REGRESSÃO LOGÍSTICA

É uma técnica recomendada para situações em que a ***variável dependente*** é de ***natureza dicotômica ou binária***. Quanto às ***independentes, tanto podem ser categóricas ou não***.

A regressão logística é um recurso que nos permite **estimar a probabilidade associada à ocorrência** de **determinado evento** em face de um conjunto de variáveis explanatórias.

Características

- Busca estimar a probabilidade da variável dependente assumir um determinado valor em função dos conhecidos de outras variáveis;
- Os **resultados da análise ficam contidos no intervalo de zero a um**.

Aplicação da Regressão Logística

- **Previsão de risco** na área tributária – calcular a probabilidade do contribuinte ser inadimplente o adimplente após o parcelamento de tributos. Dias Filho (2003).
- **Utilizado para classificar** se a empresa encontra-se no grupo de empresas solvente ou insolvente. Matias (2002).
- **Determinar quais características** levam as empresas adotarem o *balanced scorecard*. Wanderley (2002).

Vantagens do Modelo Logístico

- Facilidade para lidar com variáveis independentes categóricas.
- Fornece resultados em termos de probabilidade.
- Facilidade de classificação de indivíduos em categorias.
- Requer pequeno número de suposições.
- Alto grau de confiabilidade.

Regressão Logística
x
Regressão Linear

- Logística: Variável dependente é categórica;
- Linear: Utiliza o método dos mínimos quadrados;
- Logística: Utiliza o método da máxima verossimilhança;

O Modelo Regressão Logística

Função Logística

Na regressão logística, **a probabilidade de ocorrência de um evento pode ser estimada diretamente**. No caso da variável **dependente** *Y* **assumir apenas dois possíveis estados (1 ou 0)** e haver um conjunto de *p* variáveis independentes X_1 , X_2 , ... , X_p, o modelo de regressão logística pode ser escrito da seguinte forma:

$$P(Y=1) = \frac{1}{1+e^{-g(x)}}$$

onde,

$g(x) = B_0 + B_1X_1 + \cdots + B_pX_p$

Os coeficientes

Os coeficientes B_0 , B_1 , ... , B_p **são estimados a partir do conjunto dados, pelo método da máxima verossimilhança**, em que encontra uma combinação de coeficientes que **maximiza a probabilidade** da amostra ter sido observada. Considerando uma certa combinação de coeficientes B_0 , B_1 , ... , B_p e variando os valores de *X*. Observa-se que a **curva logística tem um comportamento probabilístico no formato da letra *S***, o que é uma característica da regressão logística. (Hosmer e Lemeshow ,1989)

