Progetto di “Metodi e Modelli Biostatistici per la Ricerca Clinica”, Università di Milano-Bicocca
Millone A. (846588), Rossi S. (857183)

Basso peso alla nascita e caratteristiche socio-sanitarie della madre: un modello predittivo

1 - Caratteristiche dello studio

L’obiettivo dello studio è quello di determinare un modello predittivo per il basso peso alla nascita del bambino, partendo da alcune caratteristiche socio-sanitarie della madre.
Le covariate (caratteristiche della madre) a disposizione erano le seguenti:

- *AGE*: età;
- *LWT*: peso della madre durante il suo ultimo periodo mestruale;
- *RACE*: etnia;
- *SMOKE*: abitudine al fumo;
- *PTL*: parto prematuro;
- *HT*: ipertensione;
- *UI*: presenza di irritabilità uterina;

Oltre alla variabile risposta *LOW* relativa al basso peso del nascituro.

Il modello utilizzato in questo progetto è stato il modello logistico, che ci ha permesso di determinare le stime relative alle singole variabili. Successivamente, per suddividere il dataset in train e validation abbiamo utilizzato l’approccio Split-Sample, il quale divide i dati mantenendo fisse le proporzioni degli eventi e dei non-eventi.
Dopo aver diviso il dataset, abbiamo calcolato le tre principali quantità relative ai modelli predittivi:

- *Misura dell’errore di predizione*: determinata con lo Score di Brier e l’R^2 di Nagelkerke;
- *Discriminazione*: misurata con le curve ROC (e l’AUC ossia l’area sottesa alle stesse) e il Test di DeLong;
- *Calibrazione*: Test di Hosmer-Lemeshow, Integrated Discrimination Improvement e Net Reclassification Improvement (misura dell’aumento della bontà predittiva del modello all’aggiunta di una determinata covariata).

2 - Statistiche descrittive

Utilizzando la Proc Freq di Sas abbiamo calcolato le tabelle con le frequenze per alcune coppie di variabili che abbiamo ritenuto importanti per le analisi; così facendo abbiamo anche verificato due degli assunti più importanti del modello logistico: la quasi-separation e la zero variance.

Frequency Percent Row Pct Col Pct	Table of HT by SMOKE		
	SMOKE(SMOKE)		
HT(HT)	0	1	Total
0	106 57.92 63.10 93.81	62 33.88 36.90 88.57	168 91.80
1	7 3.83 46.67 6.19	8 4.37 53.33 11.43	15 8.20
Total	113 61.75	70 38.25	183 100.00

Frequency Percent Row Pct Col Pct	Table of LOW by PTL		
	PTL(PTL)		
LOW(LOW)	0	1	Total
0	117 63.93 90.00 76.97	13 7.10 10.00 41.94	130 71.04
1	35 19.13 66.04 23.03	18 9.84 33.96 58.06	53 28.96
Total	152 83.06	31 16.94	183 100.00

Frequency Percent Row Pct Col Pct	Table of LOW by HT		
	HT(HT)		
LOW(LOW)	0	1	Total
0	122 66.67 93.85 72.62	8 4.37 6.15 53.33	130 71.04
1	46 25.14 86.79 27.38	7 3.83 13.21 46.67	53 28.96
Total	168 91.80	15 8.20	183 100.00

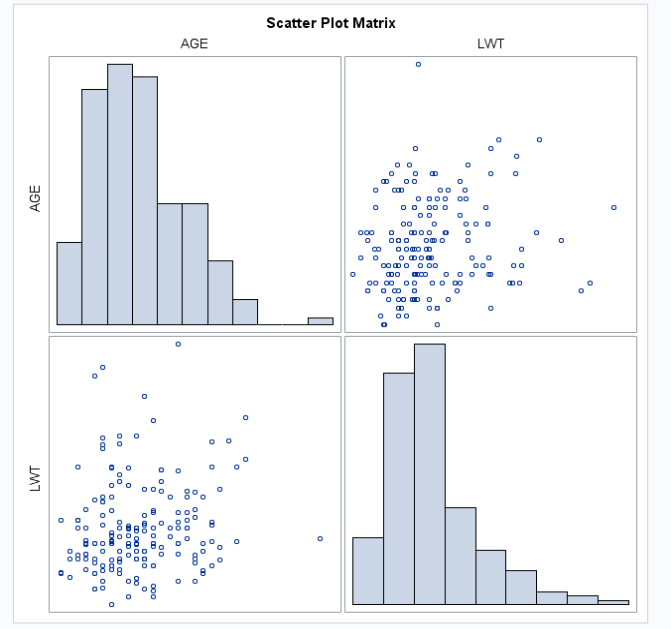
Frequency Percent Row Pct Col Pct	Table of LOW by RACE		
	RACE(RACE)		
LOW(LOW)	0	1	Total
0	75 40.98 57.69 78.95	55 30.05 42.31 62.50	130 71.04
1	20 10.93 37.74 21.05	33 18.03 62.26 37.50	53 28.96
Total	95 51.91	88 48.09	183 100.00

