Η ΔΙΕΡΕΥΝΗΣΗ ΤΗΣ ΕΓΚΑΤΑΣΤΑΣΗΣ ΣΤΗΝ ΑΓΟΡΑ ΕΡΓΑΣΙΑΣ ΤΩΝ ΑΠΟΦΟΙΤΩΝ ΤΟΥ ΑΝΩΤΕΡΟΥ ΚΥΚΛΟΥ ΤΗΣ ΔΕΥΤΕΡΟΒΑΘΜΙΑΣ ΕΚΠΑΙΔΕΥΣΗΣ ΠΟΥ ΔΕ ΣΥΝΕΧΙΣΑΝ ΣΠΟΥΔΕΣ ΣΤΗΝ ΤΡΙΤΟΒΑΘΜΙΑ ΕΚΠΑΙΔΕΥΣΗ ΜΕ ΧΡΗΣΗ ΜΟΝΤΕΛΟΥ ΛΟΓΙΣΤΙΚΗΣ ΠΑΛΙΝΔΡΟΜΗΣΗΣ.

Παναγιώτης Ρουσέας

ΓΕΝΙΚΑ ΓΙΑ ΤΗ ΛΟΓΙΣΤΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ (LOGISTIC REGRESSION)

Η λογιστική παλινδρόμηση είναι μια μέθοδος πολυπαραγοντικής στατιστικής ανάλυσης (multivariate statistical analysis) που χρησιμοποιεί ένα σύνολο ανεξαρτήτων μεταβλητών (independent variables) για τη διερεύνηση της κίνησης μιας κατηγορικής εξαρτημένης μεταβλητής (dependent variable).

Η λογιστική παλινδρόμηση (Logistic Regression) είναι χρήσιμη σε καταστάσεις στις οποίες επιθυμούμε την πρόβλεψη της ύπαρξης ή της απουσίας ενός χαρακτηριστικού ή ενός συμβάντος. Η πρόβλεψη αυτή βασίζεται στην κατασκευή ενός γραμμικού μοντέλου και συγκεκριμένα στον προσδιορισμό των τιμών που παίρνουν οι συντελεστές ενός συνόλου (set) ανεξάρτητων μεταβλητών που χρησιμοποιούνται ως μεταβλητές πρόβλεψης (predictor variables). Εκτός από την πρόβλεψη ένα μοντέλο λογιστικής παλινδρόμησης δίνει τη δυνατότητα να εκτιμήσουμε την επίδραση κάθε ανεξάρτητης μεταβλητής στη διαμόρφωση των τιμών της εξαρτημένης μεταβλητής.

Στη λογιστική παλινδρόμηση, σε αντίθεση με την πολλαπλή παλινδρόμηση (multiple regression) είναι δυνατό να χρησιμοποιηθούν ως εξαρτημένες μεταβλητές εκτός από αναλογικές αριθμητικές μεταβλητές (ratio scale) και κατηγορικές μεταβλητές (nominal scale).

Η πιο διαδεδομένη βιβλιογραφικά έκφραση της λογιστικής παλινδρόμησης είναι

$$\ln(ods)=a+b_1x_1+ b_2 x_2 +\ldots\ldots\ldots+ b_k x_k$$

Το δεξί μέρος της εξίσωσης δημιουργείται από ένα γραμμικό συνδυασμό των ανεξάρτητων μεταβλητών που συμμετέχουν στο μοντέλο παλινδρόμησης. Το αριστερό μέρος περιέχει τις τιμές της εξαρτημένης μεταβλητής με τη μορφή του λογαρίθμου των odds, δηλαδή του λογαρίθμου της σχέσης odds=Prob/(1-Prob). Το odds εναλλακτικά ονομάζεται logit και ο όρος Prob εκφράζει την πιθανότητα του συμβάντος του γεγονότος. Οι συντελεστές των ανεξάρτητων μεταβλητών στην εξίσωση παλινδρόμησης εκτιμούνται με βάση τη μέθοδου Μεγίστης Πιθανοφάνειας. Σύμφωνα με τη μέθοδο αυτή η τιμή των συντελεστών των ανεξάρτητων μεταβλητών είναι αυτή που κάνει τις παρατηρηθείσες τιμές της εξαρτημένης μεταβλητής πιο πιθανές, βάσει του συνόλου (set) των ανεξαρτήτων μεταβλητών.