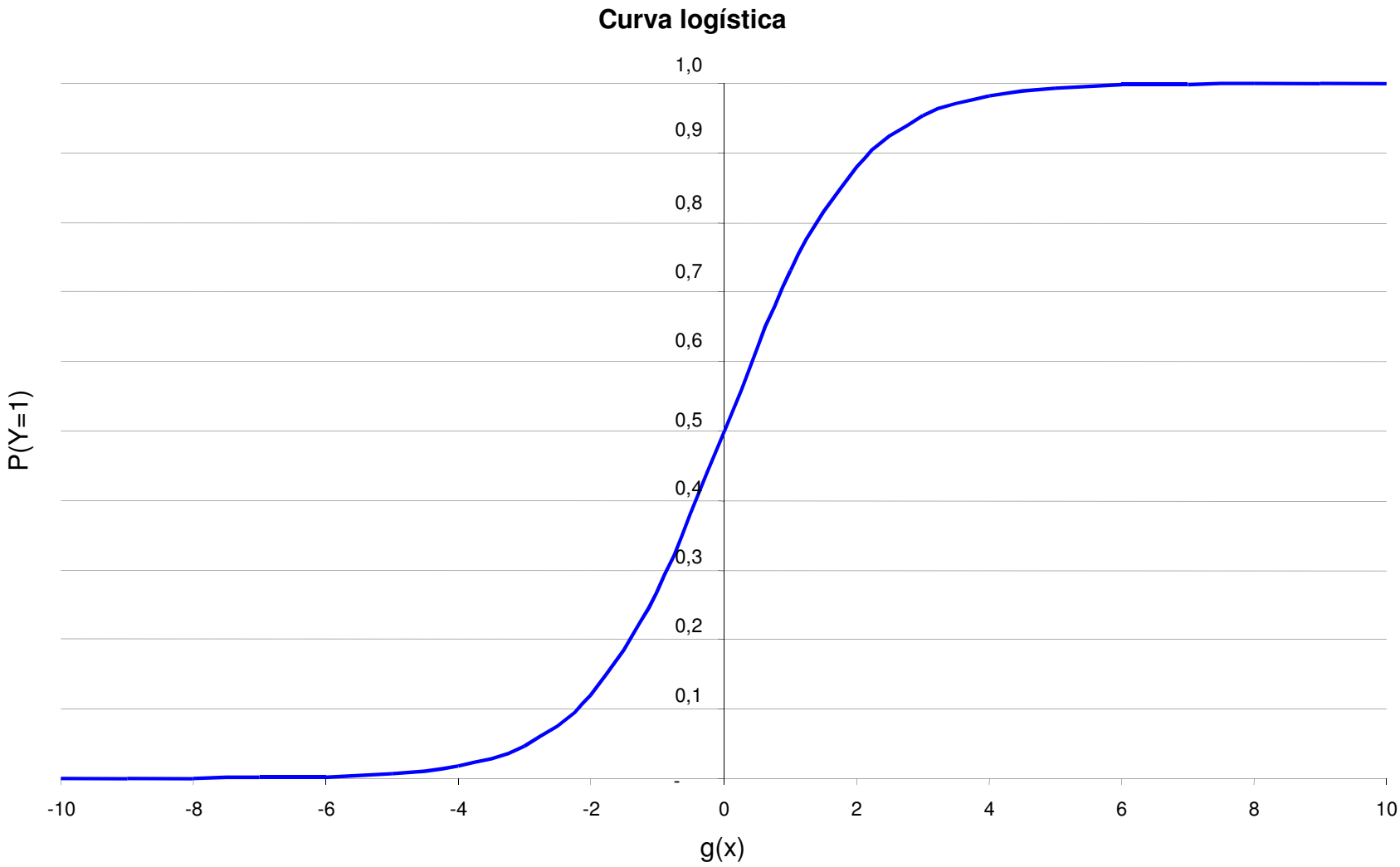
- a) Quando

$$g(x) \to +\infty \quad \text{, então} \quad \mathrm{P(Y=1)} \to 1$$

- b) Quando

$$g(x) \to -\infty \quad \text{, então} \quad \mathrm{P(Y=1)} \to 0$$

Curva da regressão logística



$$g(x) = B_0 + B_1X_1 + \cdots + B_pX_p$$

Interpretação dos Coeficientes

- Observa-se que o impacto de cada coeficiente sobre a própria razão de chances e não mais sobre a quantidade de logit.
- Identifica-se que o impacto do coeficiente da variável independente sobre a razão de chances.
- Determina-se o efeito que os coeficientes exercem sobre a chance de um evento ocorrer.
- Ressalta-se que um coeficiente:
  - positivo aumenta a probabilidade;
  - negativo diminui a probabilidade.

Classificação

- Para utilizar o modelo de regressão logística para discriminação de dois grupos, a regra de classificação é a seguinte:
  - **se *P(Y=1) > 0,5* então classifica-se Y=1**
  - **se *P(Y=1) < 0,5* então classifica-se Y=0**

- Para obter-se uma **boa estimativa da eficiência classificatória do modelo**, recomenda-se separar a amostra em **duas partes**:
  - uma parte para **estimação do modelo**, e
  - outra parte para **testar a eficiência da classificação** (*holdout sample*) (Hair et alii, 1998).

