Relatório do Exercício I Pratico Obrigatório

Programa SAS para Proc Reg

Resultado igual ao Excel o L. Office Calc

Sempre será um resultado errado, por que proc reg tem pressuposições difíceis de atingir, como ausência de outliers e distribuição normal.

Todos os dados estão em percentagem.

```
data clientes;
input B_Unit Sales Precific Qu_level
    Claims NPS Satisfac;
datalines;
1 65.98 97.8 96.77 13.58 98.9 97.??
2 15.84 98.9 98.39 12.35 97.8 98.91
3 8.89 100 100 11.11 100 100
4 12.46 98.9 95.16 12.35 96.7 96.74
5 80.67 21.98 19.35 100 2.2 21.59
6 32.17 23.08 22.58 97.53 3.3 23.91
```

```
7 23.45 24.18 24.19 96.3 2.75 25
8 89.96 24.18 19.35 95.06 2.2 26.09
9 31.43 64.84 56.45 50.62 65.93 65.22
10 11.23 65.93 51.61 49.38 71.43 66.3
11 77.46 70.33 53.23 46.91 63.74 68.48
12 23.9 68.13 51.61 45.68 61.54 67.39
13 7.4 86.81 80.65 25.93 90.11 86.96
14 0.29 87.91 79.03 24.69 85.71 85.87
15 83.42 87.91 77.42 22.22 90.11 88.04
16 100 86.81 75.81 25.93 84.62 84.78
17 15.84 98.9 98.39 12.35 97.8 28.91
;
proc print;
run;
/* input B_Unit Sales Precific Qu_level
        Claims NPS Satisfac; */;
title "Rodando proc reg, resultado igual ao
Excel";
proc reg;
  model Satisfac = Sales Precific Qu_level
```

Claims NPS;

run;

Resultado

Estimaciones de parámetro					
Variable	DF	Estimación de parámetros	Error estándar	t valor	Pr > |t|
Intercept	1	133.29163	573.35854	0.23	0.8204
Sales	1	0.04255	0.15554	0.27	0.7895
Precific	1	-0.40787	5.46122	-0.07	0.9418
Qu_level	1	-0.45457	0.97682	-0.47	0.6508
Claims	1	-0.96649	4.65783	-0.21	0.8394
NPS	1	0.49011	1.25901	0.39	0.7045

Podemos observar que nenhum p valor (Pr>/t/) é menor que 0,05, assim nenhuma variável preditora impacta na satisfação do cliente.

Falaríamos de forma equivocada para Dona Luiza que nenhuma variável preditora (Sales, Presific, etc.) impacta na satisfação do cliente.

Erro produzido por rodar proc reg, o que sempre funciona é rodar proc robustreg, faremos uma regressão robusta múltipla, que pertence a ciência de dados ou estatística robusta.

Programa para rodar robustreg

```
/* input B_Unit Sales Precific Qu_level
        Claims NPS Satisfac; */;
title "Rodando proc robustreg, resultado muito melhor ao do Excel";
proc robustreg;
   model Satisfac = Sales Precific Qu_level
        Claims NPS / diagnostics;
run;
```

Resultado do programa robustreg

Estimaciones de parámetro							
Parámetro	DF	Estimación	Error estándar	Límites de confianza al 95%		Chi-cuadrado	Pr > ChiSq
Intercept	1	56.9427	25.9621	6.0579	107.8275	4.81	0.0283
Sales	1	0.0020	0.0070	-0.0118	0.0158	0.08	0.7789
Precific	1	0.2437	0.2473	-0.2410	0.7283	0.97	0.3245
Qu_level	1	0.1104	0.0442	0.0237	0.1970	6.22	0.0126
Claims	1	-0.4243	0.2109	-0.8377	-0.0109	4.05	0.0442
NPS	1	0.1180	0.0570	0.0062	0.2297	4.28	0.0385
Scale	1	0.8244					

Podemos observar que 3 variáveis preditoras foram significativas (p valor < 0,05)

São Qu_level (nível de qualidade dos produtos) p valor 0,0126 ou nas empresas é melhor falar que a confiança para afirmar que essa variável preditora impacta na satisfação do cliente é confiança=(1-0,0126)*100 = 98,7 %. A confiança tem que ser maior que 95% para a variável ter impacto, o que é a mesma coisa que dizer: a margem de erro (p valor) foi menor que 0,05.

Agora o nível de qualidade teve como coeficiente de regressão o valor 0,1104 assim cada unidade a mais no nível de qualidade aumenta 0,1104 a satisfação do cliente.

A segunda variável preditora que impactou a satisfação foi Claims (reclamações), com uma probabilidade, confiança de =(1-0,0442)*100= 95,6%.

O coeficiente de regressão foi -0,4243,

assim para cada uma unidade de aumento das reclamações temos uma perda de 0,4243% na satisfação do cliente.

A terceira variável significativa estatisticamente foi NPS (se o cliente recomendaria a empresa para um amigo ou parente). Com uma margem de confiança de =(1-0,0385)*100=

96%. O coeficiente de regressão foi 0,1180, assim para cada uma unidade de aumento das reclamações temos um aumento na satisfação do cliente de 0,1180%.

