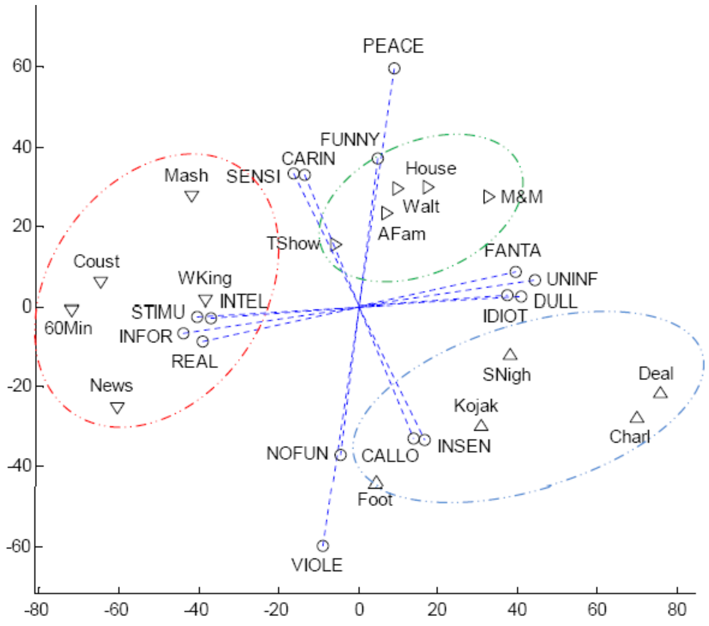


BIG DATA
Parte I - Borra

Statistica Multivariata I

Simone Borra

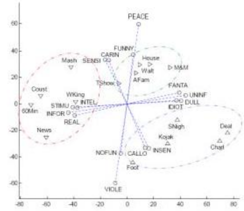
Introduzione e obiettivi

La statistica multivariata si occupa di analizzare e studiare in modo simultaneo un set di *k* **variabili** su un campione di *n* **unità**:

– **simultaneo** indica che lo studio è congiunto, non una variabile alla volta;
– **set di k variabili**: in linea di massima *k* è inteso maggiore o uguale a 3 – nel caso di 2 variabili si utilizzano modelli di correlazione o di regressione semplice.

Obiettivi:

– individuare e misurare i **legami** tra le variabili;
– **segmentare** (raggruppare) le unità in sottoinsiemi simili;
– ricercare **regolarità o tendenze** nei dati;
– definire le **gerarchie** tra variabili.

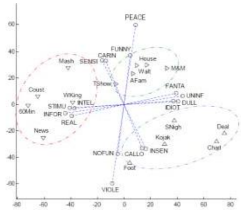


Indice

I modelli e le tecniche che tratteremo sono:

Modelli inferenziali per l’analisi della dipendenza (I)
- modelli di regressione multipla;
- modelli di analisi della varianza.

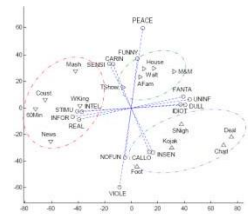
Tecniche esploratorie di riduzione dimensionale (II)
- riduzione variabili: analisi in componenti principali;
- riduzione unità: analisi dei gruppi (*cluster analysis*).



La struttura dati

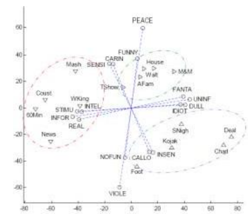
- Il set di dati è solitamente molto “ricco”, caratterizzato da un numero ampio di variabili.

- In linea generale la struttura dati può presentare variabili di diversa natura: quantitative o qualitative.



Esempio dati

UNITA'	PREZZO	CV	SICUREZZA	CONFORT	CONSUMO
MAZDA 3	17.000	110	4	3	8,0
MEGANE	16.500	115	5	4	7,5
Nissan 350 Z	32.000	180	4	3	11,0
Peugeot 307	16.900	105	4	4	7,6
OCTAVIA	21.000	130	4	4	8,4
ALFA 147	16.800	105	4	4	8,1
AUDI A3	20.700	102	4	4	6,9



Regressione lineare multipla

Scopo generale della regressione multipla è quello di studiare la relazione esistente tra una **variabile dipendente** e ***k* variabili indipendenti**, o esplicative. In termini formali:

$$y_i = f(x_{i1}, x_{i2}, \ldots, x_{ik}) + w_i$$

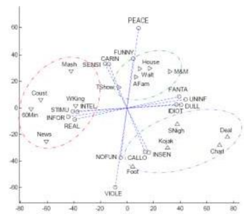
Se la regressione è di tipo lineare si ottiene:

$$y_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + \ldots + b_k x_{ik} + w_i$$

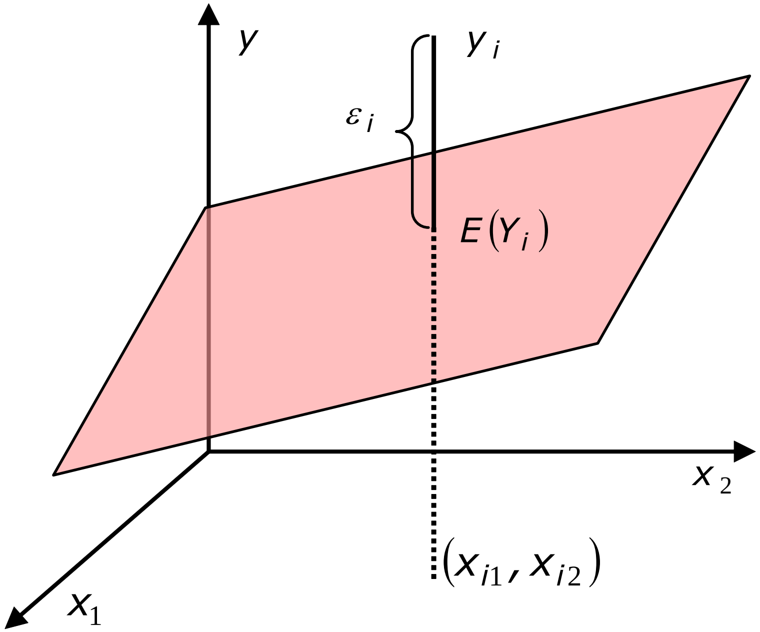
Equazione — *Assunzioni*

$E(y_i \mid x_{i1},\ldots,x_{ik}) = \mu_i = b_0 + b_1 x_{i1} + \ldots + b_k x_{ik}$ — $y_i \mid \mathbf{x}_i \sim N(\mu_i, \hat{\sigma}_w^2)$, indip.