Exemplo de aplicação no SPSS

Uma amostra aleatória de 92 clientes de uma concessionária.

Variáveis: **Status** Inadimplente: 1

Adimplente: 0

Renda Mensal,

Número de dependentes,

Vínculo Empregatício

Com vínculo: 1

Sem vínculo: 0

Construção do Modelo

- Definido **-** as variáveis dependentes e independentes.

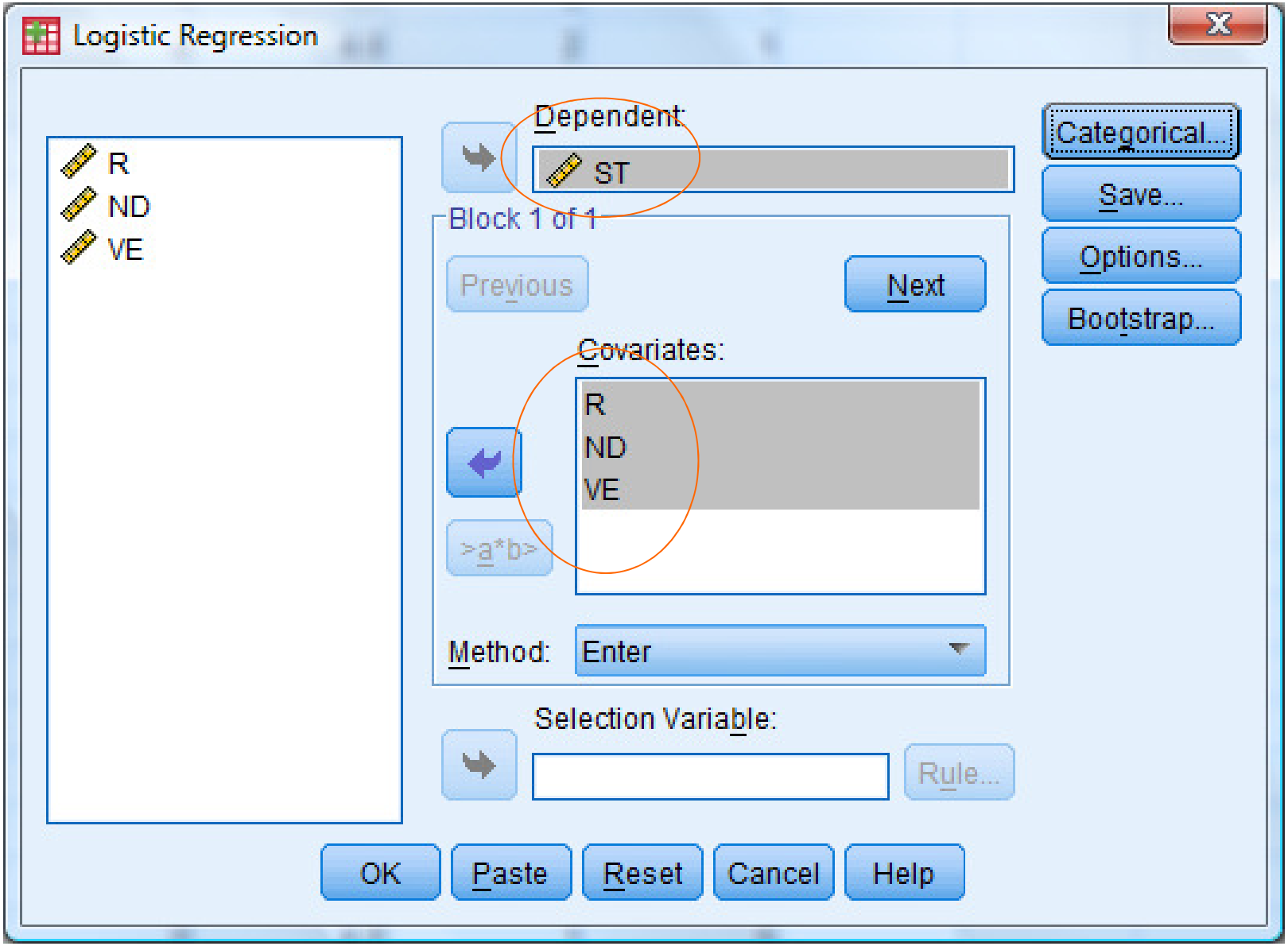
2º Passo

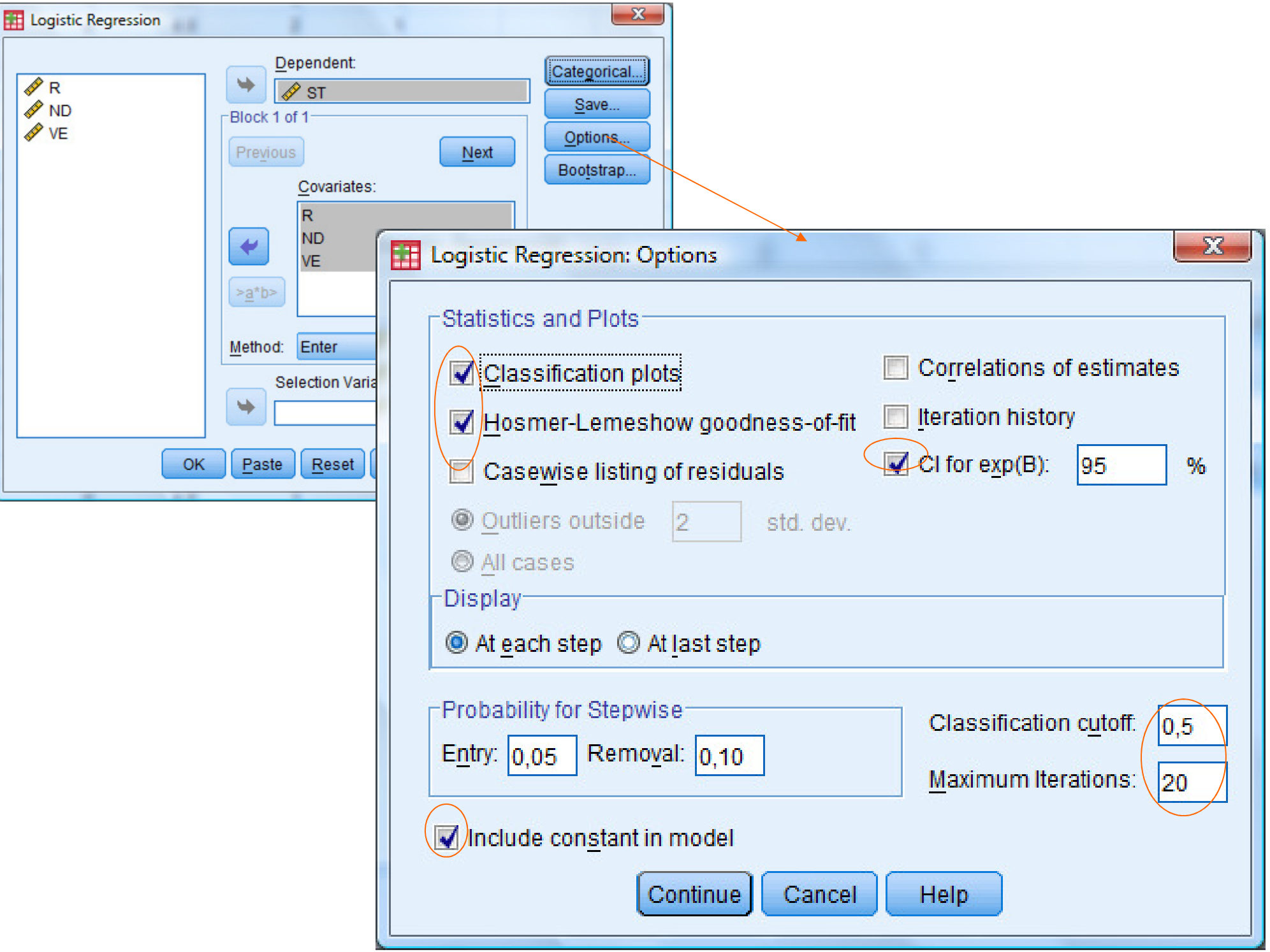
- **Selecionar -** as seguintes opções no SSPS:
- **Analyze**
    - Regression
        - Binary logistic
    - Inserir no campo **Dependent** – Variável dependente “x11”
    - Inserir no campo **Covariates** – Variáveis independentes “x1, x2, x3”
- **Options** (marcar)
    - Classificaçao do plots;
    - Hosmer-Lomeshow goodness-of- fit;
    - Include constant in model;
    - Classification cutoff: 0,5 (seleção do ponto de corte);
    - Maximum Iterations: 20;
    - CI for exp(B): 95% (intervalo de confiança de cada coeficiente estimado);
    - Método ENTER (inclusão simultânea de todas as variáveis independentes).

