Regression

Mario Guarracino

Introduction

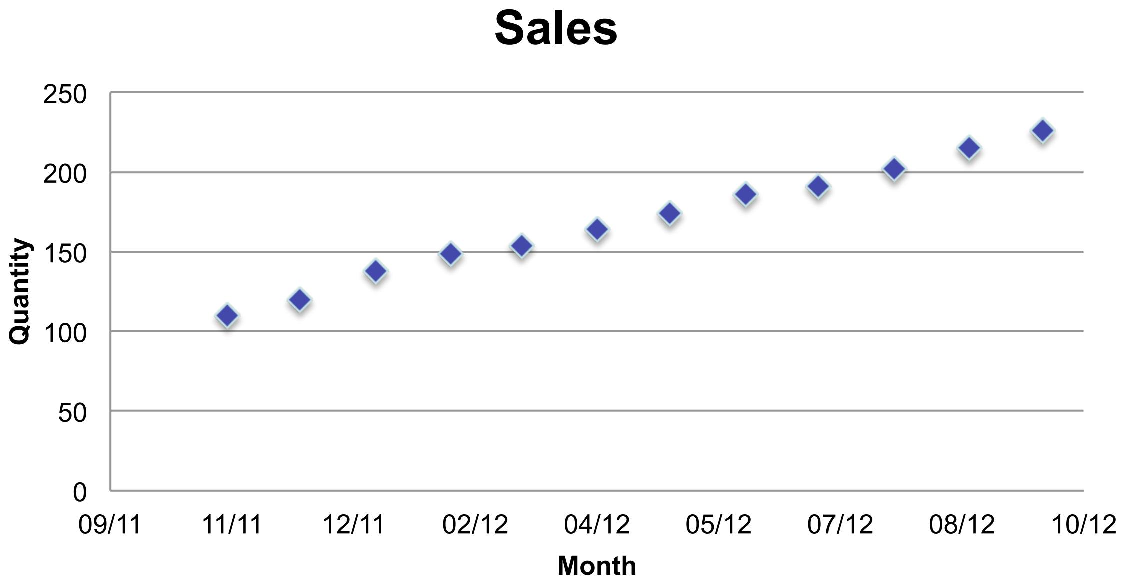
- In the last lesson, we saw how to aggregate data from different sources, identify measures and dimensions, to build data marts for business analysis.
- Some techniques were introduced to treat missing data and to detect outliers.
- In this lesson, we extend those topics introducing linear regression, an analysis method borrowed from statistics.
- It is used for two main purposes:
  - As a predictive model fitted to observed data,
  - to model the relation between an independent variable and one or more dependent variables.

Introduction

- A linear regression is a statistical model that attempts to show the relationship between at least two variables with a linear equation.
- A regression analysis involves graphing a line over a set of data points that most closely fits the overall shape of the data.
- A regression shows the extent to which changes in a "dependent variable," which is put on the y-axis, can be attributed to changes in an "explanatory variable," which is placed on the x-axis.

Example

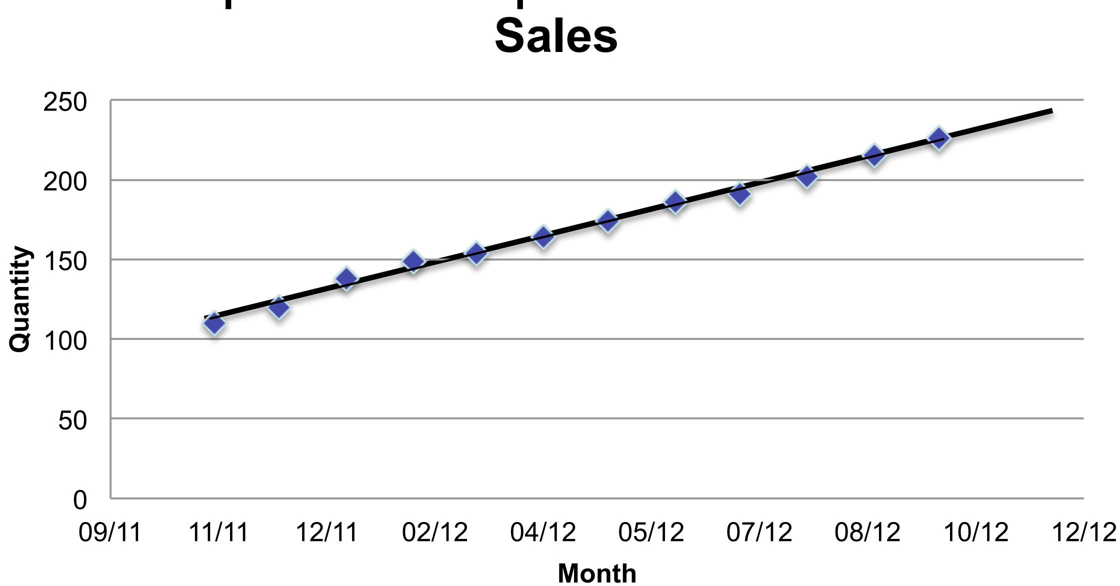
- Linear regressions can be used in business intelligence to evaluate trends and make estimates or forecasts.



