REGRESSÃO MULTIVARIÁVEL

A *Regressão Multivariável* é uma técnica de previsão de vendas que usa outras variáveis que podem ter uma influência sobre as vendas. Ela usa a relação entre dois tipos de variáveis: a *variável dependente* e as *variáveis independentes*. Por exemplo, suponha que você já sabe que as vendas dependem das alterações no PIB e da taxa de desemprego. As vendas previstas seriam a variável dependente, porque seu valor depende do valor do PIB e da taxa de desemprego que seriam as variáveis independentes. Então você precisaria determinar a intensidade da relação (correlação) entre estas duas variáveis para previsão de vendas. Se o PIB diminui em 1%, e a taxa de desemprego cai 2%, quanto irão suas vendas aumentar ou diminuir?

O resultado de uma analise por regressão multivariada poderia ser por exemplo concluir que uma cota de vendas anual razoável para uma loja é dada pela seguinte equação.

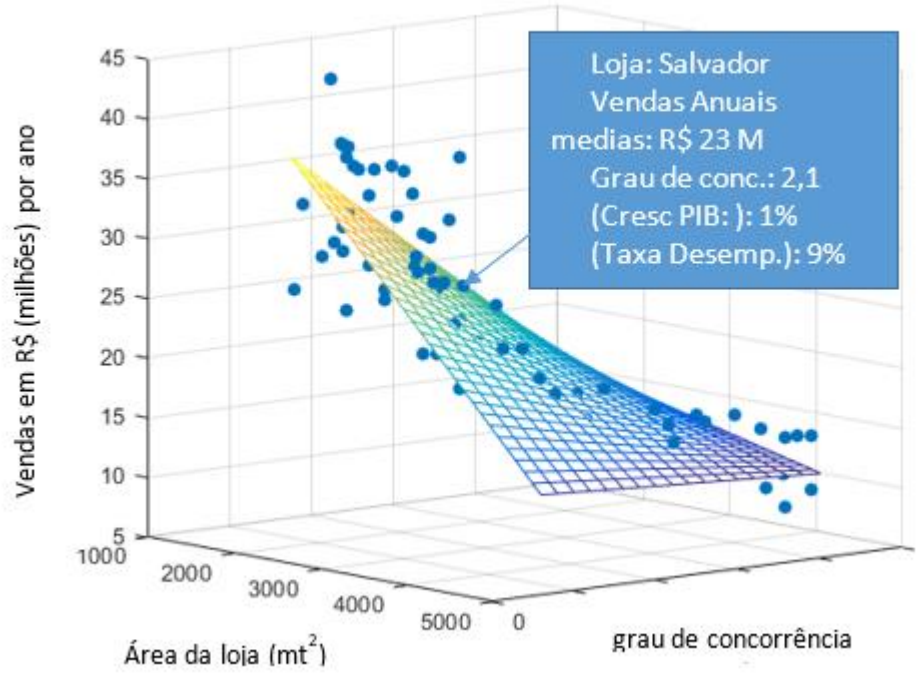
Vendas Previstas para o trimestre = 5(cresc PIB) -10*(grau da Concorrência) -100* (Taxa de Desemprego) + 1000* (Área em mt2 da loja)*

Os coeficientes 5, 10, 100, 1000 são calculados pelo modelo e definem a importância da variável associada na multiplicação, assim como se o efeito nas vendas previstas é positivo ou negativo.

Por exemplo, o coeficiente 5 do Crescimento do PIB tem influência positiva mas pequena nas vendas. O coeficiente -100 na Taxa de Desemprego tem influência negativa e relativamente grande nas vendas. Já o coeficiente 1000 na área da loja tem influência positiva e bastante grande nas vendas.

No gráfico tridimensional abaixo os pontos azuis representam lojas de uma rede de varejo. Cada ponto é associado com a venda média dos últimos três anos (descontada a inflação) que é a variável depedente e as variáveis independentes que são a área da loja em mt^2 e o grau de concorrência (em uma escala de 1 a 5).

Note que enquanto a venda é muito influenciada pela área da loja, o grau de concorrência tem menor influência. Note também que não é possível representar num gráfico tridimensional mais do que 3 variáveis, por isso não estão representados o Crescimento do PIB e a Taxa de desemprego.