Abbiamo considerato la relazione tra la variabile risposta basso peso alla nascita e le variabili parto prematuro, ipertensione ed etnia, inoltre abbiamo analizzato la relazione tra il fumo e l’ipertensione, poiché è un legame molto presente nella letteratura scientifica.
Osserviamo che per la relazione fumo-ipertensione e basso peso alla nascita-ipertensione vi sono pochi casi in cui entrambe le variabili assumo livello pari a 1 (ossia “presenza dell’evento”), la percentuale di casi aumenta per le relazioni tra basso peso alla nascita e parto prematuro (10%) ed etnia (18%). Da questi risultati possiamo assumere l’assenza di quasi-separation e zero variance.
Le altre variabili presenti nell’analisi non mostravano risultati significativi circa la loro relazione.

3 - Modello logistico

3.1 - Assunti

Prima di calcolare le stime del modello logistico, abbiamo controllato che lo stesso rispettava gli assunti.
Abbiamo verificato le distribuzioni delle variabili continue presenti nel modello. Per quanto concerne ciò, si osserva che sia LWT che AGE hanno un’asimmetria positiva con coda a destra.
Tra le due variabili non si osserva nessun legame lineare forte, l’indice di correlazione lineare è pari a 0.15, ciò viene confermato anche dai valori del VIF (Variance Inflaction Index) e del TOL mostrati di seguito.



Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Tolerance	Variance Inflation
Intercept	Intercept	1	0.83181	0.19338	4.30	<.0001	.	0
AGE	AGE	1	-0.00750	0.00625	-1.20	0.2314	0.97597	1.02462
LWT	LWT	1	-0.00283	0.00114	-2.49	0.0138	0.97597	1.02462

Inoltre, abbiamo controllato la presenza di possibili valori influenti: nessuna unità statistica presentava outliers o valori anomali. Abbiamo deciso, quindi, di mantenerle tutte, eccetto per 7 osservazioni poiché presentavano valori mancanti in alcune covariate.
Per il modello logistico, la normalità non è un assunto necessario da verificare; abbiamo provato ad utilizzare la trasformazione logistica per le uniche due variabili continue (età e peso della madre all’ultimo periodo mestruale) con lo scopo di migliorare le performance del modello, ma ciò non si è verificato. Abbiamo, quindi, deciso di continuare le analisi senza utilizzare le trasformazioni logistiche di queste covariate, in modo da avere un modello più

semplice e più facilmente interpretabile.

3.2 - Risultati del modello logistico

Prima di utilizzare il modello logistico in ambito predittivo, abbiamo calcolato le stime dei singoli coefficienti per ogni variabile, i risultati ottenuti sono i seguenti.

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	37.6941	7	<.0001
Score	35.5156	7	<.0001
Wald	27.7260	7	0.0002

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	0.6044	1.2862	0.2208	0.6384
AGE		1	-0.0467	0.0378	1.5260	0.2167
LWT		1	-0.0142	0.00728	3.7833	0.0518
RACE	1	1	0.9567	0.4142	5.3350	0.0209
SMOKE	1	1	0.7119	0.4104	3.0087	0.0828
PTL	1	1	1.5148	0.4675	10.4989	0.0012
HT	1	1	1.2726	0.6275	4.1128	0.0426
UI	1	1	0.8193	0.4816	2.8934	0.0889

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
AGE	0.954	0.886	1.028
LWT	0.986	0.972	1.000
RACE 1 vs 0	2.603	1.156	5.862
SMOKE 1 vs 0	2.038	0.912	4.555
PTL 1 vs 0	4.548	1.819	11.371
HT 1 vs 0	3.570	1.044	12.213
UI 1 vs 0	2.269	0.883	5.832

Nella prima tabella sono presenti 3 test che verificano la non nullità dell’intero modello,

svolgono il medesimo compito del “classico” Test F nel modello lineare. Tutti risultano significativi, quindi ci portano a rifiutare l’ipotesi H_0 di nullità di tutti i parametri del modello.
Le stime presenti nell’ultima tabella sono gli OR, determinati fissando come categoria di riferimento il livello 0.
Le variabili che risultano significative sono: l’etnia (OR = 2.603), il parto prematuro (OR = 4.548) e l’ipertensione (OR = 3.570). Tuttavia, gli intervalli di confidenza sono molto ampi, quindi risultano essere poco informativi.