- We can plot the sales data with monthly sales on the y-axis and time on the x-axis

Example

- Regression analysis on the sales data would produce a line that that depicts the upward trend in sales.



- After creating the trend line, the company could use the slope of the line to forecast sales in future months.

Other examples

- Analyzing the impact of price changes
  - A company changes the price of a product several times, and records the quantity sold for each price. With a linear regression (price vs quantity) can predict sells at other price levels.
- Assessing Risk
  - A health insurance company conduct a linear regression plotting number of claims per customer against age to discover whether older customers tend to make more health insurance claims.

Becoming energy efficient

- Dan, the facility manager of an office building, has decided to save some money by becoming more energy efficient.
- In December 2010, Dan spent large part of his budget on improving the building's insulation in order to *greatly* reduce the energy it took to heat the building.
- After one year, Dan compares the energy consumption:

Year	Energy consumption (kW/h)
2010	452,976
2011	445,241

- The improvement was much lower that expected.

Example

- Dan consults a colleague, who points out heating consumption is related to *heating degree days (HDD)*:
  - A 10% increase in degree days for a d/w/m/y produce 10% more heating energy consumption in the same period.
- Dan downloads the degree days for his area, and obtains:

Year	Energy consumption (kW/h)	HDD	(kW/h)/HDD
2010	452,976	3,320	136
2011	445,241	4,092	109

- In fact, the energy consumption has a 20% decrease.

Base temperature of a building

- The *base temperature* of a building is the temperature below which that building needs heating.
- It is usually determined by experience, as the temperature that better suits user needs.
- If external temperature is above base temperature, heating is not needed.
- In general, building have an internal heat gain, due to insulation and activities (sun, people, equipment,…)
- For example, if base temperature is 20C, and heat gain is 3C, heating is needed when temperature is below 17C.

Degree days

- Degree days are essentially a simplification of historical weather data.
- *Heating degree days*, or *HDD*, are a measure of how much (in degrees), and for how long (in days), outside air temperature was *lower* than the base temperature.
- HDD are used for calculations relating to the energy consumption required to *heat* buildings.
- HDD are typically computed as weekly or monthly figures.
  - Summing them together to make figures covering a longer period (e.g. sum 12 consecutive monthly HDD to make an annual degree-day total).

