Logistic Regression in SPSS

Πρόχειρες βοηθητικές σημειώσεις

Παράδειγμα

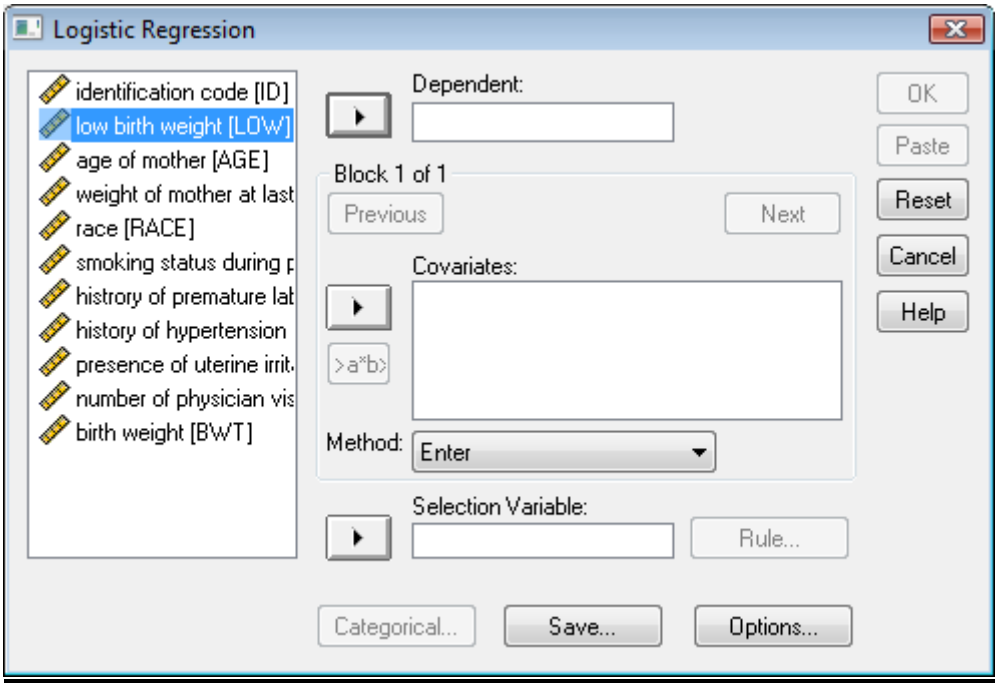
Στο αρχείο lowbirthweight.sav (πηγή Hosmer and Lemeshow (2000), http://www.umass.edu/statdata/statdata/stat-logistic.html) καταγράφονται πληροφορίες για 189 γεννήσεις καθώς και για τις μητέρες των νεογνών. Το ενδιαφέρον επικεντρώνεται στη μελέτη του φαινομένου της γέννησης νεογνών με βάρος μικρότερου των 2.500 γραμμαρίων. Το ενδιαφέρον εξηγείται διότι η θνησιμότητα των νεογνών σε τέτοιες περιπτώσεις είναι πολύ υψηλή. Οι πληροφορίες που καταγράφονται είναι οι ακόλουθες:

identification code	Κωδικός αριθμός συμμετέχοντα
Low birth weight	Αν το νεογνό ζυγίζει λιγότερο από 2500γρ
Age of mother	Ηλικία της μητέρας σε έτη
Weight of mother at last menstrual period	Βάρος της μητέρας την τελευταία εμμηνορροϊκή περίοδο
race	Φυλή, γένος.
smoking status during pregnancy	Κάπνισμα κατά την κυοφορία
History of premature labor	Ιστορικό πρόωρων γέννων
History of hypertension	Ιστορικό υπέρτασης
Presence of uterine irritability	Παρουσία ερεθιστικότητας
Number of physician visits during the first trimester	Αριθμός επισκέψεων ιατρού κατά το πρώτο τρίμηνο
birth weight	Βάρος νεογνού

<u>Προσαρμογή του μοντέλου της λογιστικής παλινδρόμησης όταν η εξαρτημένη ποιοτική μεταβλητή είναι δίτιμη.</u>

Το μοντέλο της λογιστικής παλινδρόμησης όπου η εξαρτημένη μεταβλητή είναι δίτιμη μπορεί να προσαρμοστεί τόσο από την διαδικασία Binary Logistic όσο και από την διαδικασία Multinomial Logistic Regression. Κάθε μία έχει μερικές επιλογές που δεν είναι διαθέσιμες στην άλλη. Ας προσαρμόσουμε στο συγκεκριμένο παράδειγμα το μοντέλο της λογιστικής παλινδρόμησης μέσω της διαδικασίας Binary Logistic.

1. Επιλέγουμε Analyze Regression Binary Logistic



Στο παράθυρο διαλόγου που προκύπτει τοποθετούμε στο πλαίσιο **Dependent** την κατηγορική (δίτιμη αφού επιλέξαμε Binary Logistic) μεταβλητή, ενώ στο πλαίσιο **Covariates** τις ανεξάρτητες μεταβλητές, οι οποίες είναι είτε κατηγορικές με k το πλήθος δυνατές τιμές είτε συνεχείς.

Από το πλαίσιο Method επιλέγουμε τη μέθοδο με την οποία υπεισέρχονται οι ανεξάρτητες μεταβλητές στην ανάλυση. Οι διαθέσιμες μέθοδοι είναι οι ακόλουθες:

- <u>Enter.</u> Η διαδικασία κατά την οποία όλες οι μεταβλητές υπεισέρχονται σε ένα βήμα, σε μία φάση.

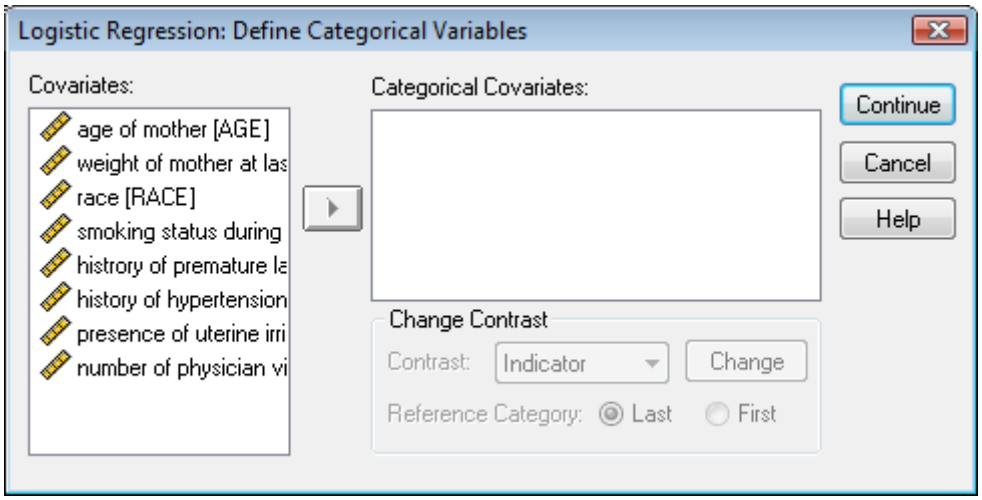
- Forward Selection (Conditional. Likelihood Ratio και Wald). Τρεις διαφορετικές μέθοδοι επιλογής των μεταβλητών που θα υπεισέρθουν στο μοντέλο με 3 διαφορετικά κριτήρια (δε θα υπεισέρθουμε σε λεπτομέρειες).
- Backward Elimination (Conditional. Likelihood Ratio and Wald). Τρεις διαφορετικές μέθοδοι επιλογής ποιων μεταβλητών θα βγουν από το μοντέλο με 3 διαφορετικά κριτήρια (δε θα υπεισέρθουμε σε λεπτομέρειες)

Σημειώνουμε ότι μας δίνεται η δυνατότητα καθορισμού διαφορετικών μεθόδων επιλογής ανεξάρτητων μεταβλητών σε υποσύνολα αυτών από το πλαίσιο Block.