L'interpretazione è analoga a quanto visto per la regressione lineare semplice



With two explanatory variables the function describes a plane:



reg: stima e bontà

- la stima di **b** = [b_0,b_1,…,b_k] è calcolata minimizzando

$$\Sigma_i \,(y_i - b_0 - b_1 x_{i1} + \ldots - b_k x_{ik})^2$$

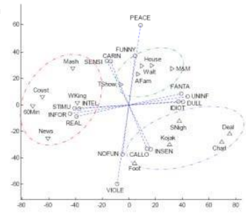
in questo modo otteniamo gli stimatori corretti

$$\hat{\mathbf{b}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

- Questo metodo di stima (detto dei minimi quadrati) coincide con quello di massima verosimiglianza nel caso di errori normalmente distribuiti.
- Lo stimatore corretto della varianza dell'errore è

$$\hat{\sigma}_w^2 = \frac{1}{n-k-1}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

- SST = SSM+SSR $\Rightarrow$ R^2 = SSM/SST = 1 – SSR/SST



Determination coefficient

We note: increasing the number of predictors in the model the value of R^2 rises too

So we conclude: It is ever better a model with more predictors!

R^2 suffers from a serious drawback when it is used for model selection: the inclusion of (possibly irrelevant) additional regressors it always produces an increase in R^2.

Then we use the adjusted determination coefficient:

$$R_c^2 = 1 - \frac{SSR/(n-k-1)}{SST/(n-1)}$$

If SSR decreases, $R^2 = 1 - \frac{SSR}{SST}$ increases,

but If K rises, $SSR/(n-k-1)$ could increases and R_c^2 decreases!

The adjusted det. Coeff. is a balance between the model goodness of fit and the model complexity (in terms of number of predictors)

Example

Some questions about the regression output:

- Is the overall model significant?
- Are the predictors significant?
- How is the fitting of the model?
- What is the interpretation of the regression coefficient for **cor**?
- What happens if we remove the variable feats?
- Does a better linear regression model exist?

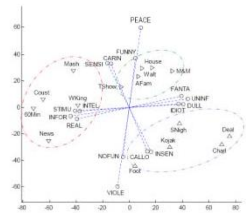
reg: esempio

Number of Observations	67

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	8478759	1695752	43.54	<.0001
Error	61	2375649	38945		
Corrected Total	66	10854409			

Root MSE	197.345	R-Square	0.781
Dependent Mean	1161.463	Adj R-Sq	0.763
Coeff Var	16.99		

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	-478.110	162.092	-2.95	0.005	-802.233	-153.988
age	1	-5.712	2.121	-2.69	0.009	-9.954	-1.470
feats	1	4.936	21.394	0.23	0.818	-37.845	47.717
ne	1	146.402	60.796	2.41	0.019	24.832	267.971
cor	1	191.379	61.859	3.09	0.003	67.684	315.073
sqft	1.000	0.987	0.101	9.770	<.0001	0.785	1.189



reg: test t e F

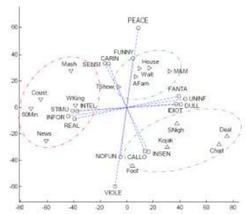
Test t (significatività delle singole variabili)

- Ipotesi nulla H_0: $b_i=0$
- Stat. test t_{oss} = Coef./(Std.Err.) ~ T_{n-k-1} se H_0 vera
- regola rifiuto $| t_{oss} | > t_{\alpha/2}$
- p-value $2\Pr\{T_{n-k-1} > | t_{oss} |\}$
- **Teorema di inversione del test.** Interessante osservare che un coef. è significativo al 100*alfa% se e solo se lo 0 non è contenuto nell’intervallo di confidenza al 100*(1-alfa)%.

Test F (significatività del modello nel complesso)

- Ipotesi nulla H_0: $b_1=b_2=\ldots=b_k=0$
- Stat. Test F_{oss}=(SSM/dfM)/(SSR/dfR) =MSM/MSR

  $F_{oss} \sim F_{k,n-k-1}$ se H_0 vera
- regola rifiuto $F_{oss} > F_{\alpha}$
- p-value $\Pr\{F_{k,n-k-1} > F_{oss}\}$



Coefficients Significance

Model Fitting

Significance of overall model

```
. reg  price age feats ne cor sqft

Source |       SS          df       MS              Number of obs =       67
-------+------------------------------              F(  5,      61) =    43.54
 Model |  8478759.27        5  1695751.85           Prob > F      =   0.0000
 Resid |  2375649.38       61  38945.0718           R-squared     =   0.7811
-------+------------------------------              Adj R-squared =   0.7632
 Total |  10854408.7       66  164460.737           Root MSE      =   197.35

--------------------------------------------------------------------------
 price |      Coef.   Std. Err.      t     P>|t|    [95% Conf.Interval]
-------+------------------------------------------------------------------
   age | -5.712149   2.121296    -2.69    0.009   -9.953942   -1.470356
 feats |  4.935951   21.39437     0.23    0.818   -37.84473    47.71663
    ne |  146.4015   60.79612     2.41    0.019    24.83218    267.9709
   cor |  191.3786   61.85902     3.09    0.003    67.68387    315.0734
  sqft |  .9871056   .1010516     9.77    0.000    .7850405    1.189171
 _cons | -478.1104   162.0919    -2.95    0.005   -802.2331   -153.9876
--------------------------------------------------------------------------
```

Model selection

We could consider **all possible subsets** among the set of X variables and identify a good subset accordig to some criterion (f.e. adjusted R^2)

In most circustances the all-possible-regressions approach is infeasible.
For istance, when there are 10 variables there would be 2^{10}-1=1023 possible regression models!

For three predictors:

(Y, X_1) (Y, X_2) (Y, X_3) (Y, X_4)
$(Y, X_1\ X_2)$ $(Y, X_1\ X_3)$ $(Y, X_1\ X_4)$ $(Y, X_2\ X_3)$ $(Y, X_2\ X_4)$ $(Y, X_3\ X_4)$
$(Y\ X_1\ X_2\ X_3)$ $(Y\ X_1\ X_2\ X_4)$ $(Y\ X_2\ X_3\ X_4)$ $(Y\ X_1\ X_3\ X_4)$
$(Y\ X_1\ X_2\ X_3\ X_4)$

Model selection – automatic search procedure

Backward selection

- Select a significance level to stay in the model (e.g. SLS=0.20, generally .05 is too low, causing too many variables to be removed)
- The procedure begins with the model containing all potential X variables (*full model*) and identifies the predictor with the largest p-value:
- if p-value >SLS, that X variable is dropped.
- The model with k-1 variables is then fitted and the next candidate for dropping is identified.