Example

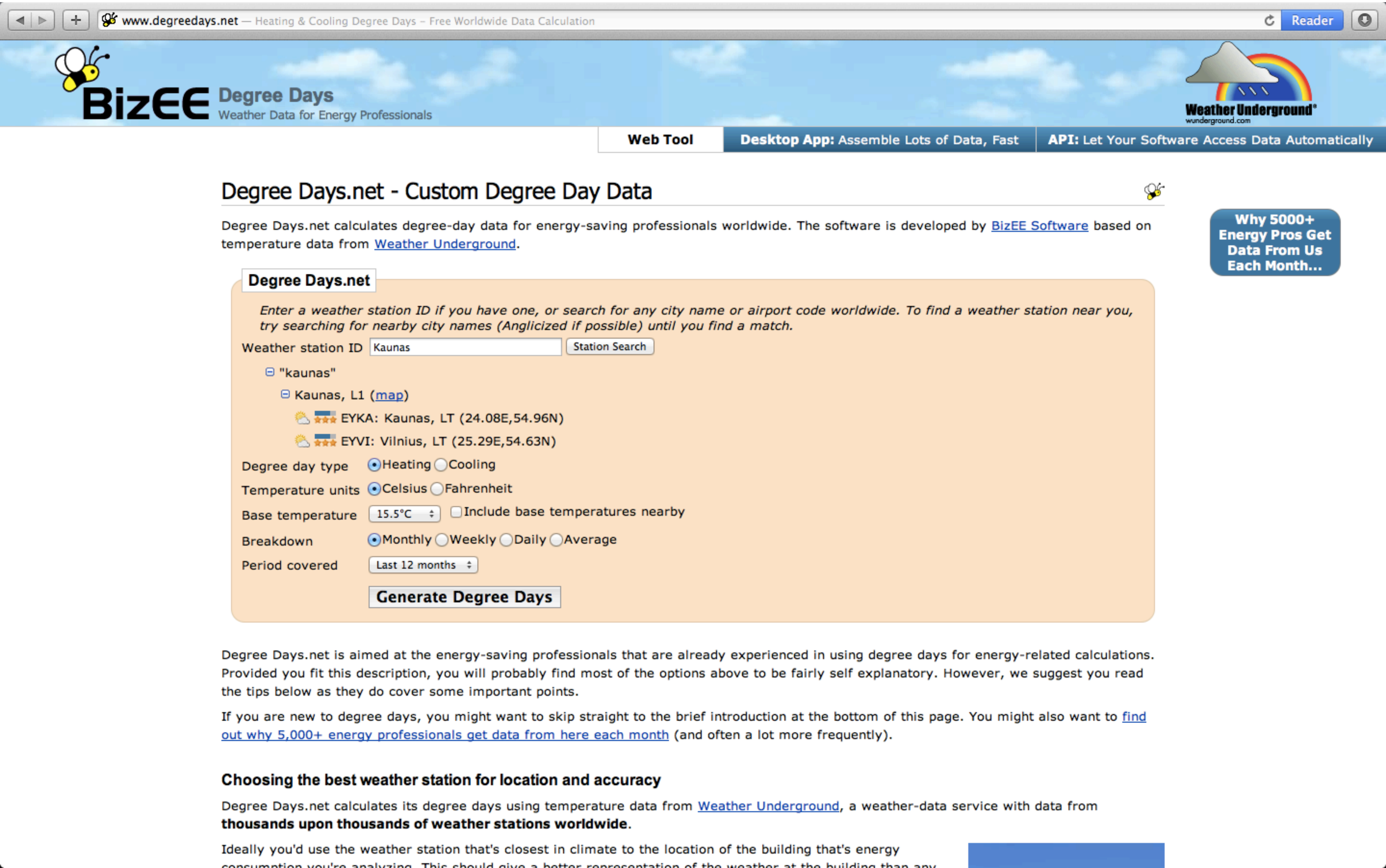
- Suppose the base temperature of a building is 17C.
  - Day 1: external temperature is constantly 16C for all day:
    - 1 degree * 1 day = 1 HDD
  - Day 2: external temperature is 2 degrees below the base temperature, we have:
    - 2 degrees * 1 day = 2 HDD
  - Day 3 outside temperature is 17C
    - 0 degree * 1 day = 0 HDD
  - Day 4 outside temperature is 19C
    - 0 degree * 1 day = 0 HDD
  - Day 5: 15C from 00:00 to 12:00, 16C from 12:00 to 24:00
    - (2 degrees * 0.5 days) + (1 degree * 0.5 days) = 1.5 HDD

Example

- ➢ Day 6: 16C from 00:00 to 06:00, 15C from 06:00 to 12:00, 14C from 12:00 to 18:00, and 13C from 18:00 to 24:00:
  - (1 degree * 0.25 days) + (2 degrees * 0.25 days) + (3 degrees * 0.25 days) + (4 degrees * 0.25 days) = 2.5 HDD
- ➢ Day 7: 13C from 00:00 to 00:30, 12.9C from 00:30 to 01:00, 12.9C from 01:00 to 01:30, 12.8 from 01:30 to 02:00,...
  - (3 degrees * 1/48 days) + (3.1 degrees * 1/48 days) + ... = 1.9 HDD

- We expect the heating energy consumption on each of those days to vary proportionally to the heating degree days.
- It is possible to add HDD to obtain values for longer periods.