Base de dados

1	ST	R	ND	VE	ST	R	ND	VE	ST	R	ND	VE	ST	R	ND	VE
2	0	2,5	3	1	0	8,2	5	0	0	1,7	4	0	0	5	3	0
3	1	1,7	3	1	1	1,8	1	1	1	1,7	2	1	1	2,2	3	0
4	0	4	2	1	1	2,5	1	1	1	1,3	3	1	1	1,3	3	1
5	1	2,3	2	1	1	2,2	3	1	0	2,5	1	1	1	1,7	3	1
6	1	3,7	4	0	0	4	1	0	0	3,5	2	0	0	3	2	0
7	0	4,8	1	0	0	4,2	1	0	0	5,6	3	0	0	3	2	1
8	1	1,9	3	0	0	3,7	1	0	0	3,8	2	0	0	3,5	2	1
9	0	5,3	2	1	1	2,4	2	1	0	4	0	0	0	5,8	2	1
10	1	3,1	4	1	1	1,6	3	1	1	2,5	1	1	0	4,8	1	0
11	1	1,9	3	1	1	2	1	1	1	1,2	2	0	1	2,3	3	1
12	1	2,3	4	1	1	2,5	3	1	0	3	1	0	1	2,6	2	1
13	0	3,6	1	0	0	3,8	1	0	0	3	1	0	1	1,8	2	1
14	0	4,7	2	1	0	4,3	2	0	1	2,1	2	1	1	2,9	2	1
15	0	5,8	2	0	1	2	2	1	0	2,5	1	0	0	3,2	1	0
16	0	6	4	0	0	5,2	2	0	0	2,9	1	0	0	4,2	1	0
17	0	3,9	3	1	1	2,4	3	0	0	4	3	0	0	2,6	1	0
18	1	2,4	4	1	0	2,6	4	0	0	3,2	3	0	0	6	1	0
19	1	1,7	4	1	0	1,3	2	1	1	1,2	2	1	1	4,5	3	1
20	0	3,7	2	0	0	3,8	1	1	0	3,5	3	0	1	1,3	2	1
21	0	4,8	1	0	0	4,5	0	1	0	4	1	0	1	2,4	2	1
22	0	3,2	2	1	0	3	0	1	1	2,3	3	1	0	4,3	2	0
23	1	2,7	3	1	1	2,1	2	1	0	2,9	4	0	1	1,8	0	1
24	1	1,2	3	1	1	1,9	2	1	1	2,4	2	1	0	2,4	2	0
25																