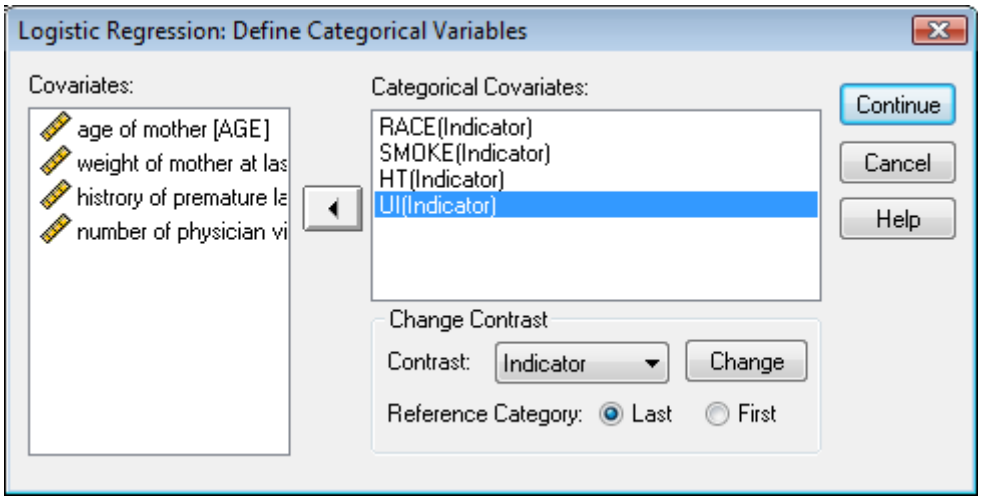
Τοποθετώντας στο πλαίσιο Selection Variable μία μεταβλητή και δίνοντας έναν κανόνα π.χ. age>35 σημαίνει ότι ζητούμε την προσαρμογή του μοντέλου μόνο για εκείνες τις πειραματικές μονάδες που ικανοποιούν αυτόν τον κανόνα. Όμως τα στατιστικά και τα αποτελέσματα της ταξινόμησης δίνονται και για τις υπόλοιπες πειραματικές μονάδες. Επομένως έτσι έχουμε τη δυνατότητα διαχωρισμού του δείγματος σε training και validation.

Τέλος για να υπεισέρθει στο μοντέλο η αλληλεπίδραση κάποιων μεταβλητών αρκεί να τις επιλέξουμε όλες ταυτόχρονα (πατώντας το Shift) και έπειτα να πατήσουμε το πλαίσιο a*b.

2. Πατώντας το πλαίσιο Categorical



δηλώνουμε τις όποιες ανεξάρτητες μεταβλητές είναι κατηγορικές. Τότε δημιουργούνται k-1 δείκτριες μεταβλητές.



Change Contrast. Επιτρέπει την αλλαγή της μεθόδου αντιθέσεων (contrast). Είναι δυνατές οι ακόλουθες.

- Indicator. Δημιουργούνται με τη λογική της παρουσίας ή όχι ενός χαρακτηριστικού του μέλους. Στον πίνακα που μας δίνεται ως Contrast matrix η κατηγορία αναφοράς είναι εκείνη που όλα τα στοιχεία της γραμμής είναι ίσα με το μηδέν.

- Simple. Κάθε κατηγορία της ανεξάρτητης κατηγορικής μεταβλητής συγκρίνεται με την κατηγορία αναφορά.
- Difference. Κάθε κατηγορία της ανεξάρτητης μεταβλητής εκτός από την πρώτη συγκρίνεται με τη μέση επίδραση των προηγούμενων.
- Helmert. Κάθε κατηγορία της ανεξάρτητης μεταβλητής εκτός από την τελευταία συγκρίνεται με τη μέση επίδραση των προηγούμενων.
- Repeated. Κάθε κατηγορία της ανεξάρτητης εκτός από την πρώτη συγκρίνεται με την προηγούμενή της.
- Polynomial. Ορθογώνιες πολυωνυμικές αντιθέσεις. Είναι διαθέσιμες μόνο για αριθμητικές-ποσοτικές μεταβλητές.
- Deviation. Κάθε κατηγορία της ανεξάρτητης μεταβλητής εκτός από την κατηγορία αναφοράς συγκρίνεται με την συνολική επίδραση.

Η προεπιλεγμένη μέθοδος είναι η Indicator. Σε περίπτωση επιλογής μίας εκ των Deviation, Simple, ή Indicator, επιλέγουμε και αν θα είναι η κατηγορία

αναφοράς το πρώτο ή το τελευταίο επίπεδο της κατηγορικής μεταβλητής. Σημειώνεται ότι η διαφοροποίηση υλοποιείται και επιτυγχάνεται μόνο όταν πατήσουμε το πλαίσιο Change.

Χωρίς να προχωρήσουμε σε άλλες επιλογές ας δούμε τα αποτελέσματα που θα προέκυπταν:

Case Processing Summary

Unweighted Cases(a)		N	Percent
Selected Cases	Included in Analysis	189	100,0
	Missing Cases	0	,0
	Total	189	100,0
Unselected Cases		0	,0
Total		189	100,0

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

Ο πίνακας **Case Processing Summary** μας ενημερώνει για το πλήθος των πειραματικών μονάδων που λαμβάνουν μέρος στην ανάλυση, ενώ ο πίνακας **Dependent Variable Encoding** για την κωδικοποίηση της εξαρτημένης μεταβλητής. Εδώ παράμεινε ίδιος αλλά αν αρχικά η μεταβλητή είχε κωδικοποιηθεί με τιμές 3 και 4 θα γινόταν αλλαγή σε 0,1. Τέλος στον πίνακα Categorical Variables Encoding δίνονται οι συχνότητες για τις κατηγορικές μεταβλητές του μοντέλου καθώς και η κωδικοποίηση των δείκτριων μεταβλητών που θα χρησιμοποιηθούν. Έτσι προκύπτει ότι race (1) είναι για τους λευκούς κοκ.

Dependent Variable Encoding

Original Value	Internal Value
>=2500g	0
<2500g	1

Categorical Variables Codings

		Frequency	Parameter coding (1)	(2)
race	White	96	1,000	,000
	Black	26	,000	1,000
	Other	67	,000	,000
history of hypertension	no	177	1,000	
	yes	12	,000	
smoking status during pregnancy	no	115	1,000	
	yes	74	,000	

Έπειτα ακολουθούν τα αποτελέσματα του πρώτου βήματος προσαρμογής του μοντέλου υπό τον τίτλο Block 0: Beginning Block. Καθώς επιλέξαμε τη μέθοδο Enter το αρχικό μοντέλο που προσαρμόζεται έχει μόνο το σταθερό όρο.

Στον πίνακα Classification Table (a,b) μας δίνονται οι ακόλουθες πληροφορίες. Τα 130 νεογνά που δεν έχουν βάρος μικρότερο των 2500 γραμμαρίων ς ταξινομούνται (στήλη Predicted) ως νεογνά με βάρος μεγαλύτερο των 2500 γραμμαρίων (ποσοστό ορθής ταξινόμησης 100%), ενώ τα 59 νεογνά με βάρος μικρότερου των 2500 γραμμαρίων ταξινομούνται και αυτά ως νεογνά με βάρος μεγαλύτερο των 2500 γραμμαρίων (ποσοστό ορθής ταξινόμησης 0%). Δηλ. στο πρώτο βήμα όλα τα νεογνά ταξινομούνται ως νεογνά με βάρος μεγαλύτερο των 2500 γραμμαρίων με συνολικό ποσοστό ορθής ταξινόμησης 68,8%.