Kauno HDD


www.degreedays.net — Heating & Cooling Degree Days – Free Worldwide Data Calculation
Reader
BizEE Degree Days
Weather Data for Energy Professionals
Weather Underground
wunderground.com
Web Tool
Desktop App: Assemble Lots of Data, Fast
API: Let Your Software Access Data Automatically
Degree Days.net - Custom Degree Day Data
Why 5000+ Energy Pros Get Data From Us Each Month...
Degree Days.net calculates degree-day data for energy-saving professionals worldwide. The software is developed by BizEE Software based on temperature data from Weather Underground.
Degree Days.net
Enter a weather station ID if you have one, or search for any city name or airport code worldwide. To find a weather station near you, try searching for nearby city names (Anglicized if possible) until you find a match.
Weather station ID
Kaunas
Station Search
"kaunas"
Kaunas, L1 (map)
EYKA: Kaunas, LT (24.08E,54.96N)
EYVI: Vilnius, LT (25.29E,54.63N)
Degree day type
Heating
Cooling
Temperature units
Celsius
Fahrenheit
Base temperature
15.5°C
Include base temperatures nearby
Breakdown
Monthly
Weekly
Daily
Average
Period covered
Last 12 months
Generate Degree Days
Degree Days.net is aimed at the energy-saving professionals that are already experienced in using degree days for energy-related calculations. Provided you fit this description, you will probably find most of the options above to be fairly self explanatory. However, we suggest you read the tips below as they do cover some important points.
If you are new to degree days, you might want to skip straight to the brief introduction at the bottom of this page. You might also want to find out why 5,000+ energy professionals get data from here each month (and often a lot more frequently).
Choosing the best weather station for location and accuracy
Degree Days.net calculates its degree days using temperature data from Weather Underground, a weather-data service with data from thousands upon thousands of weather stations worldwide.
Ideally you'd use the weather station that's closest in climate to the location of the building that's energy

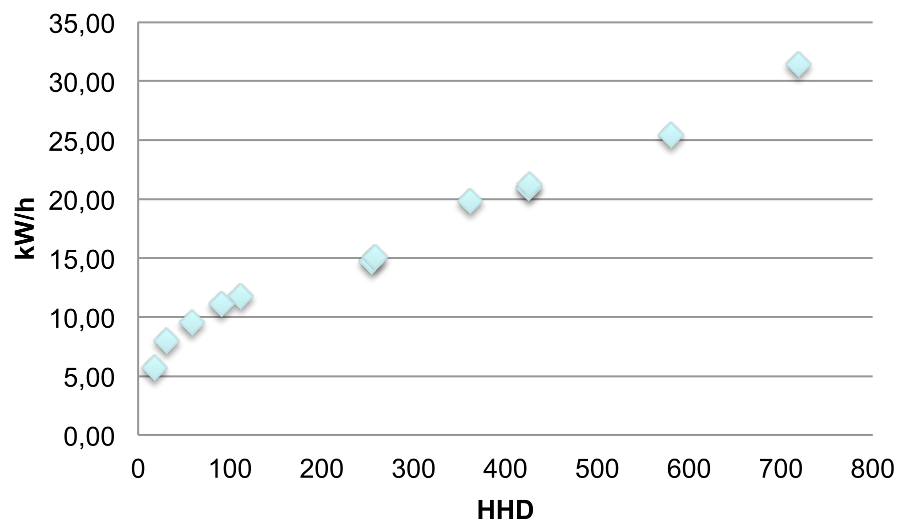
Linear regression analysis

- Linear regression analysis is now used as a monitoring and targeting technique.
- Central to this is the assumption that energy consumption is caused by a "*driving factor*" (or "*driver*") - in the case of heating or cooling, the degree days.
- So, for a heated building, it is assumed that the energy consumption required to heat that building for any particular period is proportional to (or *driven by*) the number of heating degree days over that period.
- Typically you would select a "*baseline*" set of energy consumption data: this would usually be weekly or monthly data from the past year or two.

