series of other changing variables (known as independent variables).



Simple Linear Regression

- Statistical method that allows us to summarize and study relationships between two continuous variables
  - One variable, denoted as x, as the independent (predictor) variable
  - The other variable, denoted as y, as the dependent (response) variable

Parameters for line:

- In mathematics, a line needs two parameters:

$$y = mx + b$$

- *m* is the slope, *b* is the y-intercept
- In regression, the parameters take different names:
- $h(x) = \Theta_0 + \Theta_1 x$
- $h(x)$ is the predicted value for x
- Θ_0 , Θ_1 are the coefficients.

Linear Regression with one variable

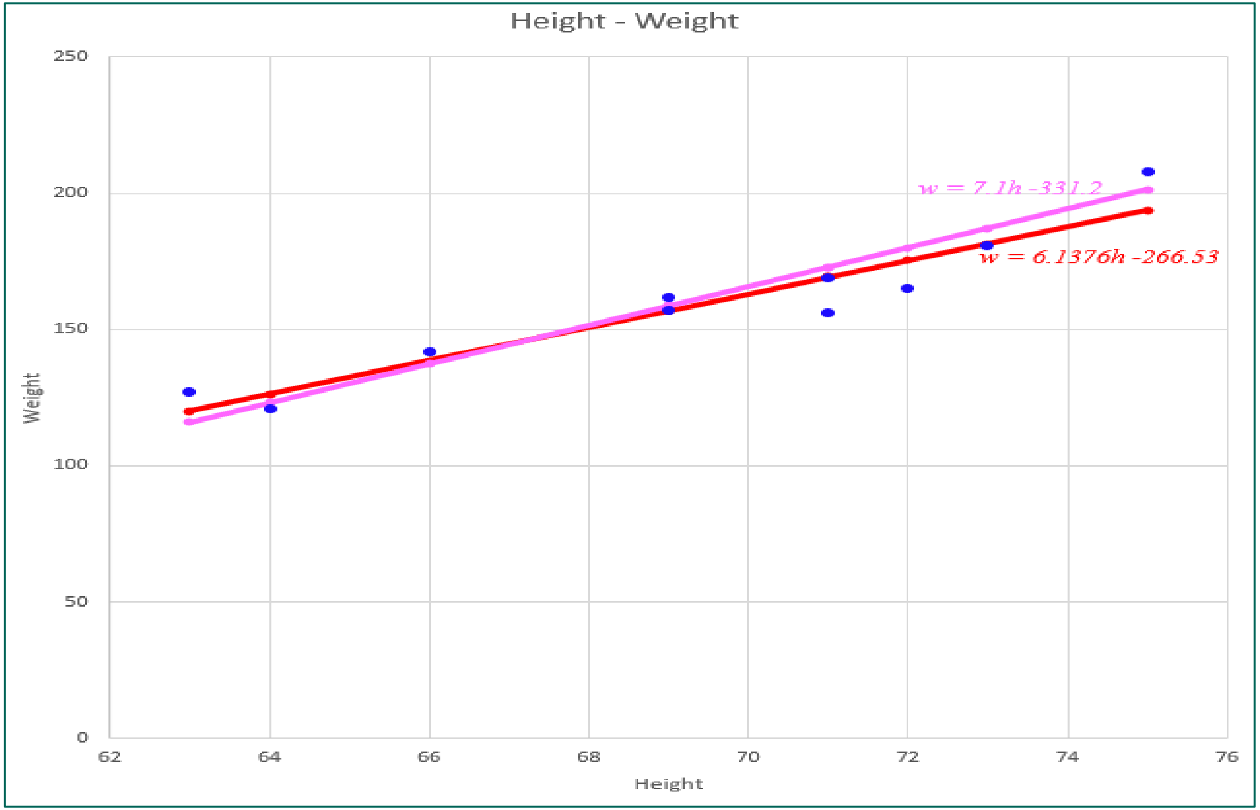
- Try to fit a best-fit line to a data set. This line is then used to predict real values for continuous output.
- Need a training set:
  - x – an input variable
  - y – The output variable.
  - h is a function that maps x -> y
  - $h(x) = \Theta_0 + \Theta_1 x$, or $y = mx + b$
- Also called Univariate linear regression.