4 - Modello predittivo

Adesso valuteremo le capacità predittive dei seguenti modelli :

- *LOW = RACE + HT + PTL* (modello con le sole variabili risultate significative dal modello logistico visto sopra);
- *LOW = LWT + RACE + HT + PTL* (modello con le variabili precedenti e la variabile relativa al peso della madre all’ultimo periodo mestruale poiché ritenuto un fattore importante);
- *LOW = LWT + AGE + SMOKE + RACE + HT + PTL + UI* (modello completo contenente tutte le variabili in analisi).

4.1 - Errore di predizione

Per valutare la capacità predittiva del modello abbiamo utilizzato le misure di R^2 di Nagelkerke e il Brier Score.
La prima misura indica la capacità predittiva del modello di spiegare la varianza di LOW attraverso le covariate inserite nel modello, mentre il Brier Score confronta ogni valore i-esimo dell’evento osservato *y* con la rispettiva probabilità predetta i-esima.

Di seguito mostriamo le misure di predizione del modello con quattro covariate (prima tabella) e il modello completo (seconda tabella).

Fit Statistics for SCORE Data

Data Set	Total Frequency	Log Likelihood	Error Rate	AIC	AICC	BIC	SC	R-Square	Max-Rescaled R-Square	AUC	Brier Score
WORK.VALID	60	-27.8162	0.2000	65.63233	66.74344	76.10406	76.10406	0.23275	0.334204	0.850205	0.149042

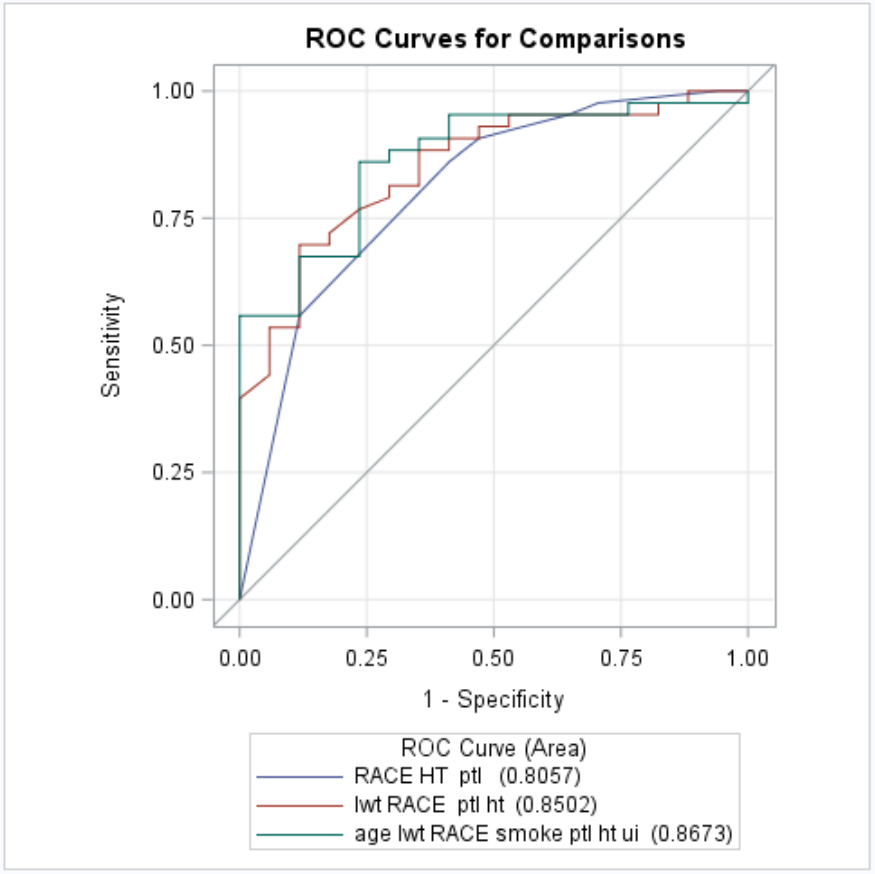
Fit Statistics for SCORE Data

Data Set	Total Frequency	Log Likelihood	Error Rate	AIC	AICC	BIC	SC	R-Square	Max-Rescaled R-Square	AUC	Brier Score
WORK.VALID	60	-26.3910	0.1833	68.78205	71.60558	85.53681	85.53681	0.268346	0.385315	0.867305	0.139827

Notiamo che i valori del Brier Score sono leggermente più bassi nel modello con tutte le covariate rispetto all’altro modello; si ricorda che questa misura assume valori nell’intervallo [0, 0.25], il modello per avere una buona predizione dovrebbe avere dei valori prossimi a 0.
Inoltre l’indice R^2 di Nagelkerke risulta maggiore nel modello con tutte le covariate rispetto all’altro.
Con le analisi successive andremo a verificare se effettivamente il modello completo è il migliore anche sotto gli altri punti di vista (discriminazione e calibrazione), rispetto ai due modelli con meno covariate.