- The procedure continues until no further X variable can be dropped.

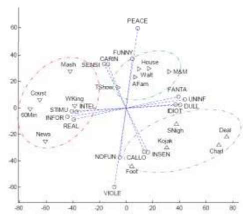
reg: backward elimination

Per selezionare le variabili significative si parte dal modello completo, si elimina la variabile “meno significativa” (i.e. p-value più alto), si stima nuovamente il modello. La procedura termina quando tutte le variabili hanno un p-value inferiore ad una certa soglia.

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	2	8042152	4021076	91.51	<.0001
Error	64	2812257	43942		
Corrected Total	66	10854409			

Root MSE	209.6223	**R-Square**	0.7409
Dependent Mean	1161.463	**Adj R-Sq**	0.7328
Coeff Var	18.04813		

Parameter Estimates							
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**	**95% Confidence Limits**	
Intercept	1	-596.21613	156.4608	-3.81	0.0003	-908.782	-283.65
sqft	1	1.07648	0.10129	10.63	<.0001	0.87413	1.27883
cor	1	215.05618	64.38887	3.34	0.0014	86.42462	343.6877



Multicollinearity

If standard error of $\hat{\beta}_j$ increases, the t-test, $t_{obs} = \hat{\beta}_j \big/ st.err(\hat{\beta}_j)$ decreases, the p-value rises and augments the probability of II error type (Accept null hypothesis when it is false). So many variables are not included in the model.

Consequence of multicollinearity :
many different regression functions providing equally good fits to the data.

X1	X2	Y	Fitted values (a)	Fitted values (b)
2	6	23	23	23
8	9	83	83	83
6	8	63	63	63
10	10	103	103	103

a) $Y=-87+X_1+18X_2$
b) $Y=-7+9X_1+2X_2$
$X_2=5+0.5X_1$

VIF

the **Variance Inflation Factor (VIF)** quantifies the severity of multicollinearity in an OLS regression analysis.
Given K different predictors,
An index of multicollinearity between the variable Xj and the other explanatory variables is:

$$VIF_j = \frac{1}{1-R_j^2}$$

where R_j^2 is the determination coefficient for the model:

$$X_j = \beta_0 + \beta_1 X_1 + \cdots + \beta_{j-1} X_{j-1} + \beta_{j+1} X_{j+1} + \cdots \beta_k X_k + \varepsilon$$

If VIF>5 then multicollinearity is high (corresponding to $R_j^2 = 0.80$)

The **Tollerance** is **1/VIF**

Considering CARS dataset

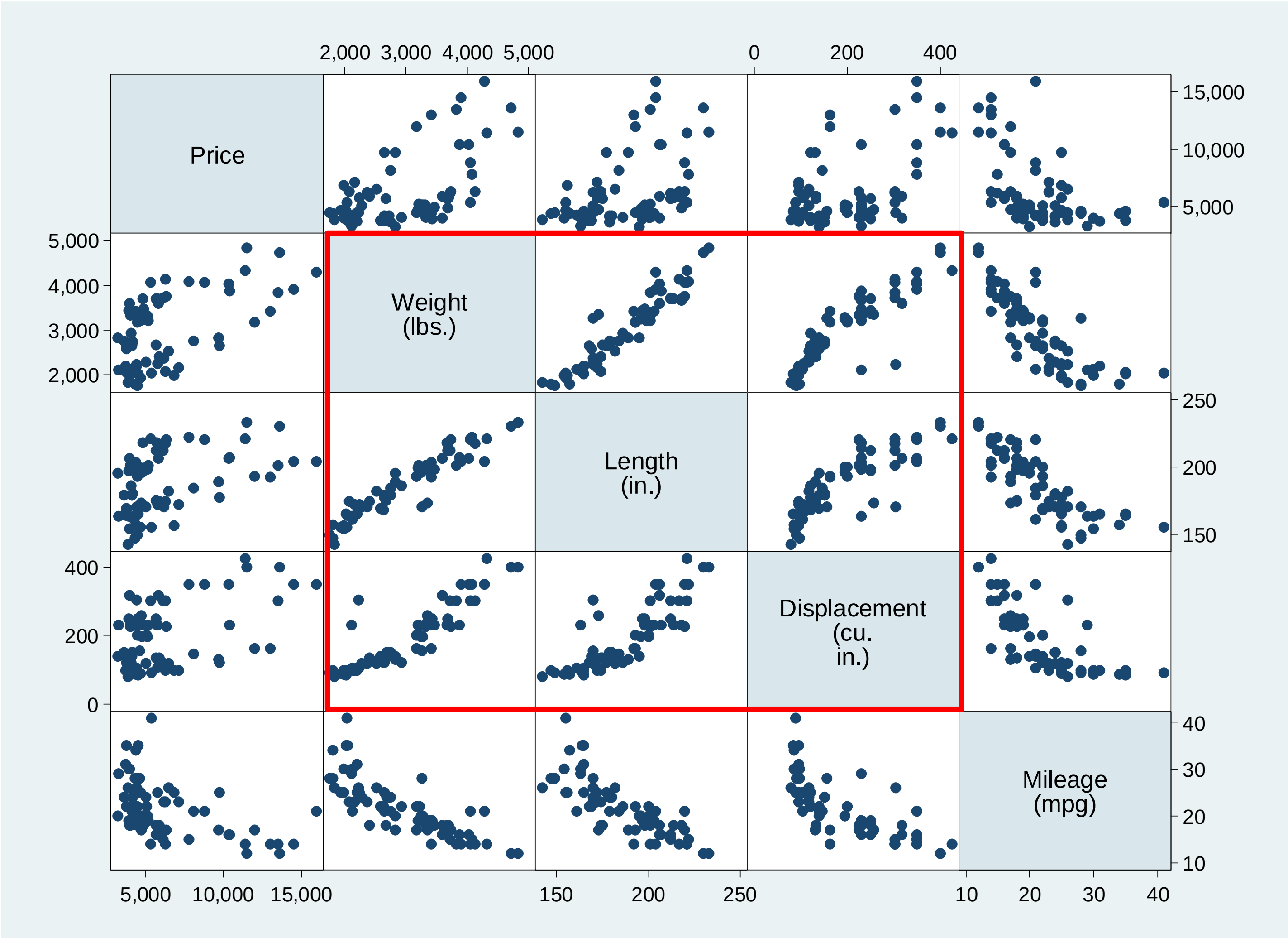
```
. estat vif

    Variable |       VIF       1/VIF
-------------+----------------------
      weight |     21.15    0.047278
      length |     14.35    0.069674
displacement |     10.07    0.099344
        turn |      4.89    0.204495
  gear_ratio |      3.46    0.289314
         mpg |      3.09    0.323292
       trunk |      2.90    0.344269
    headroom |      1.81    0.551859
       rep78 |      1.46    0.685881
-------------+----------------------
    Mean VIF |      7.02
```

