Regressione Logistica

Esercizio 1

Data set:
Nel data set heart.txt (o heart.sav) sono contenute informazioni riguardo 302 pazienti che hanno avuto infarto e 160 che non hanno avuto infarto in uno studio retrospettivo di uomini, i risultati dello studio sono pubblicati in Rousseauw et al, 1983, South African Medical Journal

South African Heart Disease (Rousseauw et al. 1983)

- **Pressione**
- **Fumo** tabacco cumulativo (kg)
- **Colest** low-density lipoprotein (colesterolo)
- **Adipos**
- **Familiarità** Presente/Assente
- **TypeA** type A behaviour (più alto lo score maggiore segnale di stress)
- **Obesità** BMI
- **Alcohol** consumo di alcohol (gr/die)
- **età**
- **Infarto** variabile di risposta

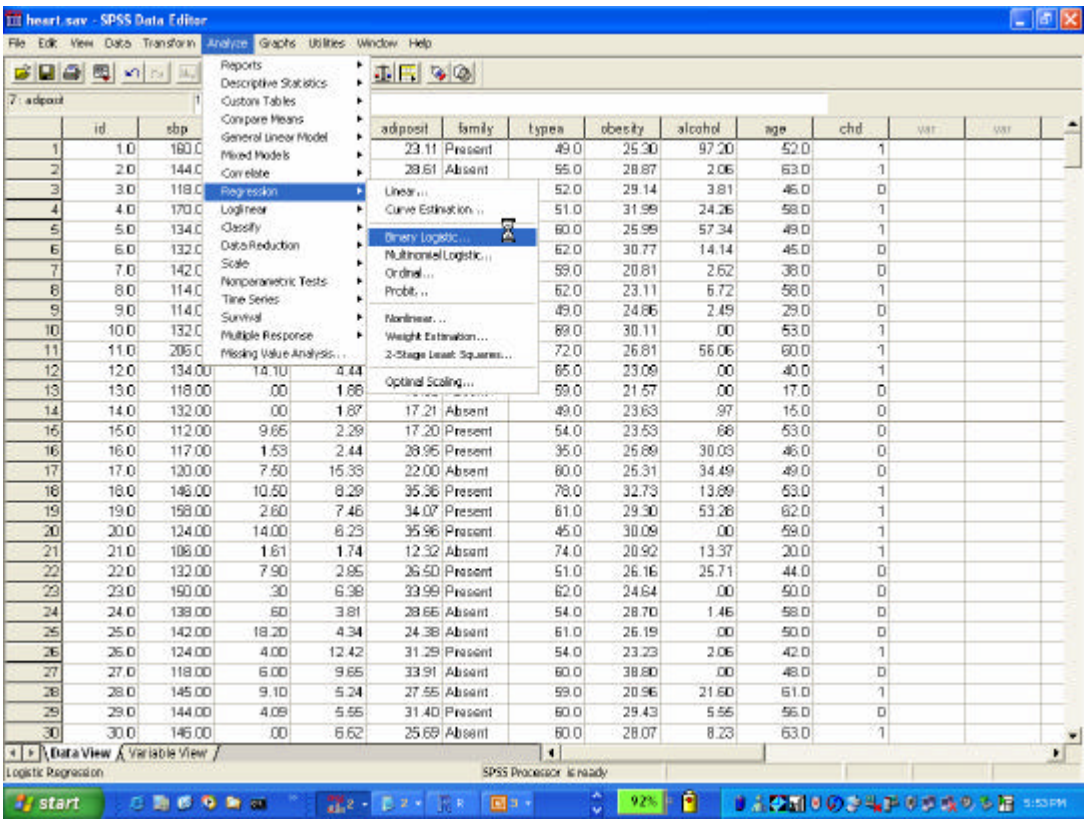
Quesiti
Siamo in grado di predire la presenza o assenza di infarto basandoci sulle variabili disponibili? Possiamo inoltre quantificare il rischio di infarto, ad esempio, di un fumatore rispetto ad un non fumatore?