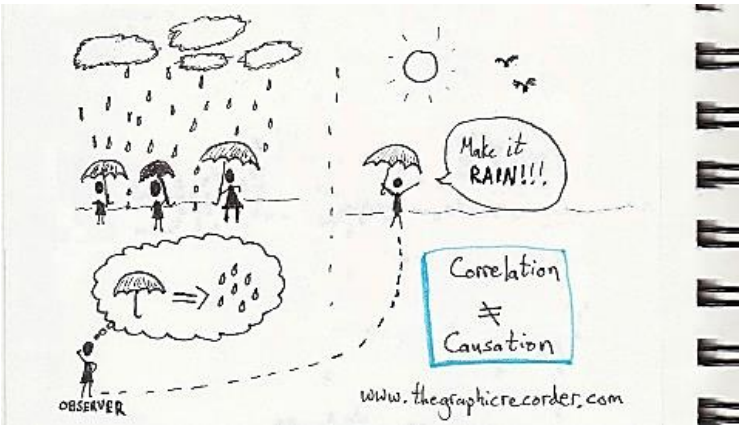


Após concluir quais são as variáveis independentes estatisticamente significativas (no exemplo acima temos quatro delas) , a pergunta natural é o quão bem a regressão representa a realidade?

Para fazer isso você deve analisar cuidadosamente cada variável que está influenciando o modelo e levar em conta os fatores citados a seguir.

Correlação não é causalidade

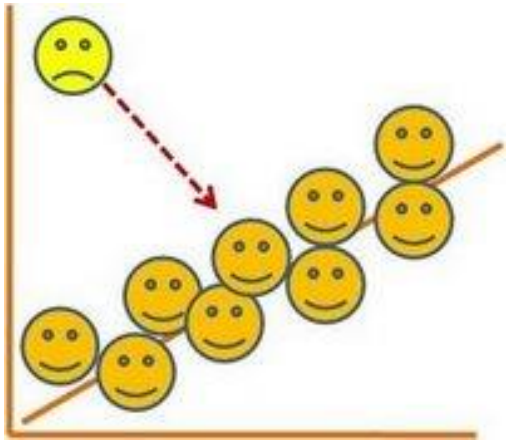
É importante não confundir correlação com causalidade. Por exemplo, é possível correlacionar o número de afogamentos em um uma praia com o número de sorvetes vendidos num dado período. O modelo pode dar previsões razoáveis, não porque sorvetes causam afogamentos, mas porque as pessoas tomam mais sorvetes em dias quentes quando elas também são mais propensas a nadar. Então as duas variáveis (vendas de sorvete e afogamentos) estão correlacionadas, mas uma não está causando a outra. Entretanto as correlações não deixam de ser úteis para a previsão, mesmo quando não há nenhuma relação causal entre as duas variáveis.



Via de regra um modelo melhor é possível se um mecanismo causal pode ser determinado. Neste exemplo, tanto as vendas de sorvete e afogamentos serão afetados pela temperatura e pelos números de pessoas que visitam a praia. Então um modelo melhor para afogamentos provavelmente poderia incluir temperaturas e o número de visitantes e excluir as vendas de sorvetes.

Dados Incomuns (*Outliers*)

Um *outlier* é um ponto de dados que é muito diferente do resto dos dados. Por exemplo, ao selecionar uma amostra das vendas de 30 lojas de uma rede, 29 delas estão entre R$ 150 e R$ 300 mil reais mensais e uma 30ª loja com vendas de R$ 3.000.000. Esta última provavelmente é um *outlier.* Na regressão, *um outlier* é um ponto de dados que está longe da linha de regressão em comparação com o resto dos dados.

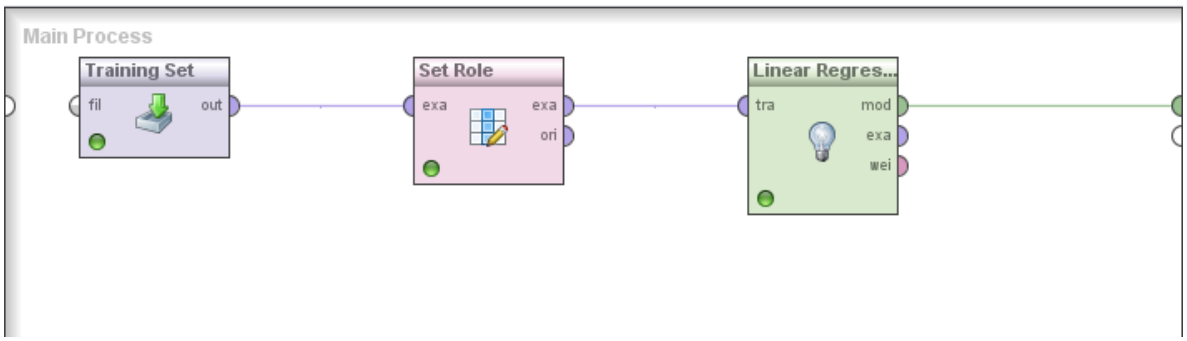


Outliers podem ocorrer por acaso em qualquer distribuição, mas eles geralmente indicam ou um erro de medição.