Scatter plot

- For each figure of energy consumption, you need a corresponding figure for the degree days.

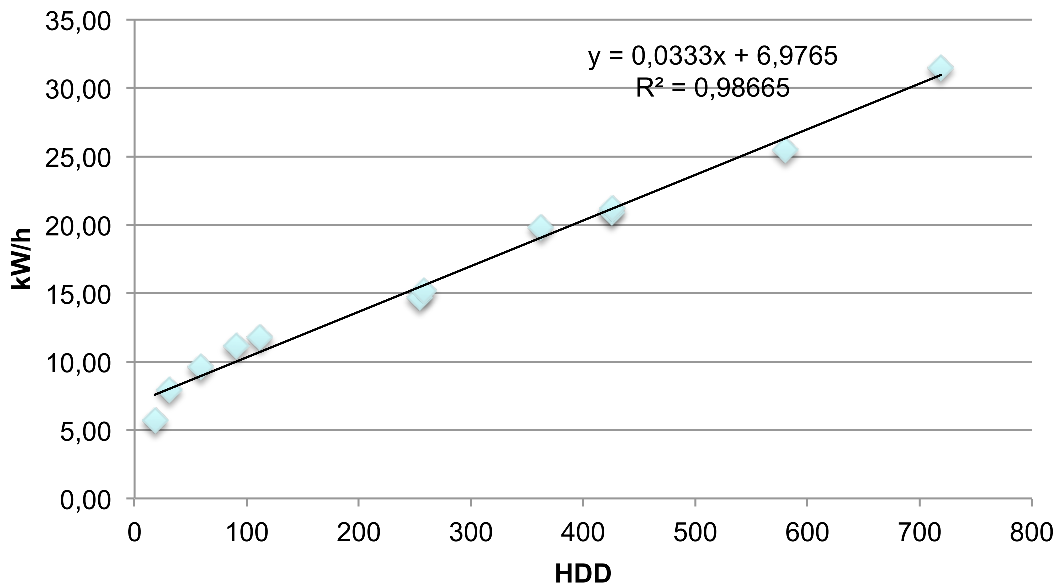
Month	HDD	KW/h
11/2011	362	19,82
12/2011	425	20,91
01/2012	580	25,47
02/2012	719	31,45
03/2012	426	21,21
04/2012	254	14,72
05/2012	112	11,80
06/2012	59	9,58
07/2012	18	5,71
08/2012	31	7,97
09/2012	91	11,13
10/2012	258	15,13



We can now correlate these two sets of figures

Regression plot

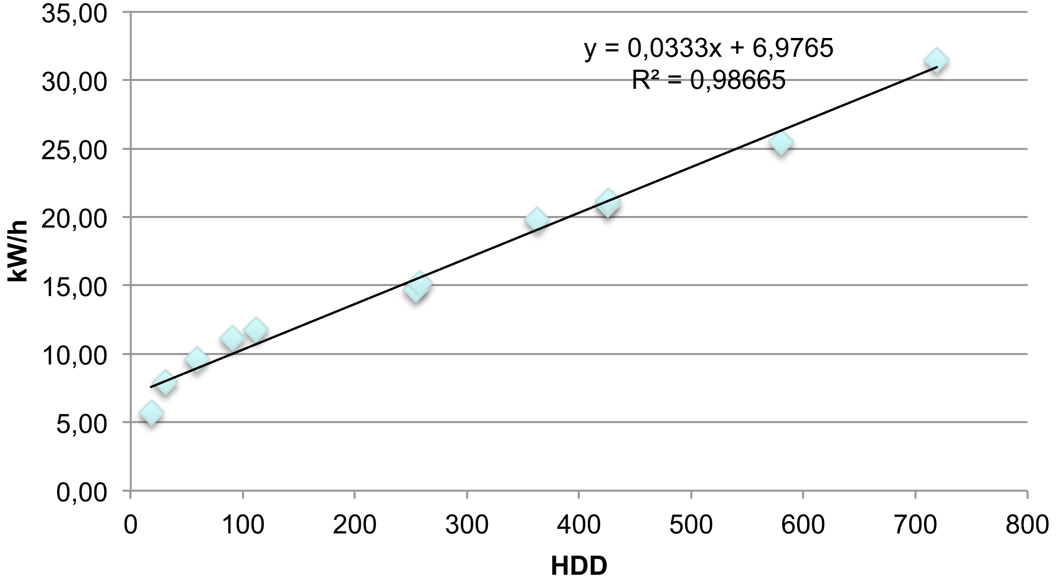
- The "*regression line*" is the line of best fit through the points in the scatter chart.