Considering CARS dataset (SAS output)

Parameter Estimates										
Variable	**Etichetta**	**DF**	**Parameter Estimate**	**Standard Error**	**Valore t**	**Pr > |t|**	**Tolerance**	**Variance Inflation**	**95% Confidence Limits**	
Intercept	Intercept	1	11063	7305.30297	1.51	0.1353	.	0	-3554.80871	25681
mpg	Mileage (mpg)	1	-114.74282	79.17245	-1.45	0.1526	0.32329	3.09318	-273.16653	43.68089
rep78	Repair Record 1978	1	710.87920	322.11717	2.21	0.0312	0.68588	1.45798	66.32422	1355.43417
headroom	Headroom (in.)	1	-725.63614	416.65992	-1.74	0.0868	0.55186	1.81206	-1559.37072	108.09844
trunk	Trunk space (cu. ft.)	1	70.11347	103.63287	0.68	0.5013	0.34427	2.90471	-137.25542	277.48236
weight	Weight (lbs.)	1	4.03447	1.53188	2.63	0.0108	0.04728	21.15167	0.96919	7.09975
length	Length (in.)	1	-84.38974	43.98228	-1.92	0.0599	0.06967	14.35262	-172.39808	3.61860
turn	Turn Circle (ft.)	1	-207.48012	131.49742	-1.58	0.1200	0.20449	4.89010	-470.60584	55.64560
displacement	Displacement (cu. in.)	1	16.63013	8.99497	1.85	0.0695	0.09934	10.06602	-1.36876	34.62903
gear_ratio	Gear Ratio	1	1642.58757	1061.14969	1.55	0.1270	0.28931	3.45645	-480.76806	3765.94320

Considering CARS dataset



reg: modelli lineari e non lineari

Importante precisare che il modello è lineare nei parametri ma non necessariamente nelle variabili.

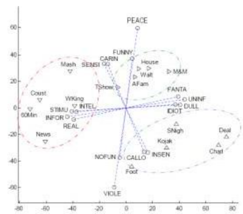
Esempi

$E(y\,|x) = b_0 + b_1 x$ lin. parametri, lin. variabili

$E(y\,|x) = b_0 + b_1 x + b_2 x^2$ lin. parametri, non lin. variabili

$E(y\,|x) = b_0 + b_1 \log(x)$ lin. parametri, non lin. variabili

$E(y\,|x) = b_0 x^{b_1}$ non lin. parametri, non lin. variabili



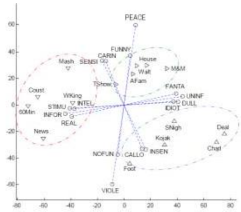
Inserimento di una variabile qualitativa

Esempio. Consideriamo la variabile titolo di studio con modalità: Analfabeta, Licenza, Diploma, Laurea. Questa può essere codificata in 3 variabili dummy scegliendo come riferimento la categoria Analfabeta.

	d_{i1}	d_{i2}	d_{i3}
Licenza	1	0	0
Diploma	0	1	0
Laurea	0	0	1
Diploma	0	1	0
Laurea	0	0	1
Analfabeta	0	0	0

categorie

variabili dummy



Esempio: variabili

Una banca vuole capire la relazione che intercorre tra

Y fattore a 2 livelli: *buen – mal*, se restituisce il prestito rispettando i termini;

e

Cuenta fattore 3 levelli: *no - good running - bad running*, qualità del conto corrente;

Mes durata del prestito in mesi;

Ppag fattore 2 levelli: *pre buen pagador - pre mal pagador*, storia passata del cliente;

Uso fattore 2 livelli *privado - profesional*, utilizzo del prestito;

DM ammontare del prestito in marchi;

Sexo fattore 2 levelli: *mujer - hombre*, sesso;

Estc fattore 2 levelli: *no vive solo - vive solo*, stato civile.

Esempio: dati

y	Frequency	Percent
buen	700	70
mal	300	30

cuenta	Frequency	Percent
bad running	332	33.2
good running	394	39.4
no	274	27.4

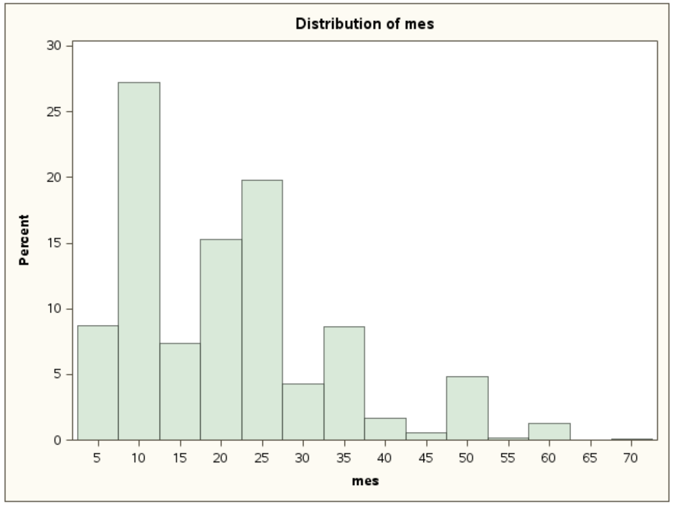
sexo	Frequency	Percent
hombre	598	59.8
mujer	402	40.2