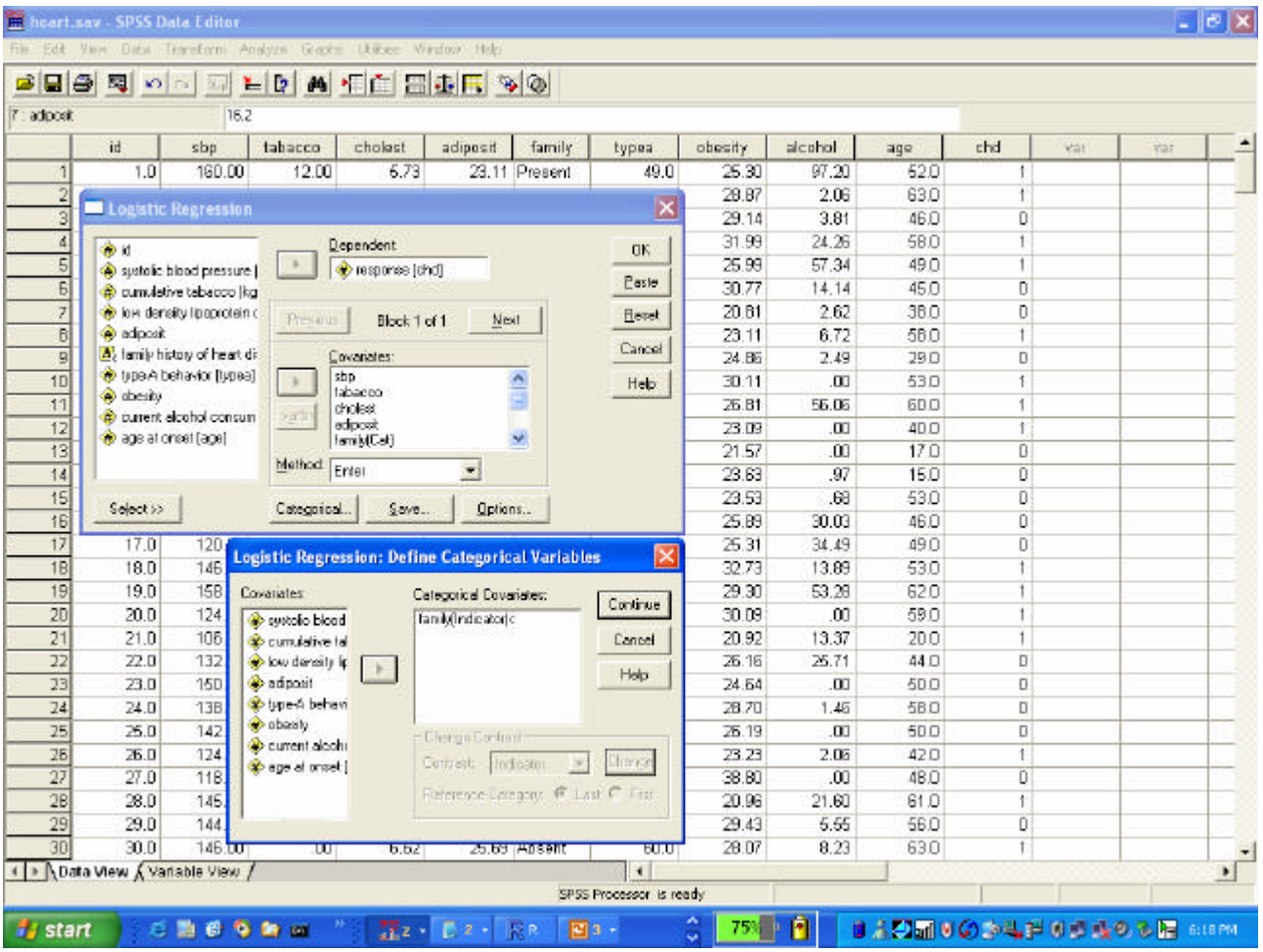
Analisi

Dal menu **Analyse**, selezioniamo **Regression**. SPSS permette di scegliere tra diverse opzioni; volendo stimare un modello di regressione logistica, scegliamo **Binary Regression**

E' necessario ora individuare la variabile dipendente (presenza o assenza di infarto) e le variabili concomitanti (dette con un inglesismo "covariate"; per ora selezioniamo tutte le variabili disponibili).

La variabile 'family' è una variabile qualitativa; selezionando "il bottone" **Categorical** si ha la possibilità di specificare che 'family' sia effettivamente variabile qualitativa (nel caso in cui la variabile sia già classificata come variabile stringa sarà automaticamente considerata qualitativa).


heart.sav - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
Reports
Descriptive Statistics
Custom Tables
Compare Means
General Linear Model
Mixed Models
Correlate
Regression
Loglinear
Classify
Data Reduction
Scale
Nonparametric Tests
Time Series
Survival
Multiple Response
Missing Value Analysis...
Linear...
Curve Estimation...
Binary Logistic...
Multinomial Logistic...
Ordinal...
Probit...
Nonlinear...
Weight Estimation...
2-Stage Least Squares...
Optimal Scaling...
7 : adiposit
id sbp adiposit family typea obesity alcohol age chd
Data View
Variable View
Logistic Regression
SPSS Processor is ready
start