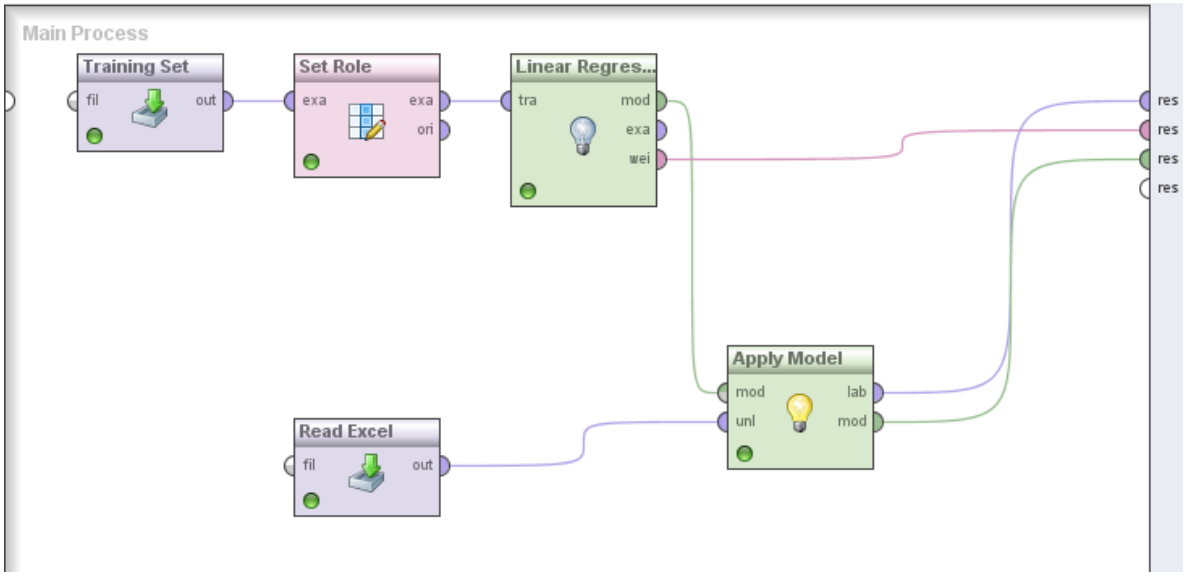
REGRESSÃO MULTIVARÁVEL NA PRÁTICA

Se você tem dúvidas sobre o que é um processo de análise preditiva leia primeiro "Implementação de um processo de análise preditiva"

O RapidMiner fornece uma ferramenta simples para regressão. O primeiro passo é importar os dados para treinar o modelo, usando o operador de leitura apropriado. Então você altera o tipo de atributo de seu campo de destino (variável dependente) para *"label"*, e adiciona o operador de Regressão Linear para gerar o modelo, como na figura abaixo.



Agora, você pode importar os dados de teste e usar o operador *Apply Model* para prever os resultados. O modelo é mostrado na figura a seguir.



Ao conectar a porta *weight* do operador de Regressão Linear à porta da janela do processo você terá os pesos (coeficientes) das variáveis independentes em uma tabela separada. Neste exemplo o modelo contém 5 variáveis independentes e uma variável dependente (que queremos prever).