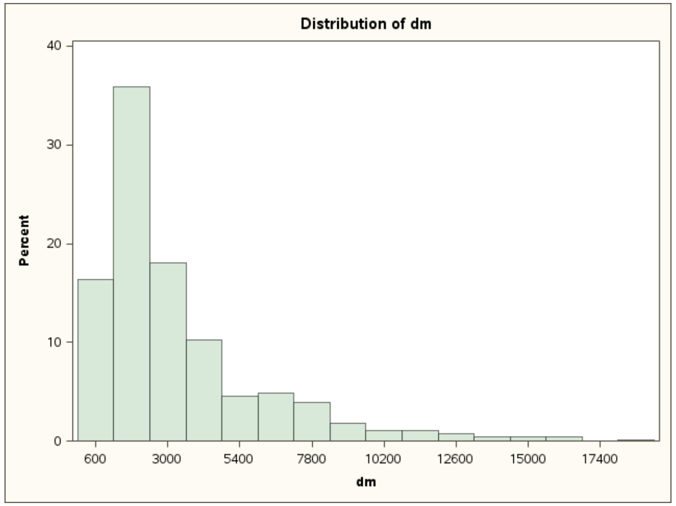
uso	Frequency	Percent
privado	657	65.7
profesional	343	34.3

ppag	Frequency	Percent
pre buen paga	911	91.1
pre mal pagad	89	8.9

estc	Frequency	Percent
no vive solo	640	64
vive solo	360	36

Variable	Mean	Std Dev	Minimum	Maximum	N	Lower Quartile	Median	Upper Quartile
mes	20.9	12.1	4	72	1000	12	18	24
dm	3271.3	2822.8	250	18424	1000	1365	2319.5	3972.5





Esempio: ricodifica variabili

y	Frequency	Percent
buen	700	70
mal	300	30

cuenta	Frequency	Percent
bad running	332	33.2
good running	394	39.4
no	274	27.4

sexo	Frequency	Percent
hombre	598	59.8
mujer	402	40.2

uso	Frequency	Percent
privado	657	65.7
profesional	343	34.3

ppag	Frequency	Percent
pre buen paga	911	91.1
pre mal pagad	89	8.9

estc	Frequency	Percent
no vive solo	640	64
vive solo	360	36

y2	1	mal
	0	buen

cuenta1	1	bad running
	0	otherwise

cuenta2	1	good running
	0	otherwise

estc1	1	no vive solo
	0	vive solo

ppag1	1	pre buen pagador
	0	pre mal pagador

sexo1	1	hombre
	0	mujer

uso1	1	privado
	0	profesional

Modello logit di regressione multipla

Per studiare la relazione che intercorre tra una variabile dicotomica e *k* variabili esplicative possiamo estendere il modello LOGIT visto precedentemente come

Equazione *Assunzioni*

$$\vartheta_i = \frac{\exp(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_K x_{iK})}{1 + \exp(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_K x_{iK})}$$

$y_i | \boldsymbol{x}_i \sim \text{Be}(\vartheta_i)$, indip.

o equivalentemente

$$\text{logit}(\vartheta_i) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_K x_{iK}$$

$y_i | \boldsymbol{x}_i \sim \text{Be}(\vartheta_i)$, indip.

Esempio: stima

Testing Global Null Hypothesis: BETA=0			
Test	**Chi-Square**	**DF**	**Pr > ChiSq**
Likelihood Ratio	204.382	8	<.0001
Score	190.987	8	<.0001
Wald	152.622	8	<.0001

Model Fit Statistics		
Criterion	**Intercept Only**	**Intercept and Covariates**
AIC	1223.729	1035.347
SC	1228.636	1079.517
-2 Log L	1221.729	1017.347

Analysis of Maximum Likelihood Estimates							
Parameter	**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**	**95% Confidence Limits**	
Intercept	1	0.67	0.3199	4.387	0.0362	0.043	1.297
cuenta1	1	-0.6346	0.1764	12.943	0.0003	-0.9804	-0.2889
cuenta2	1	-1.9517	0.2061	89.710	<.0001	-2.3555	-1.5478
dm	1	0.000032	0.000033	0.945	0.3309	-0.00003	0.000098
estc1	1	-0.3854	0.2194	3.087	0.0789	-0.8154	0.0445
mes	1	0.035	0.00785	19.914	<.0001	0.0196	0.0504
ppag1	1	-0.9884	0.253	15.268	<.0001	-1.4841	-0.4926
sexo1	1	-0.2235	0.2208	1.025	0.3115	-0.6563	0.2093
uso1	1	-0.4744	0.1605	8.740	0.0031	-0.7889	-0.1599

logit: test di Wald e LR

Test z (significatività delle singole variabili, può essere utilizzato per la backward selection)

Ipotesi nulla H_0: $b_j=0$

Stat. test $w_{oss}= (\text{Coef./(Std.Err.)})^2$

$w_{oss} \sim \chi_1$ se H_0 vera e *n* suff. elevato

regola rifiuto $w_{oss} > \chi_{1,\alpha}$

p-value $\Pr\{W > w_{oss}\}$

Test LR (significatività del modello nel complesso)

Ipotesi nulla H_0: $b_1=b_2=\ldots=b_k=0$

Stat. Test $LR_{oss}= 2Ll - 2Ll(H_0)$

$LR_{oss} \sim \chi_k$ se H_0 vera e *n* suff. elevato

regola rifiuto $LR_{oss} > \chi_{k,\alpha}$

p-value $\Pr\{\chi_k > LR_{oss}\}$

logit: bontà della classificazione

Per accertare la bontà della classificazione, costruiamo una tabella di contingenza incrociando la *Y* osservata con quella teorica ($\hat{Y} = 1 \text{ se } \hat{\vartheta}_i > \text{cut off}$)

Cut-off = 0.5

	$Y = d$	$Y = nd$	
$\hat{Y} = d$	**0.118**[1]	0.063	0.181
$\hat{Y} = nd$	0.182	**0.637**[1]	0.819
	0.3	0.7	1

Distribuzione congiunta $(\hat{Y}, Y)$

Una prima misura di bontà di classificazione è

1) *Rate of correct classification* (*RCC*). Numero casi correttamente classificati / *n*

Nell'esempio abbiamo RCC = 0.118 + 0.637 = 0.755

logit bc: sensitivity - specificity

Nel caso di eventi rari la RCC può assumere valori elevati anche per modelli poco abili nella classificazione. Se ad es. i default sono solo il 5% un modello banale che predice sempre il non default avrebbe RCC = 0.95. Per evitare questo problema si calcola (evento = default)

2) ***Sensitivity*** (abilità nel predire un *evento* correttamente) quota di *eventi* predetti come *eventi*.