Linear Regression with one variable

- To choose the best values of Θ_0 and Θ_1, we use a cost function.
- This calculates the total error between your predicted value, and the actual values. We continue to change the values until we find the minimum error.
- The h function deals with x, where the cost function deals with Θ_1.

Linear Regression - Example

Height	Weight
63	127
64	121
66	142
69	157
69	162
71	156
71	169
72	165
73	181
75	208



Which line (red or pink) is the best fit?

Red line: w = -266.53 + 6.1376h   *Pink line: w = -331.2 + 7.1h*

For the student with the height 63 inches, actual weight is 127 pounds.
Based on the red fitted line, weight is -266.53 + 6.1376 * 63 = **120.1**
Prediction Error = 127 – 120.1 = 6.9

Based on the pink fitted line, weight is -331.2 + 7.1 * 63 = **116.1**
Prediction Error = 127 – 116.1 = 10.9

A line that fits the data "best" will be the one with overall minimal prediction errors.

In order to find the overall prediction error, "least squares criterion" can be used.

Least Squares Criterion

$w = -266.53 + 6.1376h$						
		x	y_i	y_i'	$y_i - y_i'$	$(y_i - y_i')^2$
-266.53	6.1376	63	127	120.1388	6.8612	47.07607
-266.53	6.1376	64	121	126.2764	-5.2764	27.8404
-266.53	6.1376	66	142	138.5516	3.4484	11.89146
-266.53	6.1376	69	157	156.9644	0.0356	0.001267
-266.53	6.1376	69	162	156.9644	5.0356	25.35727
-266.53	6.1376	71	156	169.2396	-13.2396	175.287
-266.53	6.1376	71	169	169.2396	-0.2396	0.057408
-266.53	6.1376	72	165	175.3772	-10.3772	107.6863
-266.53	6.1376	73	181	181.5148	-0.5148	0.265019
-266.53	6.1376	75	208	193.79	14.21	201.9241
					Total	597.3863

$w = -331.2 + 7.1h$						
		x	y_i	y_i'	$y_i - y_i'$	$(y_i - y_i')^2$
-331.2	7.1	63	127	116.1	10.9	118.81
-331.2	7.1	64	121	123.2	-2.2	4.84
-331.2	7.1	66	142	137.4	4.6	21.16
-331.2	7.1	69	157	158.7	-1.7	2.89
-331.2	7.1	69	162	158.7	3.3	10.89
-331.2	7.1	71	156	172.9	-16.9	285.61
-331.2	7.1	71	169	172.9	-3.9	15.21
-331.2	7.1	72	165	180	-15	225
-331.2	7.1	73	181	187.1	-6.1	37.21
-331.2	7.1	75	208	201.3	6.7	44.89
					Total	766.51