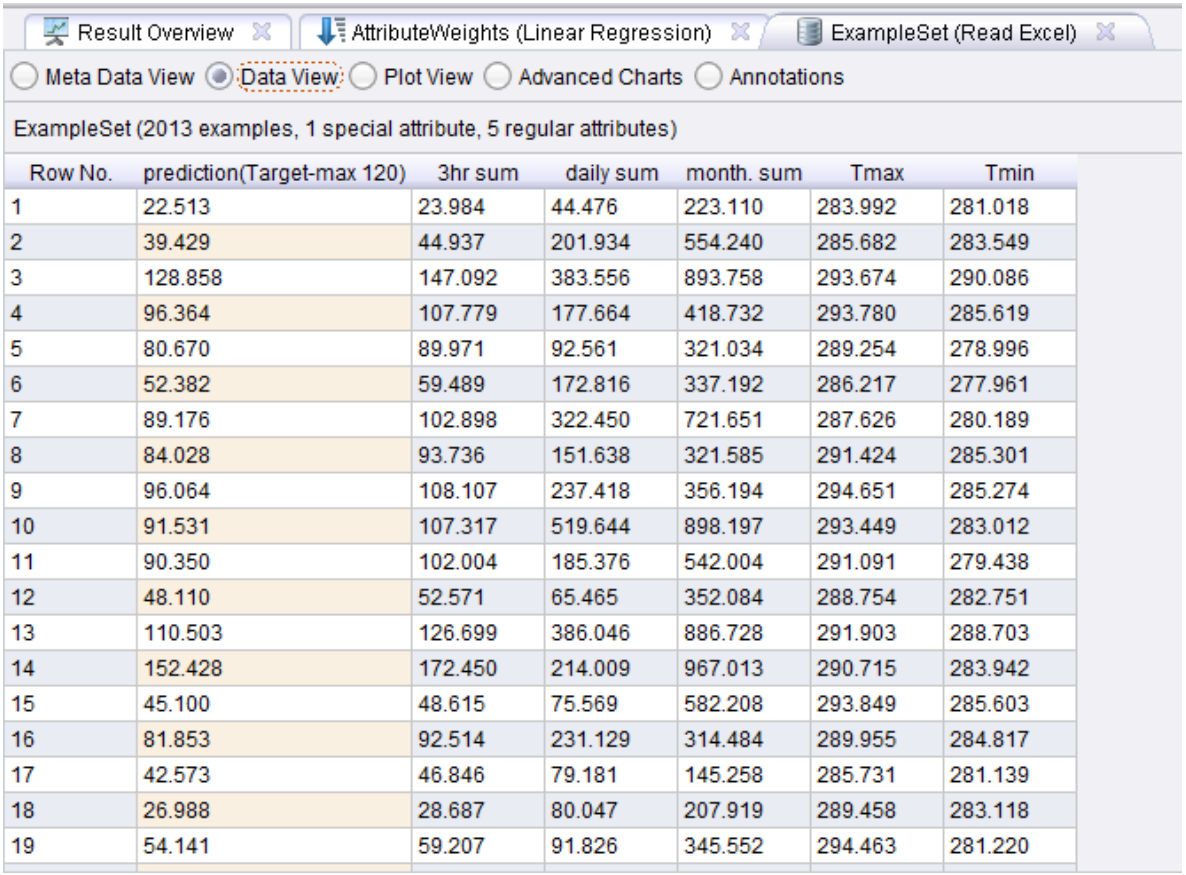
As tabelas a seguir mostram os resultados. O RapidMiner fornece as estatísticas do modelo de regressão, a equação da regressão e adiciona um campo de valores previstos para o conjunto de dados de teste. Você pode exportar os resultados para o excel.

Result Overview | AttributeWeights (Linear Regression) | ExampleSet (Read Excel)

Meta Data View | Data View | Plot View | Advanced Charts | Annotations

ExampleSet (2013 examples, 1 special attribute, 5 regular attributes)

Role	Name	Type	Statistics	Range	Missings
prediction	prediction(Target-max 120)	integer	avg = 80.005 +/- 39.402	[11.792 ; 271.600]	0
regular	3hr sum	real	avg = 90.055 +/- 45.542	[11.530 ; 311.223]	0
regular	daily sum	real	avg = 195.514 +/- 117.397	[15.497 ; 755.310]	0
regular	month. sum	real	avg = 572.571 +/- 253.852	[76.682 ; 1609.581]	0
regular	Tmax	real	avg = 291.263 +/- 3.475	[278.804 ; 299.231]	0
regular	Tmin	real	avg = 283.911 +/- 4.413	[270.000 ; 295.727]	0