Classification Table(a,b)

Observed			Predicted		
			low birth weight		Percentage Correct
			>=2500g	<2500g	
Step 0	low birth weight	>=2500g	130	0	100,0
		<2500g	59	0	,0
	Overall Percentage				68,8

a Constant is included in the model.
b The cut value is ,500

Από τον πίνακα Variables in the Equation έχουμε ότι ο σταθερός όρος (Constant) έχει μπει στο μοντέλο. Στη στήλη B δίνεται ο συντελεστής για το σταθερό όρο, το τυπικό του σφάλμα (S.E.), η τιμή του Wald στατιστικού τεστ για τον έλεγχο της υπόθεσης ότι ο σταθερός όρος είναι ίσος με το μηδέν. Οι βαθμοί ελευθερίας

αυτού του τεστ είναι ίσοι με το πλήθος των ανεξάρτητων μεταβλητών που είναι στο μοντέλο, ενώ στη στήλη Sig. μας δίνεται η p-τιμή του (άρα εδώ δεν απορρίπτεται η υπόθεση ότι ο σταθερός όρος είναι ίσος με το μηδέν, αλλά αυτή η υπόθεση δεν ενδιαφέρει τους ερευνητές). Τέλος από τη στήλη Exp(B) μας δίνεται η τιμή του exp(-0.790)=0,454 που δεν είναι τίποτε άλλο παρά το odds ratio που ισούται με το πηλίκο 59/130.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 0	Constant	-,790	,157	25,327	1	,000	,454

Από τον πίνακα Variables not in the Equation μας δίνονται ποιες από τις μεταβλητές που έχουν δηλωθεί στο πλαίσιο Covariates δεν έχουν υπεισέρθει στο μοντέλο.. Στη στήλη Score μας δίνεται η τιμή του Score στατιστικού που χρησιμοποιείται για να προβλέψει αν μία ανεξάρτητη μεταβλητή θα υπεισέρθει ή όχι στατιστικά σημαντικά στο μοντέλο. Αν η αντίστοιχη p-τιμή που δίνεται στην στήλη Sig. είναι μικρότερη του 5%=0,05 τότε η εν λόγω μεταβλητή πρέπει να μπει στο μοντέλο. Τέλος στη γραμμή Overall Statistics δίνονται τα αποτελέσματα για τον έλεγχο με το Score στατιστικό τεστ αν όλες οι ανεξάρτητες μεταβλητές θα πρέπει να μπουν στο μοντέλο.

Variables not in the Equation

			Score	df	Sig.
Step 0	Variables	AGE	2,674	1	,102
		LWT	5,438	1	,020
		RACE	5,005	2	,082
		RACE(1)	4,787	1	,029
		RACE(2)	1,727	1	,189
		SMOKE(1)	4,924	1	,026
		PTL	7,267	1	,007
		HT(1)	4,388	1	,036
		UI	5,401	1	,020
		FTV	,749	1	,387
	Overall Statistics		30,959	9	,000

Στη συνέχεια ακολουθούν τα αποτελέσματα του επόμενου βήματος (με τη μέθοδο Enter) που ονομάζεται Block 1: Method = Enter, όπου υπεισέρχονται στο μοντέλο όσες ανεξάρτητες μεταβλητές έχουν δηλωθεί στο πλαίσιο Covariates.

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	33,387	9	,000
	Block	33,387	9	,000
	Model	33,387	9	,000

Στον πίνακα Omnibus Tests of Model Coefficients μας δίνεται η τιμή καθώς και η αντίστοιχη p- τιμή του X^2 στατιστικού για τον έλεγχο ότι το συνολικό μοντέλο είναι στατιστικά σημαντικό. Δηλαδή είναι το αντίστοιχο του F-τεστ της γραμμικής παλινδρόμησης. Παρατηρούμε ότι η p-τιμή είναι μικρότερη του 0.05 επομένως το μοντέλο είναι στατιστικά σημαντικό.

Για να δούμε πως προκύπτει η τιμή αυτή. Ορίζεται ως Deviance του μοντέλου m η ποσότητα

$$D_m = -2\ln f(Y/\beta) = -2\sum_{i=1}^{n} y_i \ln(p_i) + (1-y_i)\ln(1-p_i).$$

Έτσι με D_c και D_s συμβολίζουμε την deviance του μοντέλου του σταθερού όρου και το κορεσμένο μοντέλο (το μοντέλο που οι μέσοι=data).

Η τιμή στον πίνακα Omnibus Test of model Coefficient προκύπτει από τη σχέση (βλέπε Hosmer Lemeshow (2000))

$$-2\ln\left[\frac{\text{likelihood without the variables}}{\text{likelihood with the variables}}\right] = D_o - D_1.$$

Δηλαδή μας δίνεται η διαφορά στη Deviance από το προηγούμενο βήμα. Εδώ το προηγούμενο βήμα είναι το σταθερό μοντέλο. Επομένως ελέγχουμε κατά πόσο οι ανεξάρτητες μεταβλητές βελτιώνουν την εκτίμηση.

Παρατήρηση Η τιμή του στατιστικού είναι ίδια για το Step, Block και Model καθώς δεν έχει χρησιμοποιηθεί ούτε Stepwise logistic regression ούτε Blocking. Οι βαθμοί ελευθερίας είναι τόσοι όσο το πλήθος των ανεξάρτητων μεταβλητών, λαμβάνοντας υπόψη και τις πιθανές δείκτριες που δημιουργούνται.

Στον πίνακα Model Summary μας δίνεται ότι -2Log(πιθανοφάνειας)=201,285. Επομένως $D_m = 201,285$ και άρα $D_0 = 201,285 + 33,387$. Η τιμή αυτή από μόνη της δεν είναι χρήσιμη, αλλά χρησιμεύει για τη σύγκριση των διάφορων πιθανών μοντέλων. Επιπλέον δίνονται οι τιμές των Cox and Snell R^2 (Cox and Snell, 1989, The Analysis of Binary Data, 2nd ed. London: Chapman and Hall.) καθώς και του Nagelkerke R^2 (Nagelkerke, 1991, A note on the general definition of the coefficient of determination. Biometrika, 78:3, 691-692.). Καθώς στη λογιστική παλινδρόμηση δεν υπάρχει ο συντελεστής προσδιορισμού R^2 πολλοί ερευνητές προσπάθησαν να εισάγουν κάποιο παρόμοιο συντελεστή ή ψευδοσυντελεστή. Καλό θα είναι να μην ερμηνεύονται κατά ίδιο τρόπο με το συντελεστή προσδιορισμού και να είμαστε επιφυλακτικοί στη χρήση τους.

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	201,285(a)	,162	,228

a Estimation terminated at iteration number 5 because parameter estimates changed by less than ,001.

Τρόπος υπολογισμού των Cox and Snell και Nagelkerke R Square:

$$R^2_{CS} = 1 - \exp\left(\left(D_m - D_0\right)/n\right) = 1 - \exp\left(-33.387/189\right)$$

και

$$R^2_N = \frac{R^2_{cs}}{1 - \exp\left(-D_0/n\right)} = \frac{0.162}{1 - \exp(-234,672/189)}$$

Έχουμε τώρα το νέο πίνακα Classification table όπου τώρα το συνολικό ποσοστό ορθής ταξινόμησης είναι 74,1%. Παρατηρούμε ότι το ποσοστό ορθής ταξινόμησης αυξήθηκε από 68,8% σε 74,1%.

Classification Table(a)

Observed			Predicted		
			low birth weight		Percentage Correct
			>=2500g	<2500g	
Step 1	low birth weight	>=2500g	117	13	90,0
		<2500g	36	23	39,0
	Overall Percentage				74,1

a The cut value is ,500

<u>Παρατήρηση (Αποτίμηση της προβλεπτικής ικανότητας της λογιστικής παλινδρόμησης)</u>

Επιπλέον παρατηρούμε ότι το ποσοστό ορθής ταξινόμησης εντός των νεογνών με βάρος μεγαλύτερο των 2500 γραμμαρίων είναι ίσο με 90% (specificity) ενώ το ποσοστό ορθής ταξινόμησης εντός των νεογνών με βάρος μικρότερο των 2500 γραμμαρίων είναι 39% (sensitivity).