$y_i - y_i'$: Prediction error
$(y_i - y_i')^2$: Squared prediction error

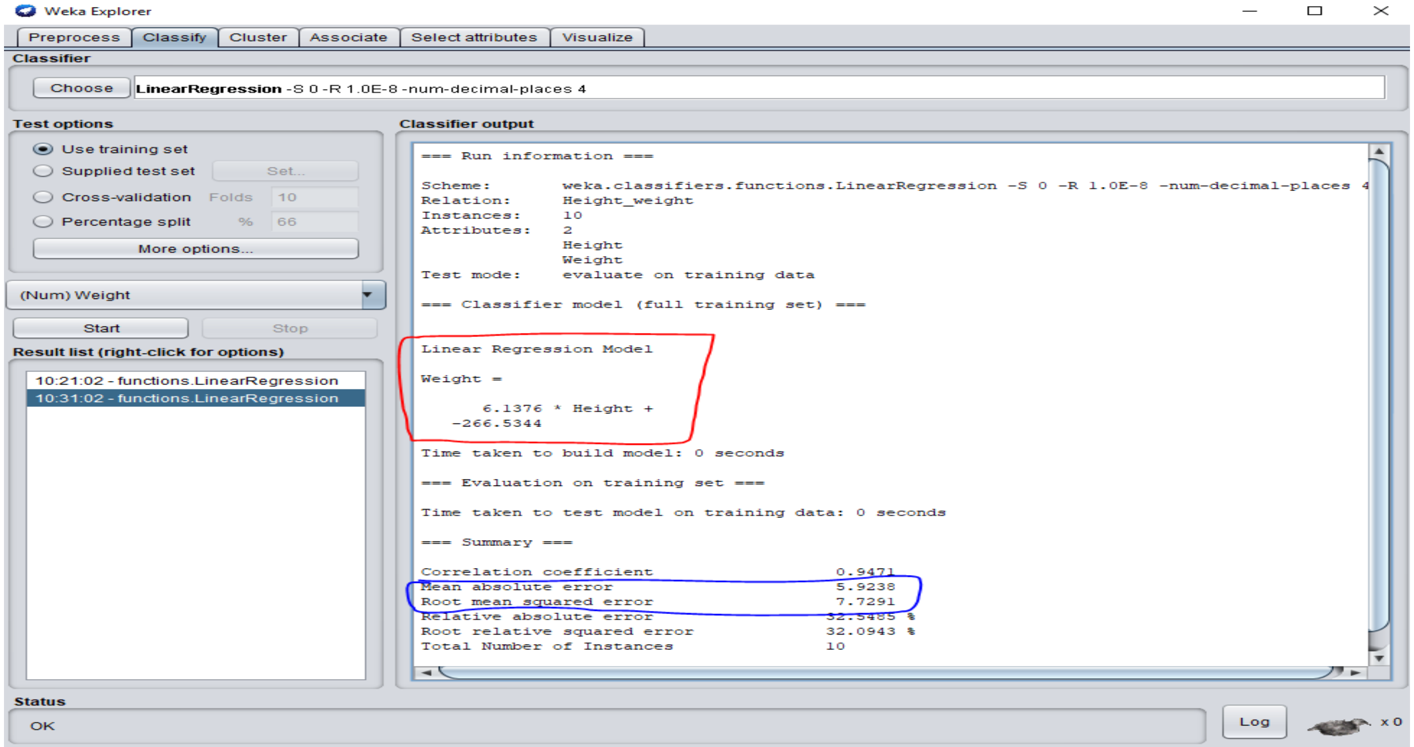
$$Overall\ squared\ prediciton\ error = \sum_{i=1}^{n} (\text{yi} - \text{yi'})2$$

Finding m and b

	x	y_i	$\bar{x}$	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$\bar{y}$	$y_i - \bar{y}$	$(y_i - \bar{y})^2$	$x_i - \bar{x}$ * $y_i - \bar{y}$
	63	127	69.3	-6.3	39.69	158.8	-31.8	1011.24	200.34
	64	121	69.3	-5.3	28.09	158.8	-37.8	1428.84	200.34
	66	142	69.3	-3.3	10.89	158.8	-16.8	282.24	55.44
	69	157	69.3	-0.3	0.09	158.8	-1.8	3.24	0.54
	69	162	69.3	-0.3	0.09	158.8	3.2	10.24	-0.96
	71	156	69.3	1.7	2.89	158.8	-2.8	7.84	-4.76
	71	169	69.3	1.7	2.89	158.8	10.2	104.04	17.34
	72	165	69.3	2.7	7.29	158.8	6.2	38.44	16.74
	73	181	69.3	3.7	13.69	158.8	22.2	492.84	82.14
	75	208	69.3	5.7	32.49	158.8	49.2	2420.64	280.44
Sqrt(Sum)					11.7516			76.15510488	847.6
								$m = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$	6.137581463

m	6.1375815	
SD	3.7161808	24.08236
Mean	69.3	158.8
$b = \bar{y} - m\bar{x}$	-266.5344	

Weka Demo for Height-Weight file



Measuring accuracy - How can you tell if your regression line is a good fit?

- Calculate the “Coefficient of determination”, the residual, or also called r^2, where r is the correlation coefficient.

- This is a number between 0 and 1, which normally means how close your data is to the line. If your data is always on the line, then $R^2 = 1$. If your data is far away from the line then R^2 will be low.