4.2 - Discriminazione

Per analizzare l’aspetto della discriminazione abbiamo sviluppato le curve ROC con i relativi contrasti per valutare la migliore combinazione di caratterisitiche operative (sensibilità e 1-specificità).



ROC Association Statistics

ROC Model	Mann-Whitney Area	Standard Error	95% Wald Confidence Limits		Somers' D	Gamma	Tau-a
RACE HT ptl	0.8057	0.0630	0.6823	0.9292	0.6115	0.7388	0.2525
lwt RACE ptl ht	0.8502	0.0522	0.7479	0.9525	0.7004	0.7052	0.2893
age lwt RACE smoke ptl ht ui	0.8673	0.0487	0.7719	0.9627	0.7346	0.7346	0.3034

ROC Contrast Test Results			
Contrast	DF	Chi-Square	Pr > ChiSq
contrasti	2	3.6279	0.1630

ROC Contrast Estimation and Testing Results by Row						
Contrast	Estimate	Standard Error	95% Wald Confidence Limits		Chi-Square	Pr > ChiSq
RACE HT ptl - lwt RACE ptl ht	-0.0445	0.0266	-0.0967	0.00777	2.7839	0.0952
RACE HT ptl - age lwt RACE smoke ptl ht ui	-0.0616	0.0328	-0.1258	0.00265	3.5313	0.0602
lwt RACE ptl ht - age lwt RACE smoke ptl ht ui	-0.0171	0.0203	-0.0569	0.0227	0.7101	0.3994

Per quanto riguarda i risultati delle curve ROC osserviamo che il modello completo assume un valore di AUC = 0.86, il modello con quattro covariate ha un AUC = 0.85, infine il modello con solo tre variabili ottiene un AUC = 0.80. Osserviamo, quindi, un discreto aumento tra il modello completo e il modello con tre variabili, il modello con quattro covariate assume valori molto vicini al modello completo.
Il test di Delong, però, non risulta significativo per nessun confronto, possiamo affermare quindi che il miglioramento tra i singoli modelli è decisamente moderato.
Andiamo a valutare l’ultimo aspetto dei modelli predittivi, ossia la calibrazione per avere un quadro completo che ci permetta di scegliere il modello finale.

4.3 - Calibrazione e scelta del modello migliore

Gli strumenti migliori per valutare quest’ultimo aspetto sono: l’indice Net Reclassification Improvement (NRI), il quale misura il miglioramento nella classificazione all’aumentare di una covariata e il Test di Hosmer-Lemeshow che misura la bontà di calibrazione generale del modello. Inoltre, abbiamo calcolato anche l’Integrated Discrimination Improvement (IDI), il quale confronta la capacità di ogni covariata aggiunta di classificare correttamente l’evento.
Sia l’NRI che l’IDI sono indici utili per confrontare due modelli, noi li abbiamo utilizzati per confrontare il modello completo e il modello con quattro covariate rispetto al modello iniziale.

AUC modello 3 cov.	AUC modello 4 cov.	Differenza AUC	P-value
0.80575	0.85021	0.0445	0.0952

Metriche	Stima puntuale	Intervallo di conf.	P-value
IDI	0.041953	[0.01112, 0.0727]	0.0075
NRI (category free)	0.67031	[0.2006, 1.1401]	0.0193
NRI (user category)	0.39808	[0.1119, 0.6843]	0.0121

Hosmer Lemeshow Test with 8 df

model	Hosmer Lemeshow Chi Square	Degree of Freedom	P-Value
Model1	3.76078	8	0.878
Model2	6.25570	8	0.6186

Il modello con quattro covariate ottiene i risultati migliori nella calibrazione. Esso ottiene un valore di NRI (per quanto riguarda “Category Free”) pari a 0.67 e una percentuale di classificazione corretta degli eventi pari al 65%.
L’Integrated Discriminant Improvement risulta significativo per il modello con quattro covariate rispetto al modello iniziale, esso ci dice che con l’aggiunta di una covariata, il modello ha migliorato le sue performance nella classificazione degli eventi.
Per quanto riguarda il Test di Hosmer Lemeshow siamo portati ad accettare l’ipotesi nulla, la quale indica che entrambi i modelli hanno una buona capacità di adattamento.