- It is often known as the "*trend line*" or the "*performance characteristic line*".

Regression plot

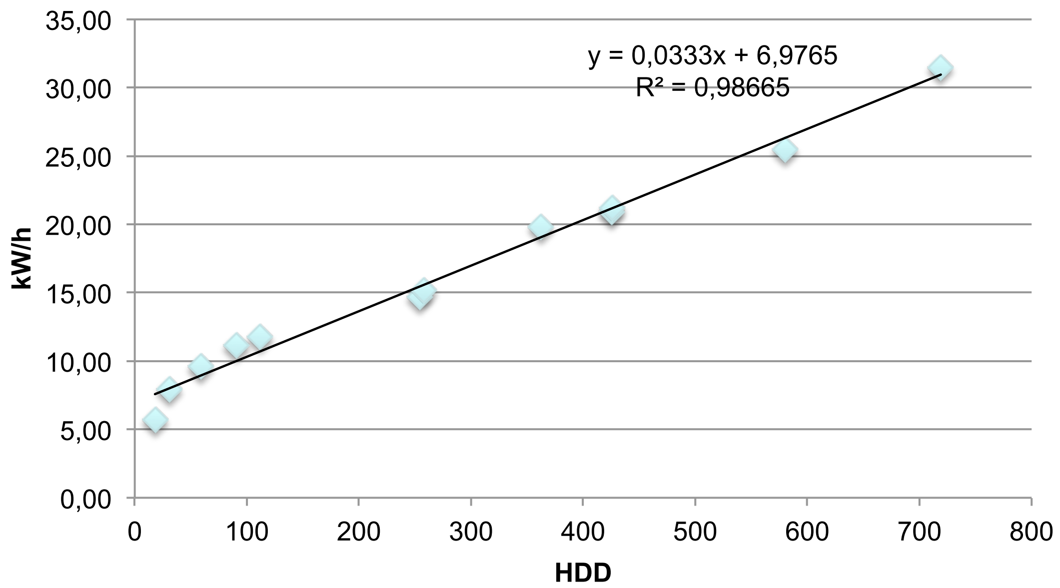
- The "*regression line*" is the line of best fit through the points in the scatter chart.



- The "*y*" corresponds to the kWh.
- The "*x*" corresponds to the degree days.

Regression plot

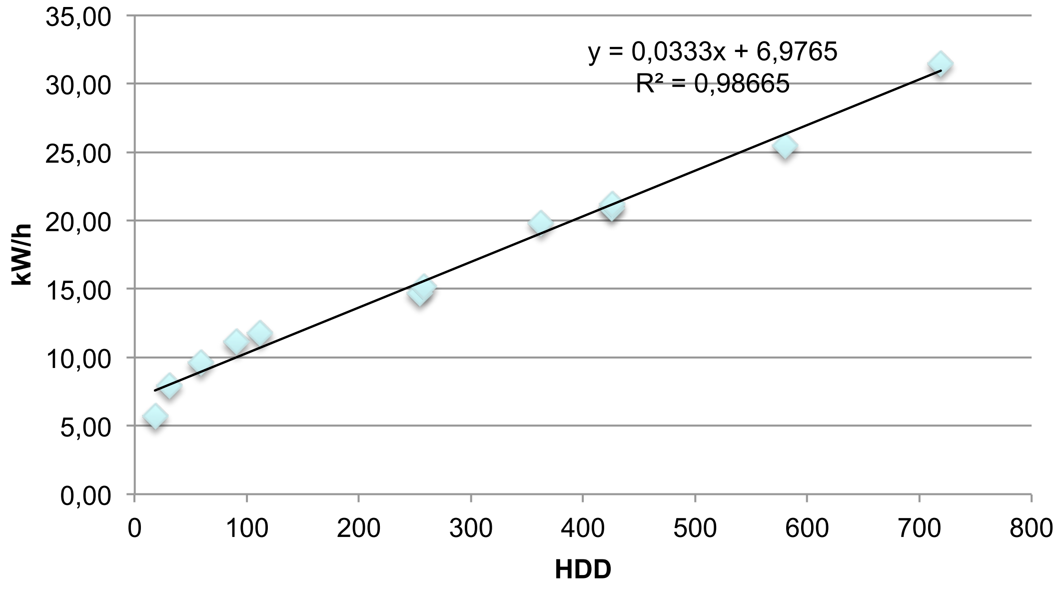
- The "*regression line*" is the line of best fit through the points in the scatter chart.