Ci chiediamo preliminarmente quali siano le variabili rilevanti per spiegare il rischio di infarto. L'opzione **Method** permette di scegliere tra 7 possibili metodi di selezione delle variabili. **I** metodo di *default* è **Enter**: in questo modo il modello considera come predittori le variabili presenti nella tabella Covariates. Altri metodi sono:
Forward Conditional, LR, Wald: il metodo forward, cioè "inserimento in avanti", inizia con un modello con solo l'intercetta e aggiunge una ad una le variabili significative. La scelta delle variabili può avvenire tramite Likelihood Ratio, Conditional Likelihood o Statistica di Wald.
Backward Conditional, LR, Wald il metodo backward, cioè "eliminazione all'indietro", inizia con un modello con tutte le variabili e rimuove una ad una le variabili non significative. La scelta delle variabili può avvenire tramite Likelihood Ratio, Conditional Likelihood o Statistica di Wald.

Procediamo selezionando **Method**=**Entry**.

Nell'output è possibile confrontare i risultati relativi a due modelli: il modello con solo intercetta:

Classification Table[a,b]

			Predicted		
			response		Percentage Correct
	Observed		0	1	
Step 0	response	0	302	0	100.0
		1	160	0	.0
	Overall Percentage				65.4

a. Constant is included in the model.

b. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 0	Constant	-.635	.098	42.199	1	.000	.530

e il modello con tutte le variabili presenti:

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	472.140	.235	.325

Classification Table[a]

			Predicted		
			response		Percentage Correct
	Observed		0	1	
Step 1	response	0	256	46	84.8
		1	77	83	51.9
	Overall Percentage				73.4

a. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	SBP	.007	.006	1.288	1	.256	1.007
	TABACCO	.079	.027	8.903	1	.003	1.083
	CHOLEST	.174	.060	8.498	1	.004	1.190
	ADIPOSIT	.019	.029	.403	1	.526	1.019
	FAMILY(1)	.925	.228	16.488	1	.000	2.523
	TYPEA	.040	.012	10.329	1	.001	1.040
	OBESITY	-.063	.044	2.021	1	.155	.939
	ALCOHOL	.000	.004	.001	1	.978	1.000
	AGE	.045	.012	13.901	1	.000	1.046
	Constant	-6.151	1.308	22.103	1	.000	.002

a. Variable(s) entered on step 1: SBP, TABACCO, CHOLEST, ADIPOSIT, FAMILY, TYPEA, OBESITY, ALCOHOL, AGE.

Dalla tabella **Classification Table**, vediamo che il 65.4% delle osservazioni risultano correttamente classificate con un modello con la sola intercetta (quindi un modello che indipendentemente dalle covariate disponibili classifica tutti i soggetti come a rischio di infarto), mentre un modello la cui predizione dipende dalle variabili disponibili classifica correttamente il 73.4% dei soggetti. Dalla tabella **Variables in the Equation** leggiamo nella seconda colonna la stima dei parametri del modello logistico; la terza colonna fornisce una stima del relativo *standard error,* in modo tale che nella sesta colonna leggiamo il *p-value* relativo alla significatività del parametro. Infine nell'ultima colonna leggiamo una stima dell'*odds ratio*.

Nel caso di studio retrospettivi caso/controllo come questo, l'*odds ratio* viene stimato con $\exp(\hat{b})$. L'*odds ratio,* come misura di associazione, approssima il rischio relativo, ossia la probabilità di avere la malattia dato che si è esposti rapportata alla probabilità di avere la malattia dato che NON si è esposti. Tra le variabili disponibili abbiamo una sola variabile di 'esposizione' binaria: 'familiarità/NON familiarità'. Leggiamo dalla tabella che la probabilità di avere un infarto dato che si ha familiarità sulla probabilità di avere l'infarto dato che NON si ha familiarità è 2.523.

Come leggere l'*odds ratio* relativo a variabili di esposizione continue? Leggiamo, ad esempio, che 1.083 è l'*odds ratio* all'aumentare di un unità (kg) del consumo di tabacco.

