Regression
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.
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.
Regression analysis on the sales data would produce a line 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:

<table>
<thead>
<tr>
<th>Year</th>
<th>Energy consumption (kW/h)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2010</td>
<td>452,976</td>
</tr>
<tr>
<td>2011</td>
<td>445,241</td>
</tr>
</tbody>
</table>

- The improvement was much lower than 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:

<table>
<thead>
<tr>
<th>Year</th>
<th>Energy consumption (kW/h)</th>
<th>HDD</th>
<th>(kW/h)/HDD</th>
</tr>
</thead>
<tbody>
<tr>
<td>2010</td>
<td>452,976</td>
<td>3,320</td>
<td>136</td>
</tr>
<tr>
<td>2011</td>
<td>445,241</td>
<td>4,092</td>
<td>109</td>
</tr>
</tbody>
</table>

- 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, buildings have an internal heat gain, due to insulation and activities (sun, people, equipment, …)
- For example, if base temperature is 20°C, and heat gain is 3°C, heating is needed when temperature is below 17°C.

Mario Guarracino
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).

Mario Guarracino
Example

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

Mario Guarracino
Example

- Day 6: 16°C from 00:00 to 06:00, 15°C from 06:00 to 12:00, 14°C from 12:00 to 18:00, and 13°C from 18:00 to 24:00:
  - \((1 \text{ degree} \times 0.25 \text{ days}) + (2 \text{ degrees} \times 0.25 \text{ days}) + (3 \text{ degrees} \times 0.25 \text{ days}) + (4 \text{ degrees} \times 0.25 \text{ days}) = 2.5 \text{ HDD}\)

- Day 7: 13°C from 00:00 to 00:30, 12.9°C from 00:30 to 01:00, 12.9°C from 01:00 to 01:30, 12.8°C from 01:30 to 02:00,...
  - \((3 \text{ degrees} \times 1/48 \text{ days}) + (3.1 \text{ degrees} \times 1/48 \text{ days}) + \ldots = 1.9 \text{ HDD}\)

- 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.

Mario Guarracino
Degree Days.net - Custom Degree Day Data

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.

**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

- **“kaunas”**
- **Kaunas, L1** (man)
- **EYKA**: Kaunas, LT (24.08E, 54.96N)
- **EYVI**: Vilnius, LT (25.29E, 54.63N)

**Degree day type**: Heating

**Temperature units**: Celsius

**Base temperature**: 15.5℃

**Breakdown**: Monthly

**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 consumption you're analyzing. This should give a better representation of the weather at the building than any.
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.
For each figure of energy consumption, you need a corresponding figure for the degree days.

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".

Mario Guarracino
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.

Mario Guarracino
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.
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.

Mario Guarracino
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.
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.

| Month starting | A | B | C  | D  | E  | F  | G  | H  | I  | J  | K  | L  | M  | N  | O  |
|---------------|---|---|----|----|----|----|----|----|----|----|----|----|----|----|
| 7             |   |   |    |    |    |    |    |    |    |    |    |    |    |    |    |
| 8             |   |   |    |    |    |    |    |    |    |    |    |    |    |    |    |
| 9             |   |   |    |    |    |    |    |    |    |    |    |    |    |    |    |
| 10            |   |   |    |    |    |    |    |    |    |    |    |    |    |    |    |
| 11            |   |   |    |    |    |    |    |    |    |    |    |    |    |    |    |
| 12            |   |   |    |    |    |    |    |    |    |    |    |    |    |    |    |
| 13            |   |   |    |    |    |    |    |    |    |    |    |    |    |    |    |
| 14            |   |   |    |    |    |    |    |    |    |    |    |    |    |    |    |
| 15            |   |   |    |    |    |    |    |    |    |    |    |    |    |    |    |
| 16            |   |   |    |    |    |    |    |    |    |    |    |    |    |    |    |
| 17            |   |   |    |    |    |    |    |    |    |    |    |    |    |    |    |
| 18            |   |   |    |    |    |    |    |    |    |    |    |    |    |    |    |
| 19            |   |   |    |    |    |    |    |    |    |    |    |    |    |    |    |
| 20            |   |   |    |    |    |    |    |    |    |    |    |    |    |    |    |
| 21            |   |   |    |    |    |    |    |    |    |    |    |    |    |    |    |
| 22            |   |   |    |    |    |    |    |    |    |    |    |    |    |    |    |

N20 = \text{SLOPE}($O8$:$O19$, $N8:$N19)

(COLUMN TITLES SHOW THE BASE TEMPERATURE IN CELSIUS)


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

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

Mario Guarracino

http://www.degreedays.net/regression-analysis
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:

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

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

Mario Guarracino
Same regressor, different data

- $y_1$ vs $x_1$
- $y_2$ vs $x_2$
- $y_3$ vs $x_3$
- $y_4$ vs $x_4$
In this lesson we learned basic concepts about linear regression.

We have seen how to apply linear regression for monitoring and targeting energy consumption.

Finally, we learned how to estimate the quality of regression and how to compute it.