3) ***Specificity*** (abilità nel predire un *non evento* correttamente) quota di *non eventi* predetti come *non eventi*.

cut-off = 0.5

	$Y = d$	$Y = nd$
$\hat{Y} = d$	$\mathbf{0.393}^{2}$	0.090
$\hat{Y} = nd$	0.607	$\mathbf{0.910}^{3}$
	1	1

Tabella profili colonna

Distribuzioni di $\hat{Y} \mid Y$

Sensitivity = 0.118/0.300 = 0.393, Specificity = 0.637/0.700 = 0.910

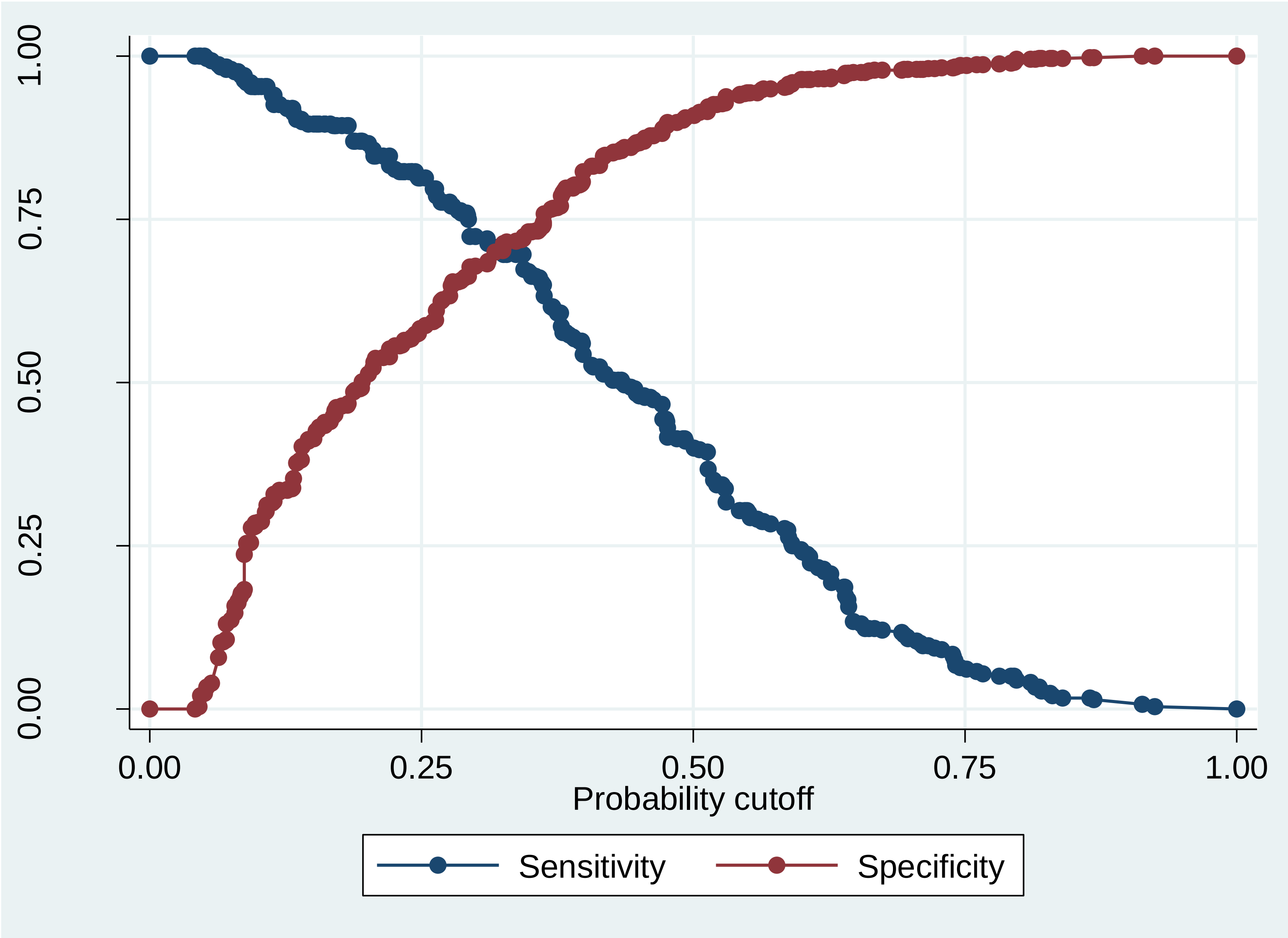
Esempio: classificazione

Classification Table									
Prob Level	Correct		Incorrect		Percentages				
	Event	Non-Event	Event	Non-Event	Correct	Sensi-tivity	Speci-ficity	False POS	False NEG
0.5	113	632	68	187	74.5	37.7	90.3	37.6	22.8

	Y=1	Y=0	
Ystim=1	113	68	181
Ystim=0	187	632	819
	300	700	1000

Tasso corretta classificazione =(113+632)/100=74,5
Sensitivity = (113/300)*100=37,7%
Specificity = (632/700)*100=90,3%

curve ss



1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
Probability cutoff
Sensitivity
Specificity

logit bc: curva ROC

Le variazioni in termini di sensitivity e specificity corrispondenti a diversi livelli di cut-off possono anche essere rappresentate mediante la curva **ROC**.