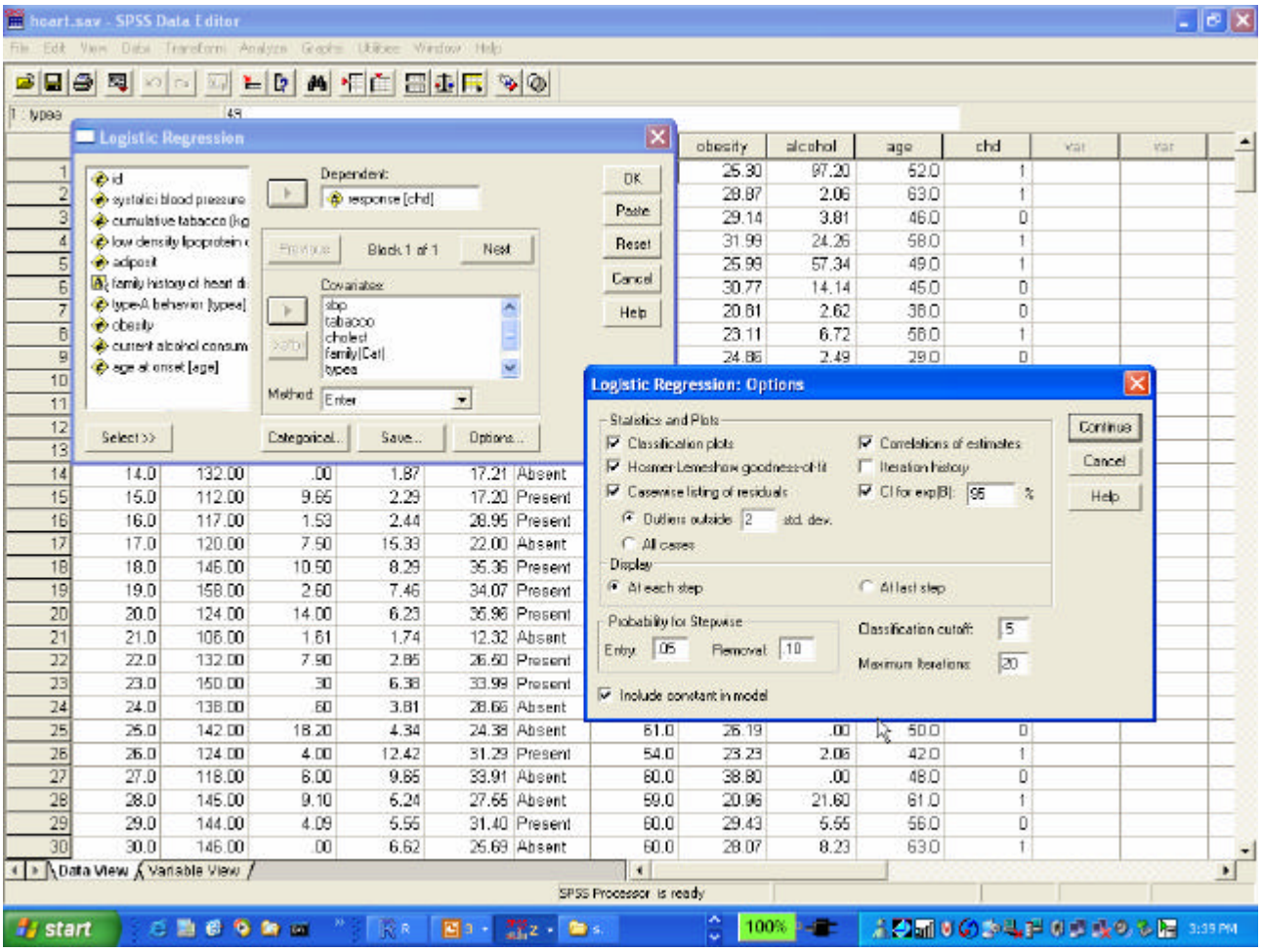
Dalla tabella leggiamo inoltre che alcool non è una variabile significativa, così come adiposità e obesità; è possibile quindi rimuovere tali variabili, ottenendo il seguente modello finale (a cui si arriva anche scegliendo altri metodi di 'entry' delle variabili):