Ισχύουν οι ακόλουθοι ορισμοί:

Accuracy: $\dfrac{\text{Correctly Classified}}{\text{Total observed}} = \dfrac{117+23}{117+23+36+13} = 0.741$

Sensitivity ή true positive fraction: $\dfrac{\text{Correctly Classified as } Y=1}{\text{Total observed as } Y=1} = \dfrac{23}{59}$

Specificity ή false positive fraction: $\dfrac{\text{Correctly Classified as } Y=0}{\text{Total observed as } Y=0} = \dfrac{117}{130}$

Positive predictive value: $\dfrac{\text{Correctly Classified as } Y=1}{\text{Total classified as } Y=1} = \dfrac{23}{36}$

Negative predictive value: $\dfrac{\text{Correctly Classified as } Y=0}{\text{Total classified as } Y=0} = \dfrac{117}{153}$.

Οι δείκτες sensitivity και specificity χρησιμοποιούνται για την αποτίμηση της ακρίβειας της πρόβλεψης. Μας δίνουν πόσα καλά γίνεται η ταξινόμηση μεταξύ πειραματικών μονάδων που ικανοποιούν ή όχι μία συγκεκριμένη συνθήκη. Η επιλογή του cut-off είναι εκείνη που καθορίζει το πλήθος των σωστών και λανθασμένων ταξινομήσεων. Εύκολα γίνεται αντιληπτό ότι καθώς αυξάνεται η sensitivity ταυτόχρονα μειώνεται η specificity. Μία πιο πλήρη περιγραφή της ακρίβειας της ταξινόμησης επιτυγχάνεται με την λεγόμενη Receiver Operating Characteristic (ROC) Curve. Είναι το γράφημα της (1-specificity) στον άξονα των X και sensitivity στον άξονα των Y για τις διάφορες τιμές του cut off point. Το εμβαδό του χωρίου κάτω από την καμπύλη (αναφέρεται και ως index of accuracy A ή *concordance index*) αποτελεί έναν δείκτη της ακρίβειας. Όσο μεγαλύτερο τόσο καλύτερη η ισχύς της πρόβλεψης. Η ROC αν είναι στη διαγώνιο σημαίνει τυχαίος ταξινομικός κανόνας (ισοδύναμος με το να ρίχνουμε ένα νόμισμα).Ο τρόπος κατασκευής του γραφήματος αυτού θα γίνει στη συνέχεια.

Στον πίνακα Variables in the Equation έχουμε το προσαρμοζόμενο μοντέλο. Πληροφορούμαστε λοιπόν για την σχέση των ανεξάρτητων μεταβλητών με την εξαρτημένη μεταβλητή, η οποία είναι στην κλίμακα Logit. Από τις τιμές των Wald στατιστικών και τις αντίστοιχες p- τιμές κρίνουμε ποιες μεταβλητές είναι στατιστικά σημαντικές. Διαπιστώνουμε ότι οι μεταβλητές age, ptl, ui ftv δεν συνεισφέρουν στατιστικά σημαντικά στο μοντέλο.

Αν σκοπός μας είναι να αποκτήσουμε ένα μοντέλο που προσαρμόζεται όσο γίνεται καλύτερα και επιπλέον ελαχιστοποιείται ο αριθμός των παραμέτρων το επόμενο λογικό βήμα είναι η προσαρμογή ενός μοντέλου που περιέχει τις μεταβλητές που είναι στατιστικά σημαντικές και η σύγκρισή του με το πλήρες μοντέλο.

Οι εκτιμητές (στήλη B) μας δίνουν την αύξηση (ή αντίστοιχα μείωση αν το πρόσημο είναι αρνητικό), στα προβλεπόμενα log odds της low birth weight=1 όταν θα έχουμε μοναδιαία αύξηση στην αντίστοιχη ανεξάρτητη μεταβλητή διατηρώντας τις υπόλοιπες σταθερές. Έτσι αύξηση του βάρους της μητέρας κατά την τελευταία εμμηνορροϊκή περίοδο κατά ένα κιλό αναμένεται να επιφέρει ελάττωση κατά 0.015 στα log odds της low birth weight=1. Στη στήλη Exp(B) δίνονται τα odds ratio για τις ανεξάρτητες μεταβλητές. Στη στήλη S.E. δίνεται το τυπικό σφάλμα του εκτιμητή των συντελεστών. Διαιρώντας την τιμή του εκτιμητή με το τυπικό σφάλμα μπορούμε να

αποκτήσουμε την τιμή ενός t στατιστικού για τον έλεγχο της υπόθεσης ότι ο συντελεστής είναι ίσος με μηδέν. Επιπλέον μπορεί να χρησιμοποιηθεί η τιμή του S.E. για την κατασκευή διαστημάτων εμπιστοσύνης για τους συντελεστές. Για παράδειγμα ένα 95% διάστημα εμπιστοσύνης για το συντελεστή της μεταβλητής lwt είναι: $-0,015 \pm 1.96 * 0.007$

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1(a)	AGE	-,030	,037	,637	1	,425	,971
	LWT	-,015	,007	4,969	1	,026	,985
	RACE			7,116	2	,028	
	RACE(1)	-,880	,441	3,990	1	,046	,415
	RACE(2)	,392	,538	,531	1	,466	1,480
	SMOKE(1)	-,939	,402	5,450	1	,020	,391
	PTL	,543	,345	2,474	1	,116	1,722
	HT(1)	-1,863	,698	7,136	1	,008	,155
	UI	,768	,459	2,793	1	,095	2,155
	FTV	,065	,172	,143	1	,705	1,067
	Constant	4,163	1,442	8,334	1	,004	64,281

a Variable(s) entered on step 1: AGE, LWT, RACE, SMOKE, PTL, HT, UI, FTV.

Ας προσαρμόσουμε στη συνέχεια το μοντέλο χωρίς τις μεταβλητές age, ptl, ui ftv. Τότε προκύπτουν τα ακόλουθα:

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	26,425	5	,000
	Block	26,425	5	,000
	Model	26,425	5	,000

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	208,247(a)	,130	,183

a Estimation terminated at iteration number 5 because parameter estimates changed by less than ,001.

Classification Table(a)

Observed			Predicted		
			low birth weight		Percentage Correct
			>=2500g	<2500g	
Step 1	low birth weight	>=2500g	123	7	94,6
		<2500g	43	16	27,1
	Overall Percentage				73,5

a The cut value is ,500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1(a)	LWT	-,018	,007	6,937	1	,008	,982
	RACE			8,095	2	,017	
	RACE(1)	-,944	,423	4,968	1	,026	,389
	RACE(2)	,344	,536	,411	1	,521	1,411
	SMOKE(1)	-1,072	,388	7,646	1	,006	,342
	HT(1)	-1,749	,691	6,411	1	,011	,174
	Constant	4,116	1,252	10,817	1	,001	61,339

a Variable(s) entered on step 1: LWT, RACE, SMOKE, HT.

Θέλουμε αρχικά να συγκρίνουμε το νέο μοντέλο με το αρχικό μοντέλο που περιέχει όλες τις μεταβλητές. Θα χρησιμοποιήσουμε το γεγονός ότι:

$$G = -2[\ln l_{νέο} - \ln l_{αρχικό}] \sim X_r^2,$$

Όπου r το πλήθος των παραμέτρων που δεν συμπεριλαμβάνονται στο νέο μοντέλο. Επομένως για το παράδειγμά μας είναι $G \sim X_4^2$, $G = 208{,}247 - 201{,}285 = 6{,}962$ με $P(G \geq 6{,}962) = 0{,}14$. Καθώς η p- τιμή είναι μεγαλύτερη από 0.05 προκύπτει ότι το μοντέλο που δεν περικλείει τις 4 προαναφερθείσες ανεξάρτητες μεταβλητές είναι το ίδιο καλό με το αρχικό.