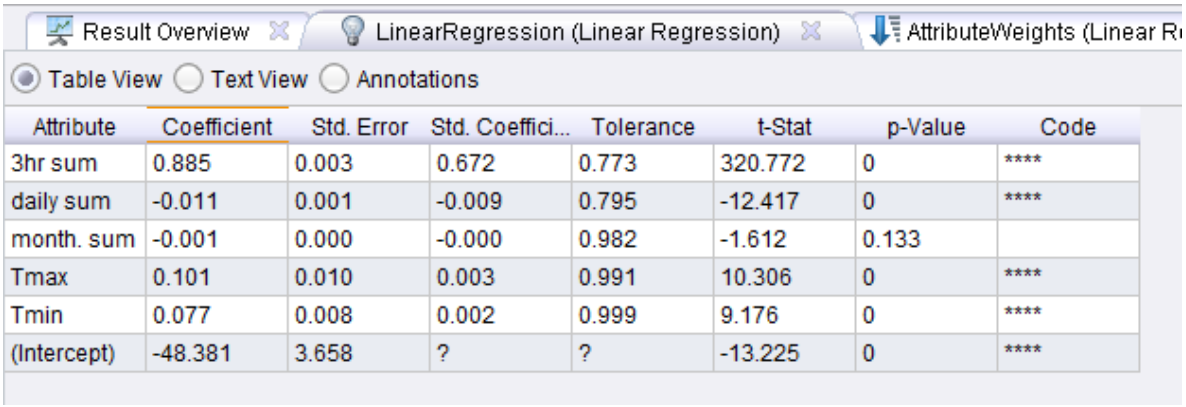


Result Overview | AttributeWeights (Linear Regression) | ExampleSet (Read Excel)

Meta Data View | Data View | Plot View | Advanced Charts | Annotations

ExampleSet (2013 examples, 1 special attribute, 5 regular attributes)

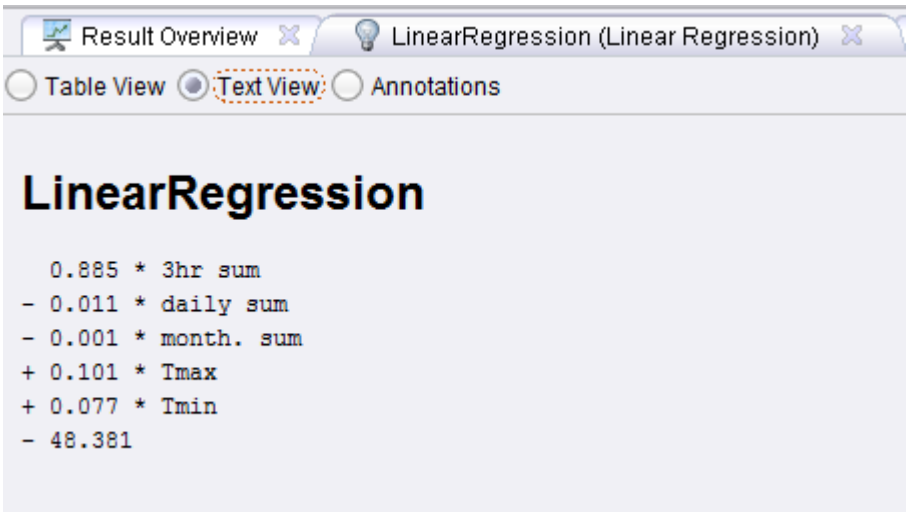
Row No.	prediction(Target-max 120)	3hr sum	daily sum	month. sum	Tmax	Tmin
1	22.513	23.984	44.476	223.110	283.992	281.018
2	39.429	44.937	201.934	554.240	285.682	283.549
3	128.858	147.092	383.556	893.758	293.674	290.086
4	96.364	107.779	177.664	418.732	293.780	285.619
5	80.670	89.971	92.561	321.034	289.254	278.996
6	52.382	59.489	172.816	337.192	286.217	277.961
7	89.176	102.898	322.450	721.651	287.626	280.189
8	84.028	93.736	151.638	321.585	291.424	285.301
9	96.064	108.107	237.418	356.194	294.651	285.274
10	91.531	107.317	519.644	898.197	293.449	283.012
11	90.350	102.004	185.376	542.004	291.091	279.438
12	48.110	52.571	65.465	352.084	288.754	282.751
13	110.503	126.699	386.046	886.728	291.903	288.703
14	152.428	172.450	214.009	967.013	290.715	283.942
15	45.100	48.615	75.569	582.208	293.849	285.603
16	81.853	92.514	231.129	314.484	289.955	284.817
17	42.573	46.846	79.181	145.258	285.731	281.139
18	26.988	28.687	80.047	207.919	289.458	283.118
19	54.141	59.207	91.826	345.552	294.463	281.220



Result Overview | LinearRegression (Linear Regression) | AttributeWeights (Linear R

Table View | Text View | Annotations

Attribute	Coefficient	Std. Error	Std. Coeffici...	Tolerance	t-Stat	p-Value	Code
3hr sum	0.885	0.003	0.672	0.773	320.772	0	****
daily sum	-0.011	0.001	-0.009	0.795	-12.417	0	****
month. sum	-0.001	0.000	-0.000	0.982	-1.612	0.133	
Tmax	0.101	0.010	0.003	0.991	10.306	0	****
Tmin	0.077	0.008	0.002	0.999	9.176	0	****
(Intercept)	-48.381	3.658	?	?	-13.225	0	****



Michel Janos

01/2015