- The figure that multiplies the *x* (0,033) represents the **gradient** of the trend line.
- The other constant (6,98) is the **intercept**. It represents the point at which the trend line crosses the y axis.

Regression plot

- The "*regression line*" is the line of best fit through the points in the scatter chart.



- The R^2 value is a measure of how good is the correlation.
- The closer the R^2 value is to 1, the better the correlation.

Energy consumption

- Once the formula of the regression line has been established, you can use it to calculate the *baseline*, or *expected*, energy consumption from the degree days.
- So, each time you obtain a new figure for the degree days (typically each week or month), you can use in the regression-line formula to get the expected energy consumption.
- You can compare this figure with the actual energy consumption for the period, to determine whether more energy was used than expected.

Optimal base temperature

- The optimal base temperature varies from building to building.
- It's difficult to estimate the correct base temperature accurately for any particular building using logic alone, so it can be helpful to make a rough estimate and then try correlating kWh with degree days calculated to various base temperatures around that point.
- R^2 gives a way to compare the strength of the different correlations.

Optimal base temperature

- Testing various base temperatures can give you a useful indication.

N20 fx =SLOPE($O$8:$O$19, N$8:N$19)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
6		(Column titles show the base temperature in Celsius)													
7	Month starting	12.5	13	13.5	14	14.5	15	15.5	16	16.5	17	17.5	18	18.5	kWh
8	1 Oct 2009	92	102	113	124	136	150	163	178	192	206	221	236	251	593
9	1 Nov 2009	140	154	169	183	198	213	228	243	258	273	288	303	318	676
10	1 Dec 2009	250	265	281	296	312	327	343	358	374	389	405	420	436	1335
11	1 Jan 2010	280	295	311	326	342	357	373	388	404	419	435	450	466	1149
12	1 Feb 2010	217	231	245	259	273	287	301	315	329	343	357	371	385	1127
13	1 Mar 2010	152	165	179	193	208	223	238	253	269	284	300	315	331	892
14	1 Apr 2010	67	78	89	101	112	125	137	151	164	178	192	206	220	538
15	1 May 2010	31	38	46	54	63	73	84	95	106	118	130	143	157	289
16	1 Jun 2010	10	14	17	21	26	32	38	45	52	59	68	76	86	172
17	1 Jul 2010	2	3	5	7	9	12	15	20	24	30	38	46	56	131
18	1 Aug 2010	3	4	5	7	9	11	14	18	22	28	34	40	47	134
19	1 Sep 2010	9	12	15	18	22	27	34	42	51	60	71	82	93	134
20	Gradient	4.253	4.06	3.871	3.713	3.56	3.435	3.317	3.225	3.127	3.053	2.986	2.927	2.874	
21	Intercept	153.4	137	121.7	105.8	90.18	71.63	53.5	31.6	12.4	-9.76	-34.3	-58.1	-84.2	
22	R2	0.96	0.963	0.965	0.967	0.967	0.967	0.966	0.965	0.963	0.962	0.96	0.958	0.956	