Variables in the Equation

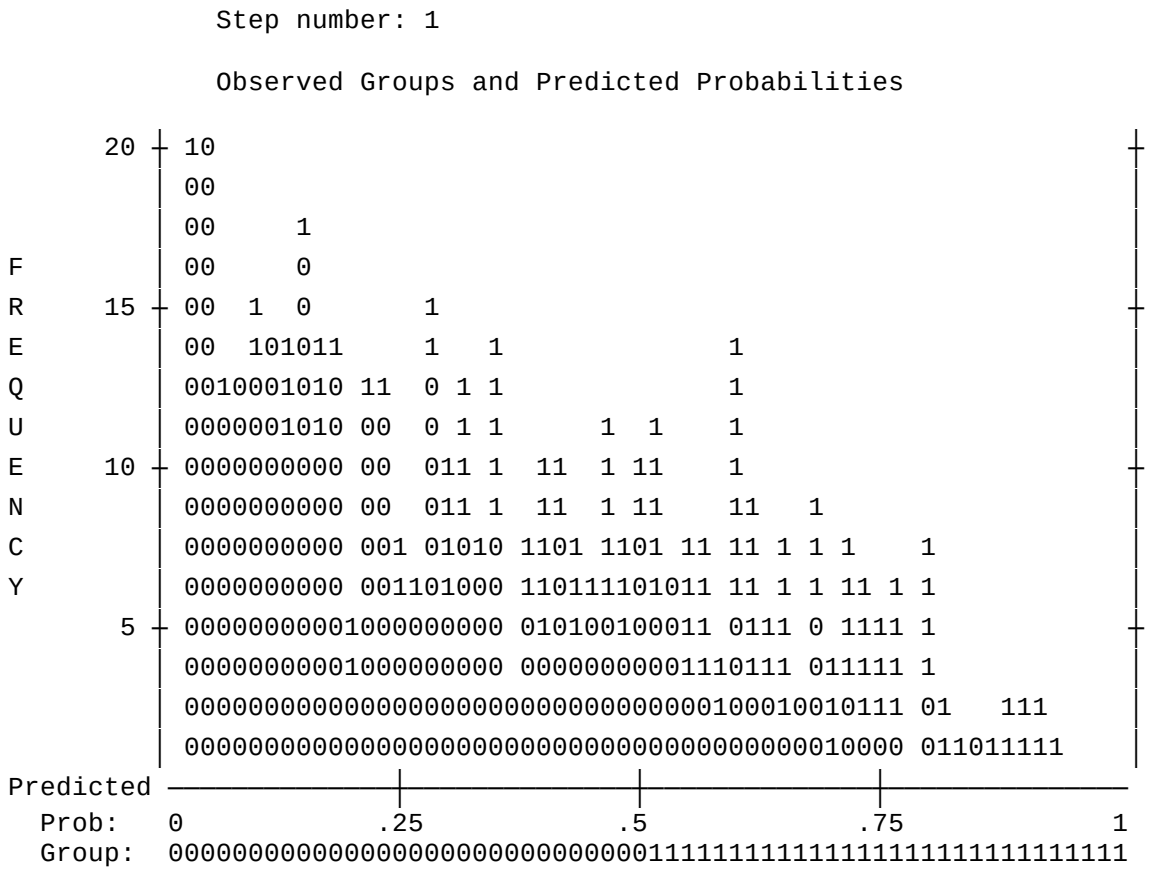
		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	SBP	.006	.006	1.028	1	.311	1.006
	TABACCO	.081	.026	9.582	1	.002	1.084
	CHOLEST	.159	.055	8.340	1	.004	1.172
	FAMILY(1)	.914	.226	16.345	1	.000	2.495
	TYPEA	.038	.012	9.487	1	.002	1.038
	AGE	.048	.011	20.279	1	.000	1.049
	Constant	-7.122	1.145	38.661	1	.000	.001

a. Variable(s) entered on step 1: SBP, TABACCO, CHOLEST, FAMILY, TYPEA, AGE.

Molte opzioni sono possibili utilizzando SPSS per la stima di un modello logistico.



Richiedendo: **Classification Plot** abbiamo il seguente istogramma che mostra il numero di osservazioni correttamente o erroneamente classificate al variare della probabilità predette di avere l'evento.



Richiedendo invece **Hosmer-Lemeshow Goodness of fit** otteniamo la seguente tabella:

Contingency Table for Hosmer and Lemeshow Test

		response = 0		response = 1		
		Observed	Expected	Observed	Expected	Total
Step 1	1	45	44.292	1	1.708	46
	2	42	42.052	4	3.948	46
	3	37	39.623	9	6.377	46
	4	39	36.888	7	9.112	46
	5	34	33.398	12	12.602	46
	6	26	29.922	20	16.078	46
	7	27	25.815	19	20.185	46
	8	25	21.788	21	24.212	46
	9	19	17.194	27	28.806	46
	10	8	11.028	40	36.972	48

Le osservazioni sono divise in decili sulla base della probabilità stimata del verificarsi o meno dell'evento, e un test di confronto tra le osservazioni di evento e non evento e i valori attesi sotto l'ipotesi di un modello che ben predice la variabile di risposta, procede dalla seguente tabella, fornendo il seguente risultato:

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	6.085	8	.638

Accettiamo l'ipotesi che il modello spieghi bene i dati.