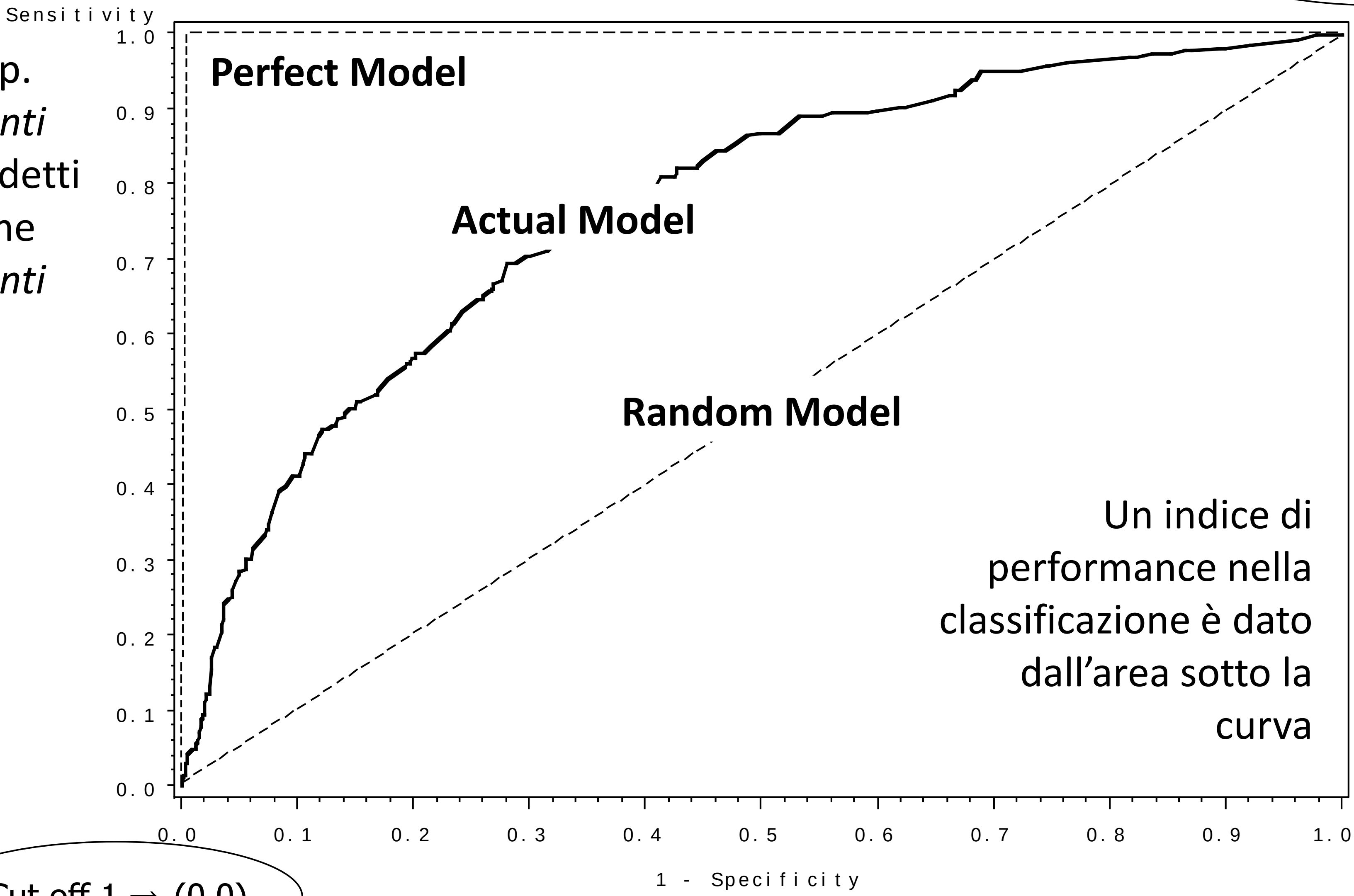
La curva ROC è costituita da tutti i punti di coordinate (sensitivity,1-specificity) che si creano considerando tutti i possibili valori di cut-off tra 0 e 1.

L'area sottesa dalla curva è un indice di bontà della classificazione che varia tra 0 e 1.

La curva ROC è' molto utile per confrontare tra loro diversi modelli di classificazione.

Curva ROC

Cut off 0 ⇒ (1,1)