Seleção das variáveis




Logistic Regression
R
ND
VE
Dependent:
ST
Block 1 of 1
Previous
Next
Covariates:
R
ND
VE
>a*b>
Method:
Enter
Selection Varia
OK
Paste
Reset
Categorical...
Save...
Options...
Bootstrap...
Logistic Regression: Options
Statistics and Plots
Classification plots
Hosmer-Lemeshow goodness-of-fit
Casewise listing of residuals
Outliers outside
2
std. dev.
All cases
Correlations of estimates
Iteration history
CI for exp(B):
95
%
Display
At each step
At last step
Probability for Stepwise
Entry:
0,05
Removal:
0,10
Classification cutoff:
0,5
Maximum Iterations:
20
Include constant in model
Continue
Cancel
Help

Saídas do SPSS

Case Processing Summary

Unweighted Cases[a]		N	Percent
Selected Cases	Included in Analysis	92	100,0
	Missing Cases	0	,0
	Total	92	100,0
Unselected Cases		0	,0
Total		92	100,0

a. If weight is in effect, see classification table for the total number of cases.

O primeiro relatório evidencia o número de casos incluídos na análise. Conclui-se que todas as observações foram aproveitadas.

Dependent Variable Encoding

Original Value	Internal Value
0	0
1	1

O quadro acima apresenta o código que o software atribui à variável dependente.

Saídas do SPSS (razão de Chances)

Classification Table[a,b]

Observed			Predicted		
			ST		
			0	1	Percentage Correct
Step 0	ST	0	51	0	100,0
		1	41	0	,0
		Overall Percentage			55,4

a. Constant is included in the model.

b. The cut value is ,500

Antes de realizar análise propriamente dita, o SPSS fornece um conjunto de dados que pode ser utilizado para fins de comparação. (Cálculo da razão de chances P/1-P = 51/41)

Saídas do SPSS

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 0	Constant	-,218	,210	1,083	1	,298	,804

Refere-se análise de cada constante no modelo.

Variables not in the Equation

			Score	df	Sig.
Step 0	Variables	R	39,112	1	,000
		ND	7,768	1	,005
		VE	33,368	1	,000
	Overall Statistics		54,573	3	,000

O quadro apresentado acima evidencia o *score* de cada variável, todas apresentam *scores* significativos.

Teste Wald

- Este teste avalia o modelo de regressão Logística como um todo, ***tem como finalidade aferir o grau de significância de cada coeficiente da equação logística, inclusive a constante.***

- Verifica se cada parâmetro ***estimado é significativamente diferente de zero***. (testa a hipótese de que um determinado coeficiente é nulo).

- Segue uma distribuição Qui-quadrado e quando a variável dependente tem um único grau de liberdade, pode-se elevar ao quadrado a razão entre o coeficiente que está sendo testado e o respectivo erro padrão.

Fórmula:

$$Wald = \frac{B_j}{SE_{Bj}}$$

Os coeficientes (*B) são divididos pelo seus respectivos erros padrão(SE).*

Step, Block e Model

Estes testes que têm como objetivo testar a hipótese de que todos os coeficientes da equação são nulos.

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	76,143	3	,000
	Block	76,143	3	,000
	Model	76,143	3	,000

H0: Todos os coeficientes da equação são nulos.

H1: Todos os coeficientes da equação não são nulos.

Todos os três testes têm a mesma finalidade.

Cox-Snell R2

Este teste é comparável ao R-quadrado da regressão linear. Ou seja, indica que **66% das variações ocorridas no Log da razão de chances** são **explicadas** pelo **conjunto** das **variáveis independentes**.

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	24,876	,666	,901

Trata-se de um ***mecanismo que pode ser utilizado para comparar o desempenho de modelos concorrentes***. Entre duas equações logísticas igualmente válidas. **Deve-se preferir o que apresente o COX-SNELL R2 mais elevado**.

Nagelkerke R2

Sua a finalidade é a mesma do cox-snell R2 Na prática a única diferença está em se fazer mais compreensível que o cox-snell R2

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	50,307[a]	,563	,754

a. Estimation terminated at iteration number 6 because parameter estimates changed by less than ,001.