Linear regression

- To compute the coefficients m and b of the linear regression

$$y = mx + b$$

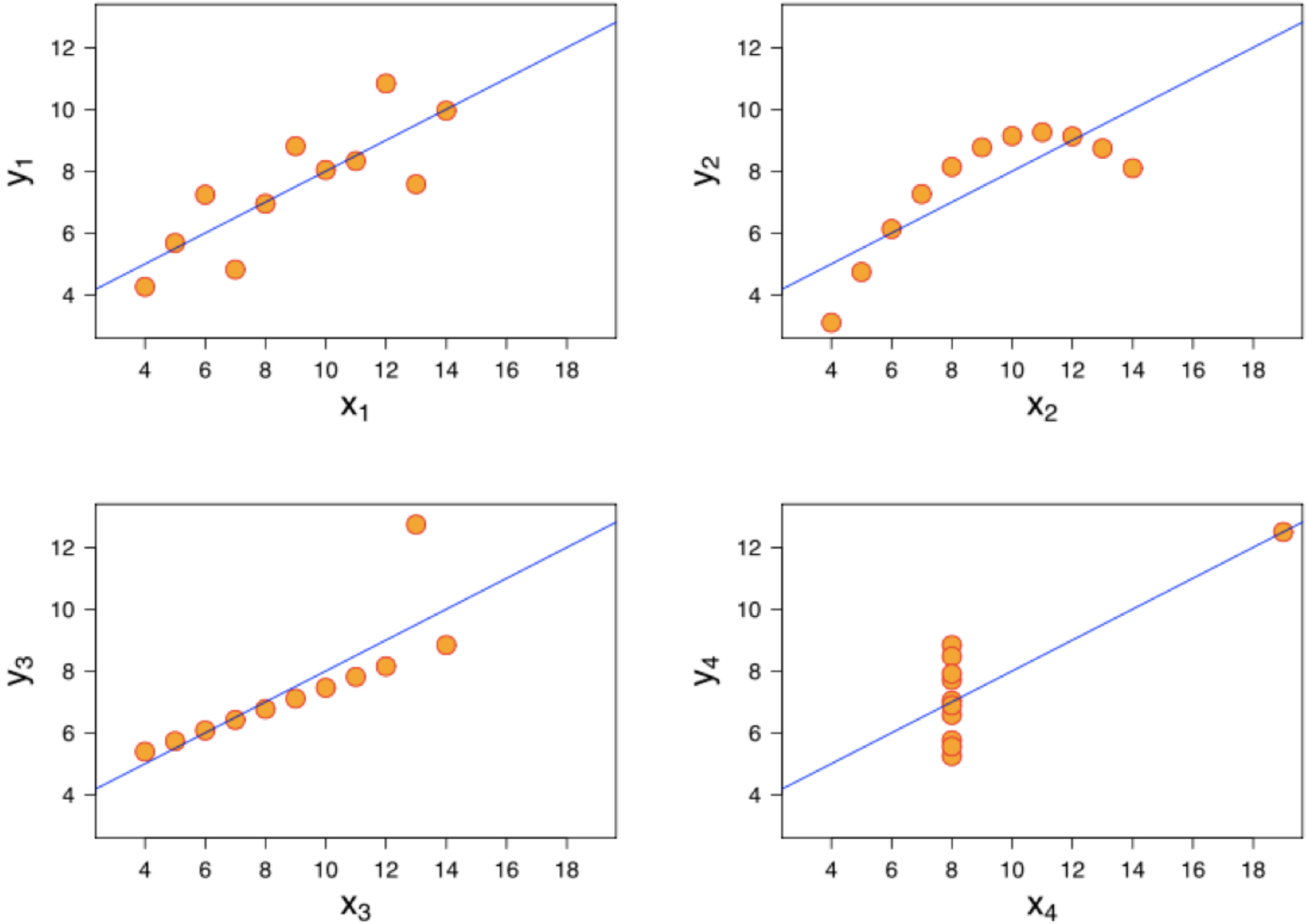
from the set of values (y_i, x_i) $i = 1, \ldots n$ we use:

$$m = \frac{\sum (x - \mu_x)(y - \mu_y)}{\sum (x - \mu_x)^2}$$

$$b = \mu_y - m\mu_x$$

$$R^2 = 1 - \frac{SS_{err}}{SS_{tot}}, \quad SS_{err} = \sum_i (y_i - y(x_i))^2, \quad SS_{tot} = \sum_i (y_i - \mu_y)^2$$

Same regressor, different data



Summary

- In this lesson we learned basic concepts about linear regression
- We have seen ho to apply linear regression for monitoring and targeting energy consumption.
- Finally we learned how to estimate quality of regression and how to compute it.