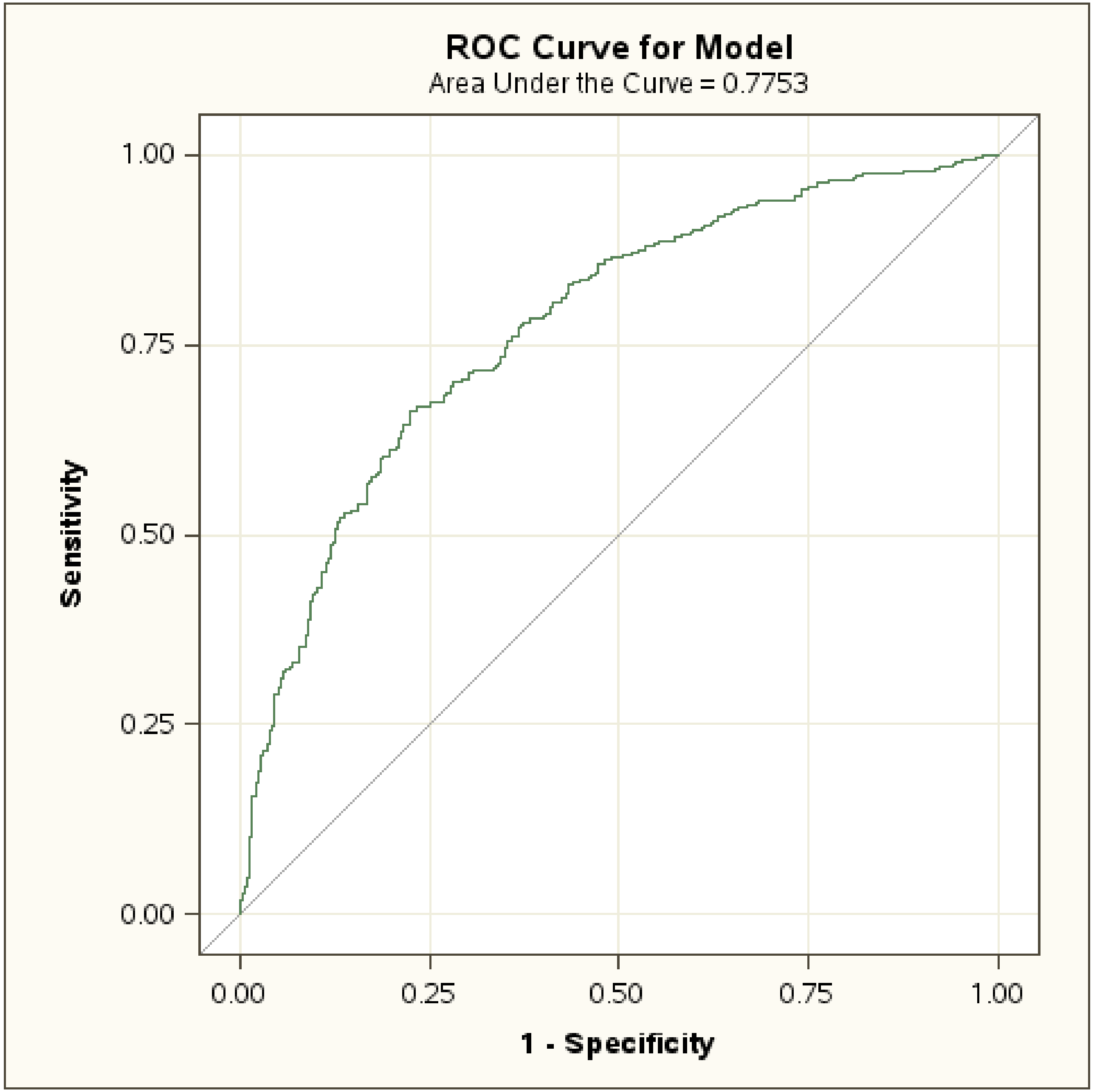
Prop. *eventi* predetti come *eventi*

Un indice di performance nella classificazione è dato dall'area sotto la curva

Cut off 1 ⇒ (0,0)

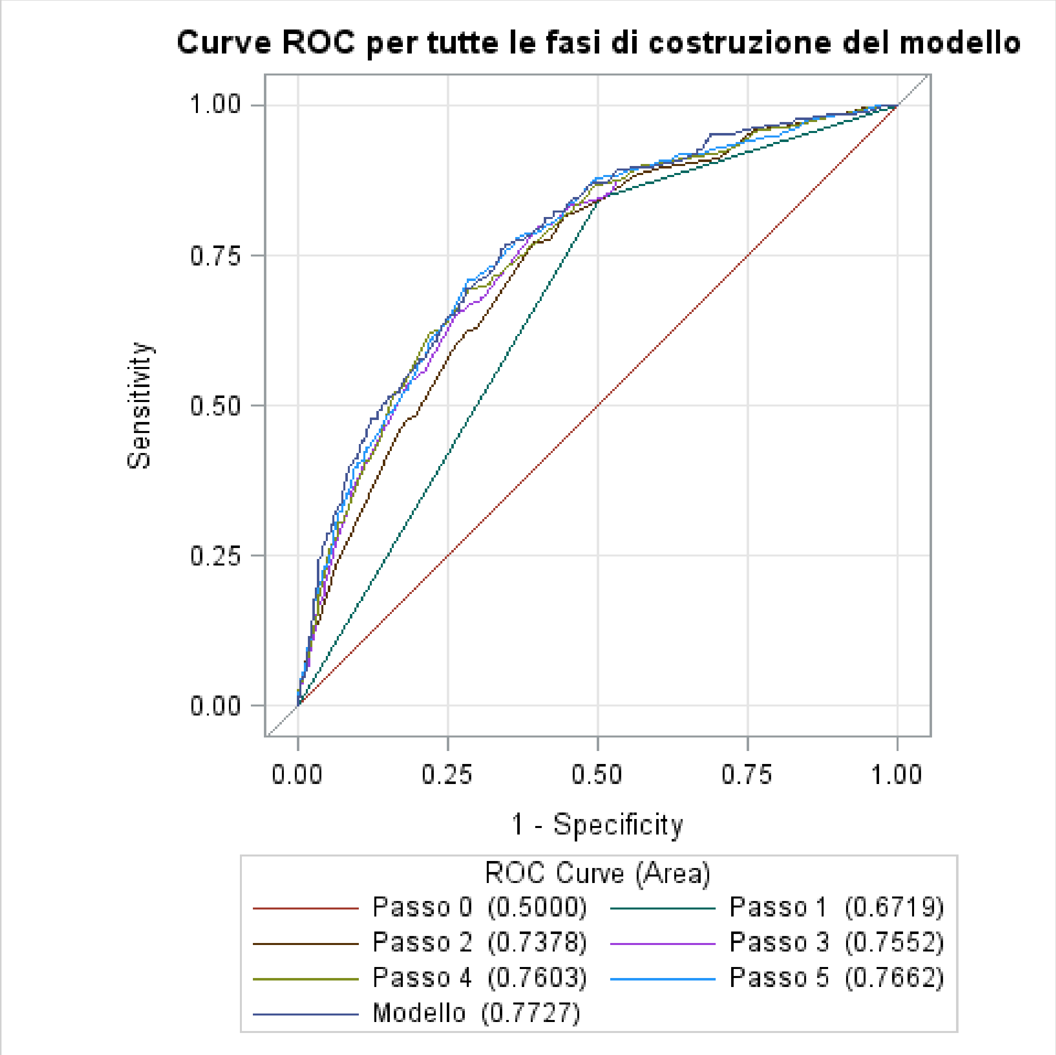
Prop. *non eventi* predetti come *eventi*

Esempio: curva ROC



Esempio: Forward selection

Summary of Forward Selection						
Step	**Effect Entered**	**DF**	**Number In**	**Score Chi-Square**	**Pr > ChiQuadr**	**Variable Label**
1	cuenta2	1	1	103.9648	<.0001	cuenta==good running
2	mes	1	2	40.952	<.0001	mes
3	ppag1	1	3	18.8406	<.0001	ppag==pre buen pagador
4	cuenta1	1	4	11.7336	0.0006	cuenta==bad running
5	estc1	1	5	10.1391	0.0015	estc==no vive solo
6	uso1	1	6	8.706	0.0032	uso==privado



Analysis of Maximum Likelihood Estimates					
Parameter	**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiQuadr**
Intercept	1	0.6437	0.3179	4.0987	0.0429
cuenta1	1	-0.6174	0.1757	12.3417	0.0004
cuenta2	1	-1.9377	0.2055	88.9342	<.0001
estc1	1	-0.5327	0.1591	11.2081	0.0008
mes	1	0.0389	0.00629	38.1125	<.0001
ppag1	1	-0.9877	0.2527	15.2807	<.0001
uso1	1	-0.4694	0.1597	8.6408	0.0033

logit bc: out of sample

Nella trattazione precedente tutte le 5 misure sono state calcolate in sample, i.e. sui dati utilizzati per stimare il modello. Questo comporta una sovrastima dell’abilità del modello nella classificazione.
Si può ovviare al problema spezzando in due parti il campione: training e test. Sulla parte training si stima il modello, le stime così ottenute vengono utilizzate per classificare le osservazioni della parte test del campione e calcolare le 5 misure.