Assim a variável mais importante foi Reclamações, coeficiente de regressão -0,4243, foi 4 vezes mais importante que nível de qualidade (0,1104) e NPS (0,1180). Levamos em consideração o valor absoluto.

Falaríamos para Dona Luiza que deve trabalhar prioritariamente as Reclamações, porque o valor absoluto do coeficiente de regressão das reclamações foi 4 vezes maior que o de nível de qualidade e NPS. Poderia utilizar Diagrama de Pareto para priorizar as reclamações.

Diagnósticos		
Obs	Residual robusto estandarizado	Valor extremo
17	-84.0830	*

Podemos observar que o SAS detectou um outlier, é a observação 17. Esse outlier não pode ir para a IA, destrói todos os algoritmos de IA. Podemos eliminar o outlier (trimming ou cutting) ou winsorizar (imputar um valor extremo dentro do contexto dos dados).

Neste casso vamos eliminar o outlier, apagaremos a observação 17. O banco de dados ficara desta forma para a IA:

Bu_Unit	Sales	Price	Qu_level	Claims	NPS
1	65,98	97,8	96,77	13,58	98,9
2	15,84	98,9	98,39	12,35	97,8
3	8,89	100	100	11,11	100
4	12,46	98,9	95,16	12,35	96,7
5	80,67	21,98	19,35	100	2,2

6	32,17	23,08	22,58	97,53	3,3
7	23,45	24,18	24,19	96,3	2,75
8	89,96	24,18	19,35	95,06	2,2
9	31,43	64,84	56,45	50,62	65,93
10	11,23	65,93	51,61	49,38	71,43
11	77,46	70,33	53,23	46,91	63,74
12	23,9	68,13	51,61	45,68	61,54
13	7,4	86,81	80,65	25,93	90,11
14	0,29	87,91	79,03	24,69	85,71
15	83,42	87,91	77,42	22,22	90,11
16	100	86,81	75,81	25,93	84,62

Agora vamos entrar na IA utilizando PCA-Biplot (IAI ou Machine Learning não Supervisionada para Redução de Dimensão e Asociação).

Visualização 6D com IAI Não S para Redução de Dimensão em 2D.

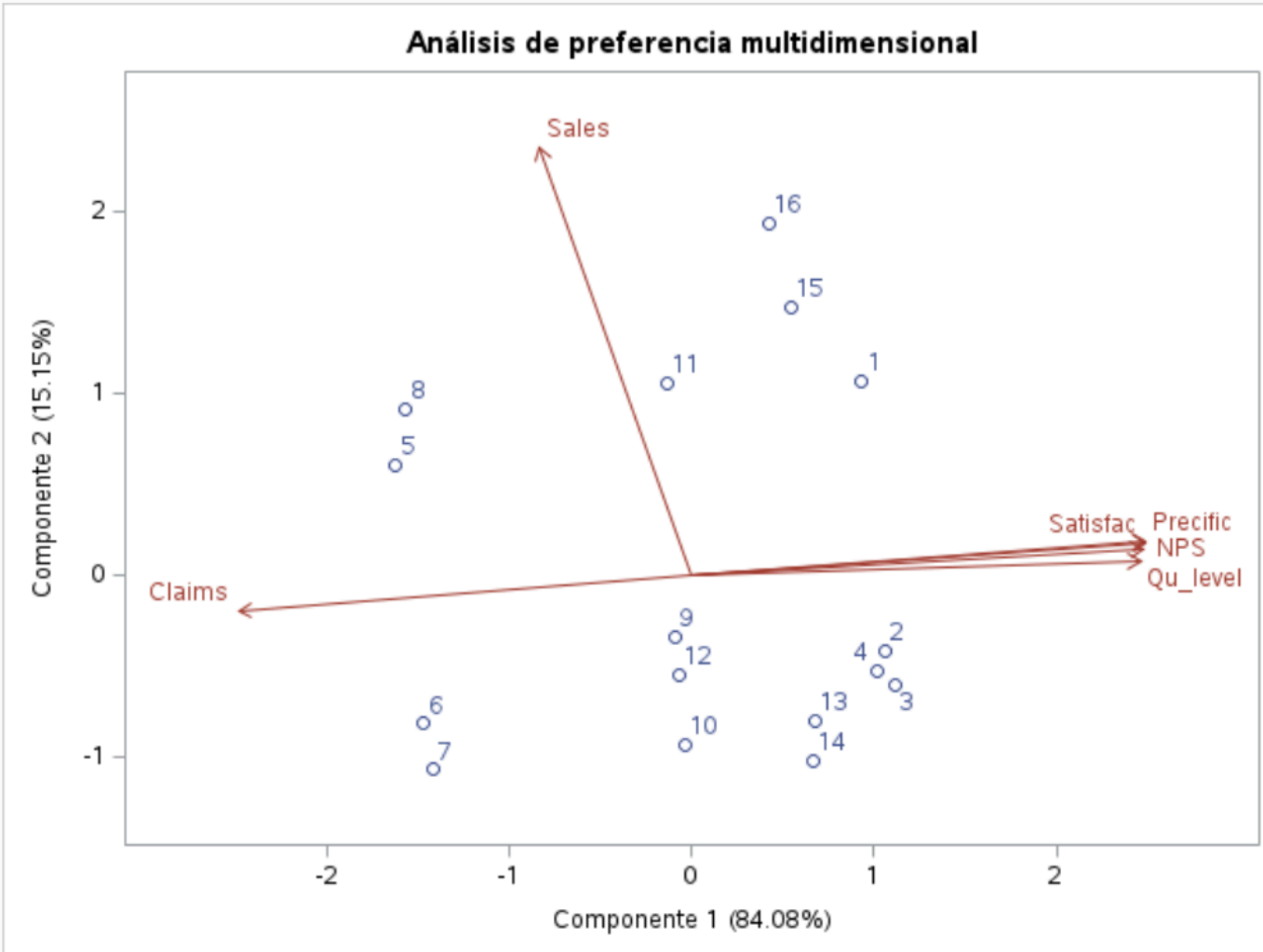
Programa SAS

```
data clientes;
input B_Unit Sales Precific Qu_level
      Claims NPS Satisfac;
datalines;
```

```
1 65.98 97.8 96.77 13.58 98.9 97.83
2 15.84 98.9 98.39 12.35 97.8 98.91
3 8.89 100 100 11.11 100 100
4 12.46 98.9 95.16 12.35 96.7 96.74
5 80.67 21.98 19.35 100 2.2 21.59
6 32.17 23.08 22.58 97.53 3.3 23.91
7 23.45 24.18 24.19 96.3 2.75 25
8 89.96 24.18 19.35 95.06 2.2 26.09
9 31.43 64.84 56.45 50.62 65.93 65.22
10 11.23 65.93 51.61 49.38 71.43 66.3
11 77.46 70.33 53.23 46.91 63.74 68.48
12 23.9 68.13 51.61 45.68 61.54 67.39
13 7.4 86.81 80.65 25.93 90.11 86.96
14 0.29 87.91 79.03 24.69 85.71 85.87
15 83.42 87.91 77.42 22.22 90.11 88.04
16 100 86.81 75.81 25.93 84.62 84.78
;
/* input B_Unit Sales Precific Qu_level
      Claims NPS Satisfac */;
proc prinqual plots=(MDPref);
```

```
   transform identity(Sales Precific Qu_level Claims
NPS Satisfac);
   ods select MDPrefPlot;
run;
```