Measuring accuracy

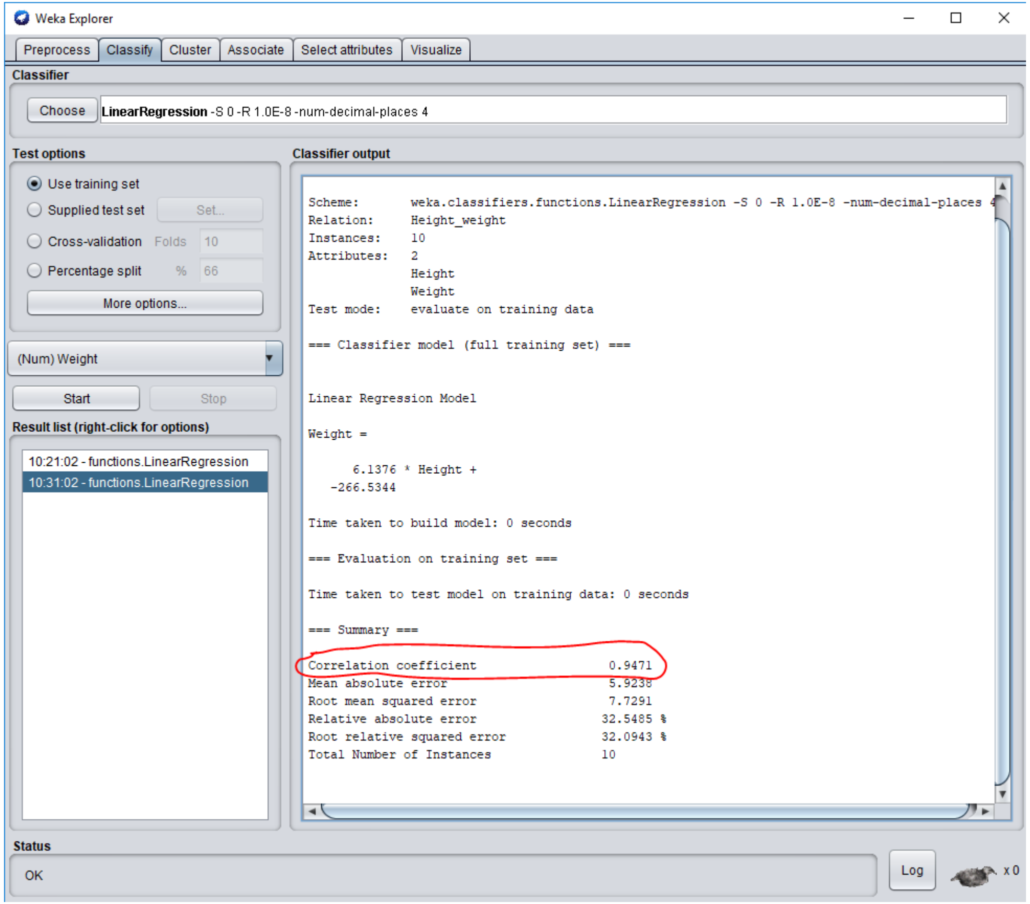
Correlation Coefficient

$$r = \frac{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} * b_1$$

where b_1 is the slope in the equation $y = b_0 + b_1x$

	x	y_i	$\bar{x}$	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$\bar{y}$	$y_i - \bar{y}$	$(y_i - \bar{y})^2$
	63	127	69.3	-6.3	39.69	158.8	-31.8	1011.24
	64	121	69.3	-5.3	28.09	158.8	-37.8	1428.84
	66	142	69.3	-3.3	10.89	158.8	-16.8	282.24
	69	157	69.3	-0.3	0.09	158.8	-1.8	3.24
	69	162	69.3	-0.3	0.09	158.8	3.2	10.24
	71	156	69.3	1.7	2.89	158.8	-2.8	7.84
	71	169	69.3	1.7	2.89	158.8	10.2	104.04
	72	165	69.3	2.7	7.29	158.8	6.2	38.44
	73	181	69.3	3.7	13.69	158.8	22.2	492.84
	75	208	69.3	5.7	32.49	158.8	49.2	2420.64
Sqrt(Sum)					11.7516			76.1551049
				Correlation Coefficient	$\frac{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} * b_1$		0.947101228	