Επομένως το νέο μοντέλο που προκύπτει είναι το ακόλουθο:

$$\ln\left(\frac{\hat{p}}{1-\hat{p}}\right) = 4.116 - 0.018LWT - 0.944RACE(1) + 0.344RACE(2) - 1.072SMOKE(1) - 1.749HT(1)$$

Έτσι για μία μητέρα με βάρος κατά την τελευταία εμμηνορροϊκή περίοδο 60 κιλά που είναι λευκή, δεν κάπνιζε κατά την εγκυμοσύνη ενώ είχε ιστορικό υπέρτασης προκύπτει η ακόλουθη πρόβλεψη:

$$\ln\left(\frac{\hat{p}}{1-\hat{p}}\right) = 4.116 - 0.018*50 - 0.944*1 + 0.344*0 - 1.072*1 - 1.749*0,$$

δηλαδή

$$\ln\left(\frac{\hat{p}}{1-\hat{p}}\right) = 4.116 - 0.9 - 0.944 - 1.072 = 1.2.$$

Επομένως αφού Ln(odds)=1.2 μετά από πράξεις έχουμε ότι

$$\hat{p} = \frac{\exp(1.2)}{1+\exp(1.2)} = 0,77.$$

Επομένως αυτή η γυναίκα έχει 77% πιθανότητα να γεννήσει παιδί με βάρος μικρότερο των 2500 γραμμαρίων.

Επιμέρους συγκρίσεις-συμπεράσματα:

α) Έστω ότι δύο γυναίκες διαφέρουν μόνο κατά την μεταβλητή LWT. Η μία είναι βαρύτερη κατά 10 κιλά. Τι συμβαίνει τότε?

Εύκολα αποδεικνύεται ότι η διαφορά στους Log(odds) της βαρύτερης από τη λεπτότερη είναι ίση με -0,18. Επομένως προκύπτει ότι

$$odds\ ratio = \frac{odds\ \text{βαρύτερης}}{\text{odds λεπτότερης}} = \exp\left(10\hat{\beta}_1\right) = \exp(-0.18) = 0,84.$$

Σχόλιο: Ένα odds ratio ίσο με ένα υποδεικνύει ότι το ενδεχόμενο που μελετούμε είναι ισοπίθανο στα δύο γκρουπ. Ένα odds ratio μεγαλύτερο (μικρότερο αντίστοιχα) του 1 ότι το ενδεχόμενο είναι πιο πιθανό στο πρώτο (στο δεύτερο αντίστοιχα) γκρουπ.

β) Σύγκριση μεταξύ καπνιστών μη καπνιστών.

Έστω ότι δύο γυναίκες διαφέρουν μόνο ως προς το αν καπνίζουν ή όχι. Τότε

$$odds\ ratio = \frac{odds\ \text{χαρακτηριστικό κωδικοποιήθηκε ως 1}}{\text{odds χαρακτηριστικό κωδικοποιήθηκε ως 0}} = 0.342$$

$$odds\ ratio = \frac{odds\ \text{μη καπνίστριας}}{\text{odds καπνίστριας}} = 0.342$$

γ) Συγκρίσεις ως προς το γένος.

Λευκοί-Άλλοι:

$$odds\ ratio = \frac{odds\ \text{λευκοί}}{\text{odds άλλοι}} = 0.389\,.$$

Μαύροι-Άλλοι

$$odds\ ratio = \frac{odds\ \text{μαύροι}}{\text{odds άλλοι}} = 1.411$$

Λευκοί-Μαύροι:

$$odds\ ratio = \frac{odds\ \text{λευκοί}}{\text{odds μαύροι}} = \frac{0.389}{1.411} = 0.27$$

δ) Συγκρίσεις ως προς το ιστορικό υπέρτασης.

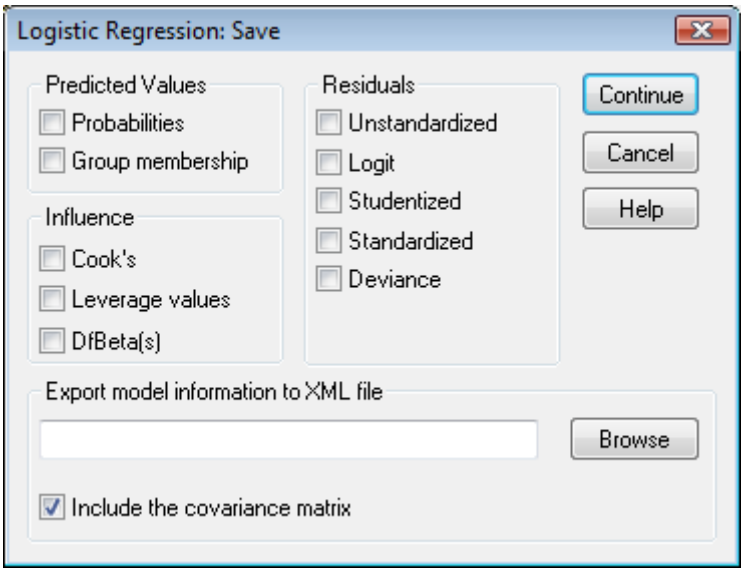
$$odds\ ratio = \frac{odds\ \text{όχι υπερτασικοί}}{\text{odds υπερτασικοί}} = 0.174$$

Τα παραπάνω odds ratio μετατρέπονται σε ποσοστά χρησιμοποιώντας τη σχέση

$$p = \frac{odds\ \text{ratio}}{1+\text{odds ratio}}\,.$$

<u>Άλλες διαθέσιμες επιλογές της διαδικασίας Binary Logistic</u>

1. Από το πλαίσιο Binary Logistic Regression Save



μας δίνεται η δυνατότητα για αποθήκευση ως νέων μεταβλητών στο αρχείο των δεδομένων μας των ακόλουθων ποσοτήτων:

Predicted Values. Αποθηκεύονται οι εκτιμώμενες τιμές του μοντέλου. Οι διαθέσιμες επιλογές είναι:.

- Probabilities. Για κάθε πειραματική μονάδα αποθηκεύεται η προβλεπόμενη πιθανότητα πραγματοποίησης του ενδεχομένου.

Για το παράδειγμά μας προκύπτει π.χ ότι για τη μητέρα με κωδικό 85 είναι 0,2873. Αυτό προέκυψε ως εξής:

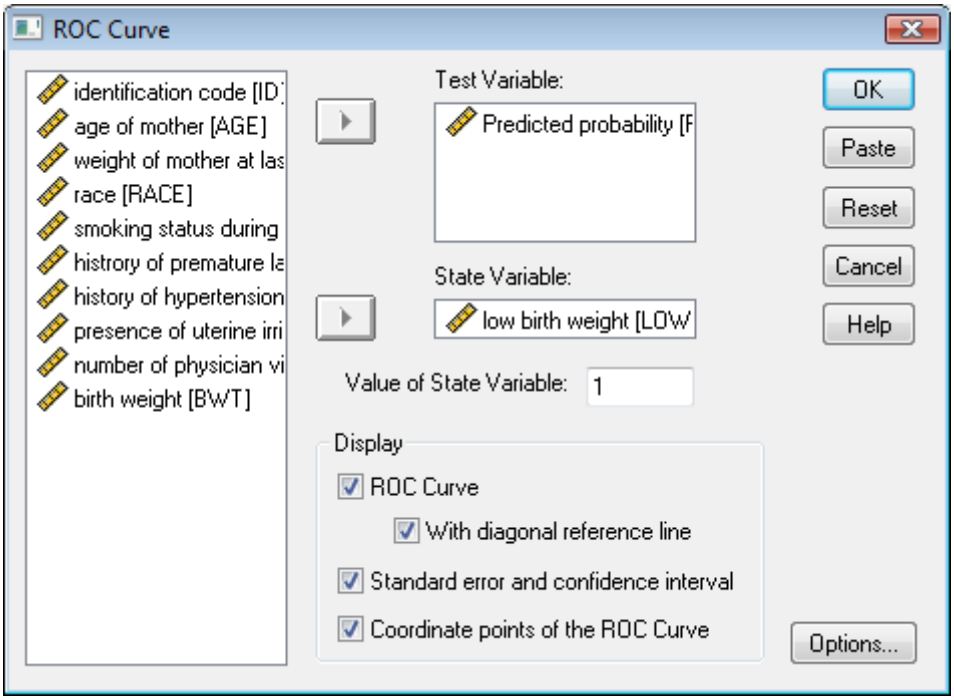
$$\ln\left(\frac{\hat{p}}{1-\hat{p}}\right) = 4.116 - 0.018*182 + 0.344*1 - 1.072*1 - 1.749*1$$
$$= 4.116 - 3.276 + 0.344 - 1.072 - 1.749 = -1.637 .$$