Resultado do PCA-Boplot



Captamos 84,08% + 15,15%= 99,15% da informação total. Assim perdemos muito pouca informação, se a

informação for maior do que 65% já é um ótimo resultado.

O maior ganho de informação está na relação de Precificação com Satisfação, vemos que o vetor da precificação teve o menor ângulo com a Satisfação do cliente, assim é a variável mais importante na satisfação, isso não foi detectado pela regressão robusta (a melhor).

Assim a IA permite visualizar o fenômeno graficamente e detectar relações que não foram detectadas pela estatística, as duas ciências são sinérgicas e completares.

Se não tivéssemos feito a IA falaríamos para a Dona Luiza que a precificação não é importante para o cliente, o que é um erro enorme, falar que o cliente não se importa com preço que vai pagar.

Sempre devemos rodar a IA para conferir o resultado da estatística.

O nível de vendas (Sales) podemos observar que forma um ângulo próximo a 90 graus com o vetor da satisfação, indicando que não tem impacto na satisfação.

Vemos que o vetor das reclamações (Caims) apresenta a mesma direção, mas sentido contrário ao vetor da satisfação (ângulo próximo a 180 graus), indicando que quando aumentam as reclamações diminui a satisfação do cliente, isso já tinha sido detectado pela estatística.

Para saber quanto impacta a Precificação na Satisfação devemos rodar regressão parcial, com esta modelo: Satisfação = a + b * Precificação.

Programa SAS para a regressão parcial

```
Title "Progrma para Regressão Parcial";
proc robustreg;
  model Satisfac = Precific / diagnostics;
run;
```

Estimaciones de parámetro							
Parámetro	DF	Estimación	Error estándar	Límites de confianza al 95%		Chi-cuadrado	Pr > ChiSq
Intercept	1	1.1583	0.7362	-0.2845	2.6012	2.48	0.1156
Precific	1	0.9795	0.0098	0.9603	0.9987	9956.12	<.0001
Scale	1	1.3027					

Podemos observar que a precificação impacta a satisfação com 99,99% de confiança e que o coeficiente de regressão foi 0,9794, o maior de todos, assim a Precificação é a variável preditora mais importante de todas e não tinha sido detectada pela estatística antes de fazermos a IA. Quando rodarmos Random Forest veremos que o algoritmo detecta a

Precificação como a principal variável preditora (independente).