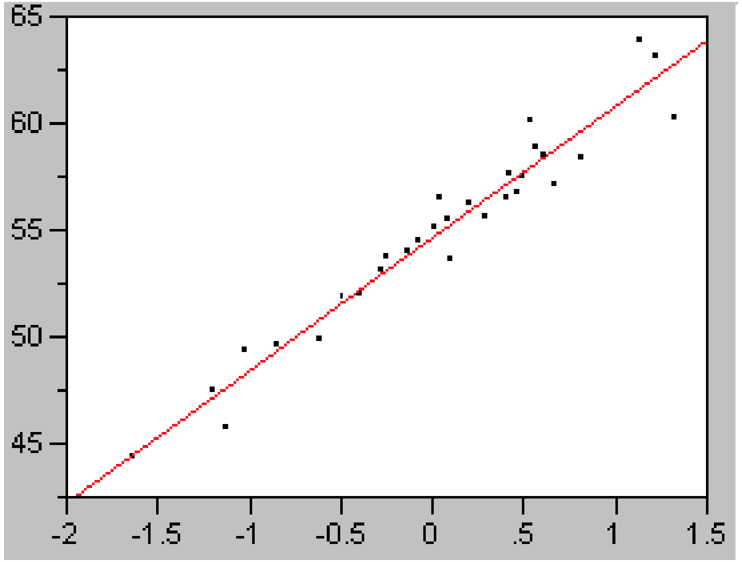
Weka



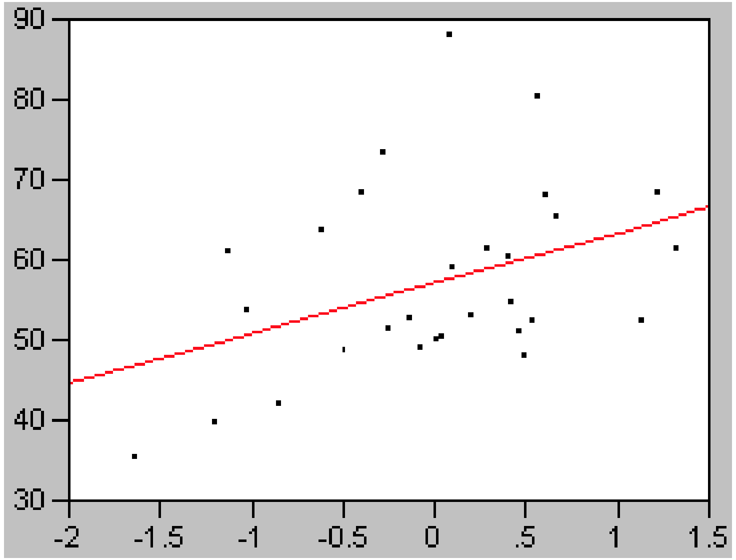
Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
LinearRegression -S 0 -R 1.0E-8 -num-decimal-places 4
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Num) Weight
Start
Stop
Result list (right-click for options)
10:21:02 - functions.LinearRegression
10:31:02 - functions.LinearRegression
Classifier output
Scheme: weka.classifiers.functions.LinearRegression -S 0 -R 1.0E-8 -num-decimal-places 4
Relation: Height_weight
Instances: 10
Attributes: 2
Height
Weight
Test mode: evaluate on training data
=== Classifier model (full training set) ===
Linear Regression Model
Weight =
6.1376 * Height +
-266.5344
Time taken to build model: 0 seconds
=== Evaluation on training set ===
Time taken to test model on training data: 0 seconds
=== Summary ===
Correlation coefficient 0.9471
Mean absolute error 5.9238
Root mean squared error 7.7291
Relative absolute error 32.5485 %
Root relative squared error 32.0943 %
Total Number of Instances 10
Status
OK
Log
x 0

Measuring accuracy

High R^2, data is close to line



Lower R^2, data is far from line



Multiple Regression Model

Linear Regression Model for cpu.arff:

class = 0.0491 * MYCT +
0.0152 * MMIN +
0.0056 * MMAX +
0.6298 * CACH +
1.4599 * CHMAX +
-56.075

The weights tells the relationship of each variable to the outcome, whether they are positive or negative.