Επομένως

$$\hat{p} = \frac{\exp(-1.637)}{1+\exp(-1.637)} = 0.16 .$$

<u>Κατασκευή ROC Curve</u>

Χρησιμοποιώντας τις αποθηκευμένες προβλεπόμενες πιθανότητες πραγματοποίησης του ενδεχομένου κατασκευάζουμε την ROC Curve. Από το κεντρικό μενού επιλέγουμε Analyze Roc Curve και στη συνέχεια τα ακόλουθα:



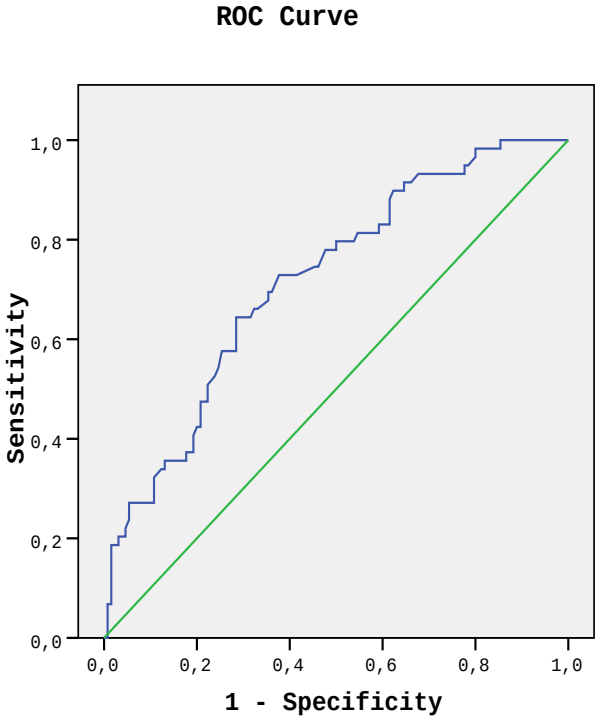
Το αποτέλεσμα είναι το ακόλουθο

Case Processing Summary

low birth weight	Valid N (listwise)
Positive(a)	59
Negative	130

Larger values of the test result variable(s) indicate stronger evidence for a positive actual state.
a The positive actual state is <2500g.

Από τον πίνακα case Processing Summary έχουμε ότι 59 νεογνά από το δείγμα μας έχουν βάρος μικρότερο από 2500 γραμμάρια και 130 βάρος μεγαλύτερο των 2500 γραμμαρίων.



Diagonal segments are produced by ties.

Area Under the Curve

Test Result Variable(s): Predicted probability

Area	Std. Error(a)	Asymptotic Sig.(b)	Asymptotic 95% Confidence Interval	
			Upper Bound	Lower Bound
,718	,039	,000	,641	,794

The test result variable(s): Predicted probability has at least one tie between the positive actual state group and the negative actual state group. Statistics may be biased.

a Under the nonparametric assumption

b Null hypothesis: true area = 0.5

Στο πλαίσιο Area έχουμε ότι η περιοχή κάτω από την ROC είναι ίση με 0.718, που σημαίνει ότι 71,8% των πιθανών ζευγαριών όπου κάποιο νεογνό έχει βάρος μικρότερο των 2500 γραμμαρίων και το άλλο μεγαλύτερο το μοντέλο θα επιφορτίσει με μεγαλύτερη πιθανότητα αυτό με βάρος μεγαλύτερο των 2500. Επιπλέον, η p-τιμή είναι μικρότερη από 0.05, το οποίο σημαίνει ότι η χρησιμοποίηση του μοντέλου είναι καλύτερη από το στρίψιμο ενός νομίσματος

Σχόλιο: Τιμές μεταξύ 0.50 το 0.75 υποδηλώνουν ένα μέτριο μοντέλο, τιμές μεταξύ 0.75 έως 0.92 ένα καλό μοντέλο, τιμές από 0.92 έως 0.97 ένα πολύ καλό μοντέλο, ενώ τιμές από 0.97 έως 1 ένα εξαιρετικό μοντέλο. Οι καμπύλες αυτές μπορούν να χρησιμοποιηθούν και για την αποτίμηση ποιου μοντέλου είναι καλύτερο για την ταξινόμηση των πειραματικών μονάδων.

Στον πίνακα Coordinates of the Curve δίνονται οι τιμές των sensitivity και 1-specificity για κάθε δυνατή τιμή του cut off point. Έτσι δίνεται η δυνατότητα επιλογής εκείνου του cut-off point, η οποία αντιστοιχεί στην επιθυμητή τιμή του δείκτη sensitivity και specificity (ή 1-specificity).

- Predicted Group Membership. Δίνεται η πρόβλεψη σε ποιο από τα δύο γκρουπ ανήκει η πειραματική μονάδα. Η πρόβλεψη αυτή στηρίζεται στις εκτιμώμενες πιθανότητες που υπολογίστηκαν πρωτύτερα. Αν είναι η πιθανότητα μεγαλύτερη του 0.5 τότε ταξινομείται στο γκρουπ όπου Y=1 δηλαδή εκεί όπου έχω βάρος μικρότερο των 2500 γραμμαρίων για το παράδειγμά μας.

Έτσι για τη μητέρα με κωδικό 85 θα προέκυπτε ότι ταξινομείται στο 0 γκρουπ. Δηλαδή ότι το νεογνό θα είχε βάρος μεγαλύτερο των 2500 γραμμαρίων. (σωστή πρόβλεψη όντως).

Από το πλαίσιο Residuals μας δίνεται η δυνατότητα για την αποθήκευση των ακόλουθων ποσοτήτων:

- Unstandardized Residuals. Η διαφορά μεταξύ της παρατηρούμενης τιμής και της τιμής που προβλέπεται από το μοντέλο.

Έτσι για την μητέρα με κωδικό 85 γνωρίζουμε ότι η παρατηρούμενη τιμής της Y είναι 0. Επομένως η τιμή στη στήλη των Unstandardized residuals είναι $0 - \text{Predicted} = 0 - 0.16 = -0.16$

- Logit Residual. Το υπόλοιπο για την πειραματική μονάδα αν η πρόβλεψη γίνει στην Logit κλίμακα. Προκύπτει από τις τιμές των Unstandardized Residuals διαιρεμένες με το γινόμενο

Προβλεπόμενες πιθανότητες *(1-προβλεπόμενες πιθανότητες).