5 - Discussione

Elenchiamo di seguito i punti di forza dello studio appena presentato.

- Il modello logistico risulta robusto, ciò è evidenziato dal fatto che rispetta gli assunti verificati all’inizio dello studio, ossia collinearità, quasi-separation, zero variance ed esclusione di casi influenti. Abbiamo provato ad effettuare la trasformazione logistica nelle variabili continue, i risultati ottenuti non si sono discostati dagli altri.
- I risultati ottenuti dal modello predittivo per quanto riguarda la bontà di discriminazione dello stesso sono ottimi; infatti, abbiamo ottenuto (con il modello ritenuto migliore, ossia quello con quattro variabili) un valore di AUC (Area Under Curve) pari a 0.85, anche il modello con le sole tre variabili significative (per il modello logistico) ha ottenuto un valore di AUC alto, pari a 0.80.
- Il modello con le quattro variabili ha ottenuto una buona performance anche nelle misure di calibrazione: la stima puntuale di NRI risulta pari a 0.67 con il relativo p-value significativo, così come l’intervallo di confidenza [0.20, 1.14]. Quest’ultimo non contiene lo zero e il suo estremo superiore assume un valore maggiore di 1, sintomo che all’aumentare di una covariata, la calibrazione del modello migliora notevolmente.

Elenchiamo di seguito i limiti dello studio appena presentato.

- La numerosità del campione analizzato è esigua per ottenere un modello predittivo robusto con dei risultati generalizzabili.
- La quantità delle covariate è limitata, essa non ci permette di avere un quadro più completo sulle caratteristiche socio-sanitarie delle donne incinte.
- I risultati mostrati potrebbero essere distorti da alcune variabili confondenti, durante l’analisi non abbiamo utilizzato nessuna tecnica per verificare la robustezza delle stime (soprattutto per la mancanza di informazioni come già sottolineato sopra).
  Le uniche due variabili potenzialmente confondenti tenute in considerazione sono

quelle relative al fumo e all’ipertensione.

In conclusione, giungiamo alle seguenti considerazioni.
Le variabili che risultano significative nel modello logistico, quindi quelle con un legame più forte con il basso peso alla nascita, sono: l’ipertensione, l’etnia e storia di un parto prematuro. Il primo modello predittivo che abbiamo stimato è stato quello con queste tre variabili, poi ad esse abbiamo aggiunto la variabile relativa al peso della madre all’ultimo periodo mestruale (ritenuto un fattore molto importante in letteratura); a questi due modelli abbiamo confrontato il modello completo con tutte le covariate.
Abbiamo ottenuto dei buoni risultati per tutti e tre i modelli, all’aumentare del numero di covariate le misure di predizione e discriminazione migliorano (come è facile immaginarsi); mentre per quanto riguarda la calibrazione, possiamo affermare che il modello migliore è quello con quattro covariate, in quanto classifica meglio gli eventi e i non-eventi.
Esso è anche la sintesi migliore tra il modello con solo tre variabili (il quale raggiunge dei risultati minori) e il modello completo, il quale ottiene i risultati migliori per quanto riguarda l’R^2 di Nagelkerke, il Brier Score e l’AUC ma risulta eccessivamente complesso (aumentando, così, il rischio overfitting) e non ottiene dei risultati di calibrazione buoni come il modello con quattro variabili.

6 - Bibliografia

[1] Dispense corso “Metodi e Modelli Biostatistici per la Ricerca Clinica”, Università degli Studi di Milano-Bicocca, prof.ssa A. Zambon;

[2] Dispense corso “Statistica Computazionale”, Università degli Studi di Milano-Bicocca, prof. P. Lovaglio;

[3] Pencina M. J., Demler O.V., et al., Novel metrics for evaluating improvement in discrimination: net reclassification and integrated discrimination improvement for normal variables and nested models, Stat Med. 2012 Jan 30; 31(2): 101–113;

[4] Ministero della Salute, La prevenzione in gravidanza, 2022, https://www.salute.gov.it/portale/alleanzaCardioCerebrovascolari/dettaglioContenutiAlleanzaCardioCerebrovascolari.jsp?lingua=italiano&id=5776&area=Alleanza%20italiana%20per%20le%20malattie%20cardio-cerebrovascolari&menu=prevenzioneMC.