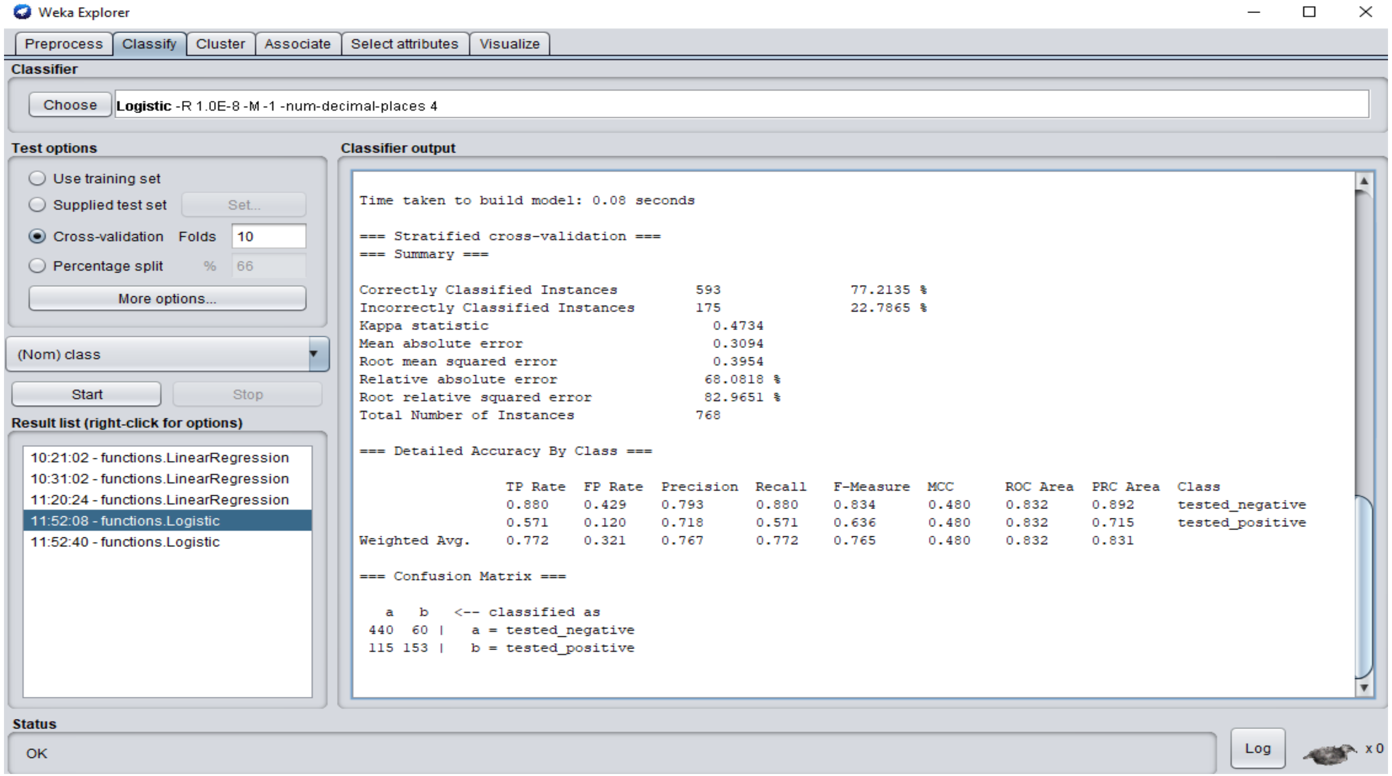
Multivariate Regression

- a technique that estimates a single regression model with more than one outcome variable.
- Example: A doctor has collected data on cholesterol, blood pressure, and weight. She also collected data on the eating habits of the subjects (e.g., how many ounces of red meat, fish, dairy products, and chocolate consumed per week). She wants to investigate the relationship between the three measures of health and eating habits.

Logistic Regression

- Models a relationship between independent (predictor) variable and a categorical response variable.
- Helps us to estimate a probability of falling into a certain level of the categorical response given a set of predictors

Weka Demo with Diabetes dataset



References

- https://www.youtube.com/watch?v=6tDnNyNZDF0
- https://www.youtube.com/watch?v=YIxoyiN8lxo
- https://www.youtube.com/watch?v=ThmZU3dTIDo
- https://onlinecourses.science.psu.edu/stat501/lesson/1