Έτσι για την μητέρα με κωδικό 85 γνωρίζουμε ότι η παρατηρούμενη τιμής της Y είναι 0. Επομένως η τιμή στη στήλη των Unstandardized residuals είναι $0 - \text{Predicted} = 0 - 0.16 = -0.16$. Τότε προκύπτει ότι η τιμή στη στήλη των Logit Residual είναι:

$$\frac{-0.16}{0.16 * (1 - 0.16)} = -1.19 .$$

- Studentized Residual. Η αλλαγή, μεταβολή στη Deviance του μοντέλου αν δεν συμπεριληφθεί η συγκεκριμένη πειραματική μονάδα.
- Standardized Residuals. Τα υπόλοιπα διαιρεμένα με έναν εκτιμητή της τυπικής τους απόκλισης. Είναι επίσης γνωστά και ως Pearson residuals ή Chi residual, έχουν μέση τιμή 0 και τυπική απόκλιση 1.
- Deviance. Τα υπόλοιπα που βασίζονται στην Deviance του μοντέλου. Για κάθε πειραματική μονάδα υπολογίζεται ένα a log-likelihood-ratio statistic, που μετρά πόσα καλά το μοντέλο προσαρμόζει την συγκεκριμένη πειραματική μονάδα. Δίνονται από τη σχέση:

$$dev_i = \begin{cases} \left\{-2\left[Y_i \ln\left(\hat{p}_i\right) + \left(1-Y_i\right)\ln\left(1-\hat{p}_i\right)\right]\right\}^{1/2}, & Y_i \geq \hat{p}_i \\ -\left\{-2\left[Y_i \ln\left(\hat{p}_i\right) + \left(1-Y_i\right)\ln\left(1-\hat{p}_i\right)\right]\right\}^{1/2}, & Y_i < \hat{p}_i \end{cases} .$$

Σχόλιο: Τα Standardized Residuals χρησιμοποιούνται κατά τον γνωστό τρόπο για τον έλεγχο ακραίων τιμών. Η ανάλυση των υπολοίπων μπορεί να μας οδηγήσει στην ανάπτυξη διαφορετικών μοντέλων για διαφορετικές ομάδες, τύπους των πειραματικών μονάδων

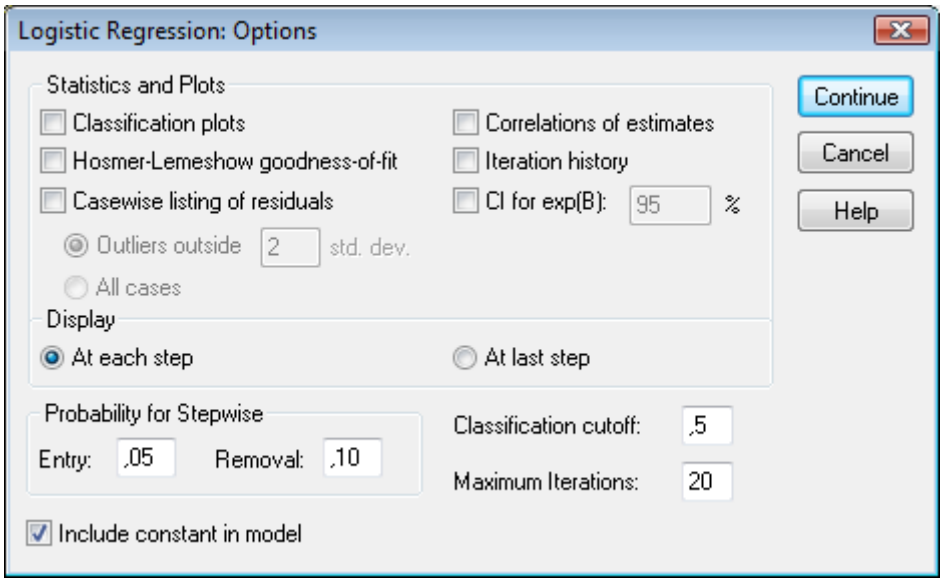
Από το πλαίσιο Influence έχουμε τη δυνατότητα αποθήκευσης μέτρων για τον έλεγχο της επίδρασης κάθε πειραματικής μονάδας στις προβλεπόμενες τιμές. Οι διαθέσιμες επιλογές είναι οι ακόλουθες:

- Cook's. Υπολογίζεται η αντίστοιχη για την λογιστική παλινδρόμηση απόσταση του Cook. Πρόκειται για ένα διαγνωστικό μέτρο που καθορίζει πόσο θα αλλάξουν τα υπόλοιπα όλα των πειραματικών μονάδων αν η συγκεκριμένη παρατήρηση δεν λαμβάνονταν υπόψη στον υπολογισμό των συντελεστών της παλινδρόμησης. (Μεγάλες τιμές αυτού του δείκτη υποδηλώνουν επηρεάζουσα παρατήρηση)
- Leverage Value. Ένα μέτρο της επίδρασης της κάθε παρατήρησης στην προσαρμογή του μοντέλου. Η τιμή αυτού του στατιστικού για οποιαδήποτε πειραματική μονάδα μπορεί να συγκριθεί με την ποσότητα $\frac{k+1}{n}$ όπου k το

πλήθος των ανεξάρτητων μεταβλητών στο μοντέλο (λαμβάνουμε υπόψη και τις πιθανές δείκτριες). Είναι αξιοσημείωτο ότι επηρεάζουσες παρατηρήσεις μπορούν μολαταύτα να έχουν μικρές τιμές στο δείκτη Leverage για περιπτώσεις όπου οι εκτιμώμενες πιθανότητες είναι μικρότερες του 0,1 ή μεγαλύτερες του 0.9. Το γράφημα των τιμών αυτού του διαγνωστικού μέτρου με τον αύξων αριθμό θα μας δώσει γρήγορα και εποπτικά τις επηρεάζουσες παρατηρήσεις.

- DfBeta(s). Η διαφορά στους εκτιμητές των συντελεστών της παλινδρόμησης αν εξαιρεθεί η συγκεκριμένη παρατήρηση. Ένα αυθαίρετο cut off point που έχει εισαχθεί ως κριτήριο για το αν μία παρατήρηση είναι επηρεάζουσα ή όχι είναι η τιμή 1.

2. Από το πλαίσιο **Binary Logistic Options** έχουμε τη δυνατότητα για τα ακόλουθα:

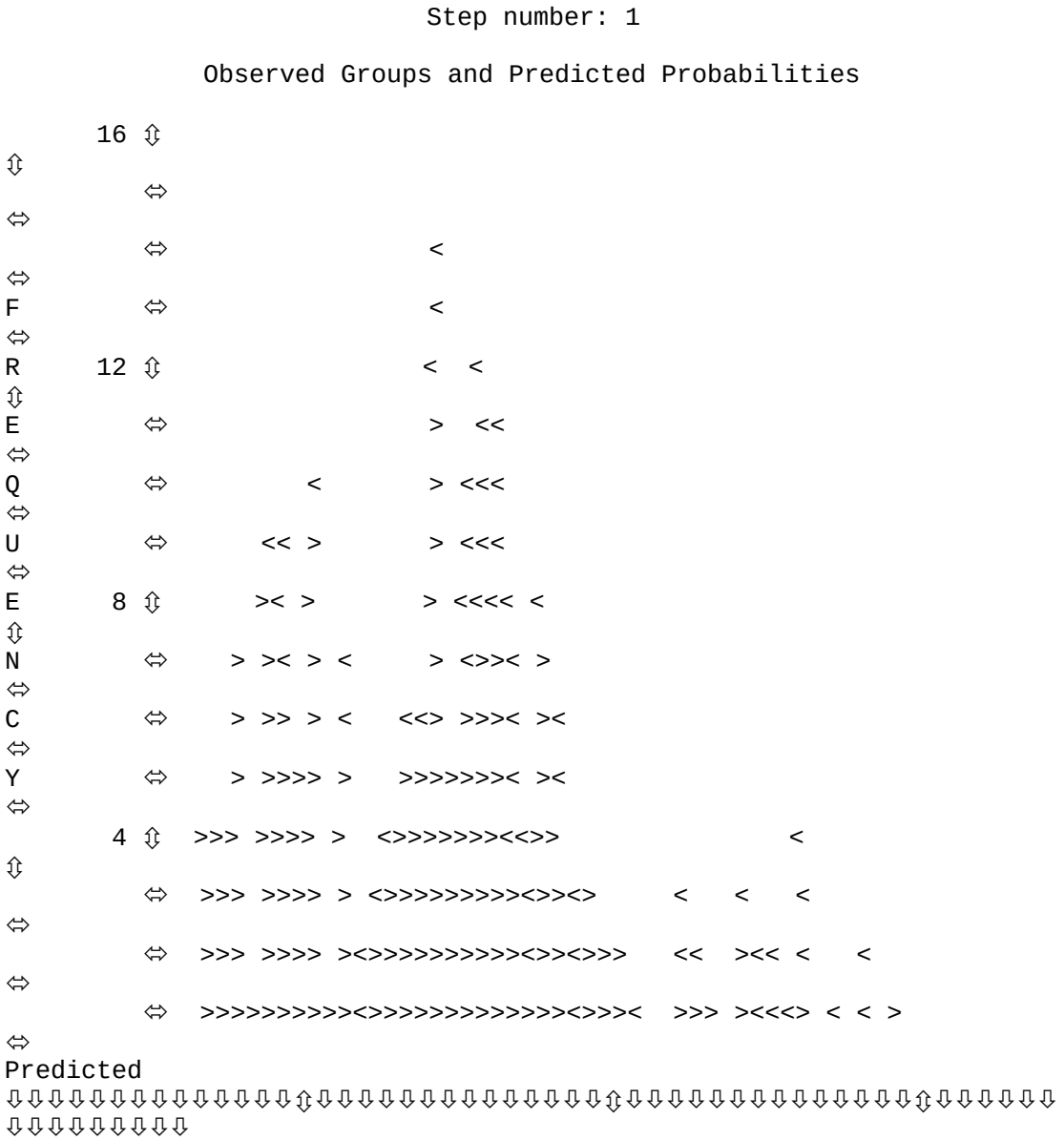


Classification plots: γράφημα των παρατηρούμενων γκρουπ και των προβλεπόμενων πιαθανοτήτων.

Hosmer-Lemeshow goodness-of-fit: Είναι ένα τεστ καλής προσαρμογής του μοντέλου διαφορετικό και πιο ανθεκτικό από το τεστ καλής προσαρμογής που χρησιμοποιείται στην λογιστική παλινδρόμηση, ειδικά για μοντέλα με συνεχείς

ανεξάρτητες μεταβλητές και μελέτες με μικρό μέγεθος δείγματος (προτείνεται να επιλέγεται).

Για το παράδειγμα μας προκύπτουν τα ακόλουθα



Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	6,521	8	,589

Επειδή η p-τιμή του τεστ είναι μεγαλύτερη από 0.05, δεν απορρίπτεται η μηδενική υπόθεση ότι δεν υπάρχει διαφορά μεταξύ των παρατηρούμενων και των εκτιμώμενων τιμών. Αυτό συνεπάγεται ότι το μοντέλο είναι στατιστικά σημαντικό.

Casewise listing of residuals: μας δίνονται ποιες πειραματικές μονάδες έχουν Studentized residuals με απόλυτη τιμή μεγαλύτερη του 2.

Casewise List(b)

Case		Observed			Temporary Variable	
	Selected Status(a)	Resid	Predicted	Predicted Group	Resid	ZResid
132	S	<**	,122	>	,878	2,686
147	S	<**	,125	>	,875	2,640
155	S	<**	,107	>	,893	2,885
183	S	<**	,121	>	,879	2,689

a S = Selected, U = Unselected cases, and ** = Misclassified cases.
b Cases with studentized residuals greater than 2,000 are listed.

Correlations of estimates:

Μας δίνονται οι συντελεστές συσχέτισεις των εκτιμητών των συντελεστών των παραμέτρων της παλινδρόμησης.

Correlation Matrix

		Constant	LWT	RACE(1)	RACE(2)	SMOKE(1)	HT(1)
Step 1	Constant	1,000	-,802	-,169	,049	-,234	-,739
	LWT	-,802	1,000	-,130	-,287	-,041	,336
	RACE(1)	-,169	-,130	1,000	,429	,484	-,005
	RACE(2)	,049	-,287	,429	1,000	,176	-,030
	SMOKE(1)	-,234	-,041	,484	,176	1,000	,018
	HT(1)	-,739	,336	-,005	-,030	,018	1,000

Iteration history

Μας δίνονται οι εκτιμητές των συντελεστών του μοντέλου της παλινδρόμησης για κάθε επανάληψη μέχρι τον τερματισμό/ Υπενθυμίζουμε ότι όταν οι συντελεστές από μία επανάληψη στην άλλη δεν μεταβάλλονται περισσότερο από 0.001 τότε τερματίζει η διαδικασία.

Iteration History(a,b,c,d)

Iteration		-2 Log likelihood	Coefficients					
		Constant	LWT	RACE(1)	RACE(2)	SMOKE(1)	HT(1)	Constant
Step 1	1	210,452	2,933	-,012	-,658	,292	-,781	-1,403
	2	208,285	3,949	-,017	-,905	,339	-1,033	-1,703
	3	208,247	4,113	-,018	-,943	,344	-1,071	-1,748
	4	208,247	4,116	-,018	-,944	,344	-1,072	-1,749
	5	208,247	4,116	-,018	-,944	,344	-1,072	-1,749

a Method: Enter
b Constant is included in the model.
c Initial -2 Log Likelihood: 234,672
d Estimation terminated at iteration number 5 because parameter estimates changed by less than ,001.

CI for exp(B)

Μας δίνονται τα κάτω και άνω όρια του 95% διαστήματος εμπιστοσύνης για το Odds ratio. Αν περιέχεται το 1 τότε τι υποδηλώνεται για την αντίστοιχη μεταβλητή;

Variables in the Equation

		95,0% C.I.for EXP(B)	
		Lower	Upper
Step 1(a)	LWT	,969	,995
	RACE		
	RACE(1)	,170	,892
	RACE(2)	,493	4,036
	SMOKE(1)	,160	,732
	HT(1)	,045	,674
	Constant		

a Variable(s) entered on step 1: LWT, RACE, SMOKE, HT.

Από το πλαίσιο Display επιλέγουμε την εμφάνιση των αποτελεσμάτων είτε για κάθε βήμα είτε για το τελευταίο βήμα.

Probability for Stepwise. Μας δίνεται η δυνατότητα να καθορίσουμε τα κριτήρια με τα οποία μία μεταβλητή εισάγετε ή βγαίνει από το μοντέλο. Πιο συγκεκριμένα μια μεταβλητή εισέρχεται στο μοντέλο αν η πιθανότητα του αντίστοιχου Score στατιστικού για αυτήν είναι μικρότερη από την τιμή στο πλαίσιο Entry value και βγαίνει από το μοντέλο αν η πιθανότητα είναι μεγαλύτερη από την τιμή στο πλαίσιο Removal value. (Entry value μικρότερη της Removal value).

Classification cutoff. Μας δίνεται η δυνατότητα καθορισμού του cut point για την ταξινόμηση των πειραματικών μονάδων. Πειραματικές μονάδες με εκτιμώμενες τιμές μεγαλύτερες του cutoff point ταξινομούνται στην κατηγορία για την οποία έχουμε Y=1. Η προεπιλεγμένη τιμή διαφοροποιείται εισάγοντας έναν αριθμό μεταξύ 0.01 και 0.99. Έτσι κάποιος που ενδιαφέρεται για το μοντέλο εκείνο που θα έχει μεγαλύτερη πιθανότητα ορθής ταξινόμησης αρκεί να βρει το cut off point για το οποίο επιτυγχάνεται αυτό.

Maximum Iterations. Μας επιτρέπει να μεταβάλλουμε τον μέγιστο αριθμό των φορών που το μοντέλο επαναλαμβάνεται μέχρι να τερματίσει.

Include constant in model. Μας επιτρέπει να καθορίσουμε αν περιλαμβάνεται σταθερός όρος στο μοντέλο που προσαρμόζεται ή όχι.

Βιβλιογραφία

1. Hosmer, D. and Lemeshow, S. (2000). Applied Logistic Regression. Wiley.
2. Afifi, A. A and Clark, V. (1998). Computer-Aided Multivariate analysis. Chapman and Hall.
3. Elizabeth R. Brown (2004). http://courses.washington.edu/b515.
4. David Garson (2007). http://www2.chass.ncsu.edu/garson/pa765/index.htm.