O **Nagelkerke R2** é uma versão do Cox e Snell adaptada para fornecer resultados entre 0 e 1. Conclui-se que o modelo é ***capaz de explicar cerca de 75,4% das variações registradas na variável dependente***

Teste Hosmer e Lemeshow

Este teste mede o grau de acurácia do modelo logístico, este indicador corresponde a um teste do qui-quadrado que consiste em dividir o número de observações em cercas de dez classes e, em seguida, comparar as freqüências preditas com as observadas. A finalidade desse teste é **verificar se existem diferenças significativas entre as classificações realizadas pelo modelo e a realidade observada**. Busca-se não rejeitar a hipótese de que não existem diferenças entre os valores preditos e observados.

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	8,169	8	,417

H0: não há diferenças significativas entre os resultados os preditos pelo modelo e os observados.

H1: há diferenças significativas entre os resultados os preditos pelo modelo e os observados.

Quadro de Classificação Final

Observa-se que o modelo apresenta-se uma classificação satisfatória, quanto ao uso de variáveis independentes como estimadores do status que o cliente poderá assumir em determinadas circunstâncias.

Classification Table[a]

Observed			Predicted		
			ST		
			0	1	Percentage Correct
Step 1	ST	0	45	6	88,2
		1	4	37	90,2
		Overall Percentage			89,1

a. The cut value is ,500

Teste Wald

- Este teste avalia o modelo de regressão Logística como um todo, ***tem como finalidade aferir o grau de significância de cada coeficiente da equação logística, inclusive a constante.***

- Verifica se cada parâmetro ***estimado é significativamente diferente de zero***. (testa a hipótese de que um determinado coeficiente é nulo).

- O teste de Wald é usado para analisar a significância, exceto nos caso em que o coeficiente é extremamente grande.

Coeficientes e Testes

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	R	-1,882	,489	14,845	1	,000	,152	,058	,397
	ND	,860	,386	4,965	1	,026	2,362	1,109	5,031
	VE	2,822	,852	10,969	1	,001	16,812	3,165	89,317
	Constant	1,478	1,657	,795	1	,373	4,383		

a. Variable(s) entered on step 1: R, ND, VE.

H0: os coeficientes são iguais a 0

H1: os coeficientes não são iguais a 0

REGRESSÃO LOGÍSTICA

ANÁLISE DISCRIMINANTE

Semelhança

- ***Ambas se enquadram na classe de métodos estatísticos multivariados***, pois relacionam um conjunto de variáveis independentes com uma variável dependente categórica. (Hair et alii, 1998; Sharma, 1996; Morgan e Griego, 1998).
- São técnicas utilizadas para ***classificação e discriminação de grupos,*** *e*m muitas situações práticas, pesquisadores desejam ***separar duas classes de objetos*** ou alocar um novo objeto em uma dessas classes;
- ***Ambas procuram encontrar uma função*** ou um conjunto de funções **que discrimine os grupos** definidos pela variável categórica ***visando minimizar erros de classificação.***

Diferença

- Em um contexto onde o conjunto de **variáveis independentes** possui um **comportamento** probabilístico de **normalidade** multivariada, a análise discriminante é ótima porque minimiza os erros de classificação (Hair et alii, 1998; Sharma, 1996).

- O modelo logístico pode ser utilizado de uma forma bem mais geral, pois **não faz suposições quanto a forma funcional das variáveis independentes** e, além disso, o número de parâmetros envolvidos no processo de estimação será provavelmente menor.

Regressão Logística x Análise Discriminante

- Comparando as duas técnicas, é consenso que a discriminação logística deve ser preferida quando as distribuições são claramente não-normais. Krzanowski (1988) Press e Wilson (1978). Hair et alii (1998)

- Motivos que levariam o pesquisador a optar pela regressão logística:
  - não é necessário supor normalidade multivariada;
  - é uma técnica mais genérica e mais robusta, pois sua aplicação é apropriada numa grande variedade de situações;
  - é uma técnica similar a regressão linear múltipla.