Η ΚΑΤΑΣΚΕΥΗ ΤΟΥ ΜΟΝΤΕΛΟΥ

ΔΕΙΓΜΑ: 4986 απόφοιτοι τεσσάρων διαφορετικών τύπων σχολείων 10 περίπου χρόνια μετά την αποφοίτησή τους.

ΤΥΠΟΣ ΣΧΟΛΕΙΟΥ

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	ΤΕΣ	971	19,5	19,5	19,5
	ΤΕΛ	1664	33,4	33,4	52,8
	ΕΠΛ	1024	20,5	20,5	73,4
	ΓΕΛ	1327	26,6	26,6	100,0
	Total	4986	100,0	100,0	

Εξαρτημένη μεταβλητή: **erg_bin** [διχοτομική (binary) μεταβλητή]

ERG_BIN ΕΡΓΑΣΙΑΚΗ ΚΑΤΑΣΤΑΣΗ

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	,0 Μη εργαζόμενος	939	18,8	18,8	18,8
	1,0 Εργαζόμενος	4047	81,2	81,2	100,0
	Total	4986	100,0	100,0	

Ανεξάρτητες μεταβλητές

A/A	NAME	SCALE	LABEL	VALUES
1	Typos_sx	Nominal	ΤΥΠΟΣ ΣΧΟΛΕΙΟΥ	1: ΤΕΣ 2: ΤΕΛ 3: ΕΠΛ 4: ΓΕΛ
2	Ekp_kat	Binary	ΕΚΠΑΙΔΕΥΣΗ-ΚΑΤΑΡΤΙΣΗ ΜΕΤΑ ΤΗΝ ΑΠΟΦΟΙΤΗΣΗ	0: ΟΧΙ 1: ΝΑΙ
3	language	Binary	ΞΕΝΗ ΓΛΩΣΣΑ	0: ΟΧΙ 1: ΝΑΙ
4	Erg_sx	Binary	ΕΡΓΑΣΙΑ ΚΑΤΑ ΤΗΝ ΔΙΑΡΚΕΙΑ ΤΩΝ ΣΠΟΥΔΩΝ ΣΤΟ ΣΧΟΛΕΙΟ	0: ΟΧΙ 1: ΝΑΙ
5	sex	Binary	ΦΥΛΟ	0: ΓΥΝΑΙΚΑ 1: ΑΝΔΡΑΣ
6	Ekp_pat	Ratio	ΕΠΙΠΕΔΟ ΕΚΠΑΙΔΕΥΣΗΣ ΠΑΤΕΡΑ	1-9
7	Grade_g	Ratio	ΒΑΘΜΟΣ ΑΠΟΛΥΤΗΡΙΟΥ ΓΥΜΝΑΣΙΟΥ	10-20
8	astikot	Nominal	ΑΣΤΙΚΟΤΗΤΑ ΠΕΡΙΟΧΗΣ ΠΑΤΡΙΚΗΣ ΚΑΤΟΙΚΙΑΣ	1: Αστική 2: Ημιαστική 3: Αγροτική
9	Thesi_pa	Nominal	ΘΕΣΗ ΠΑΤΕΡΑ ΣΤΗΝ ΑΓΟΡΑ ΕΡΓΑΣΙΑΣ	1: Μισθωτός 2: Αυτοαπασχολούμενος 3: Άλλο

Επειδή η εξαρτημένη μεταβλητή (erg_bin) είναι διχοτομική, εφαρμόζεται η Binary logistic regression.

Μέθοδος εισαγωγής των εξαρτημένων μεταβλητών: Forward (Wald)

Η κωδικοποίηση των nominal μεταβλητών φαίνεται στον πίνακα που ακολουθεί

Categorical Variables Codings

		Frequency	Parameter coding		
			(1)	(2)	(3)
TYPOS_SX ÔÕÐ Ï Ó Ó× Ï ËÅ£Ï Õ	1,00 ÔÃÓ	906	1,000	,000	,000
	2,00 ÔÂË	1594	,000	1,000	,000
	3,00 ÂÐË	986	,000	,000	1,000
	4,00 ÃÂË	1299	,000	,000	,000
THESI_PA ÈÅÓÇ ÐÁÔÅÑÍÁ ÓÔÇ Í ÁÃ Ï ÑÁ ÅÑÃÁÓ£ÁÓ	1,00 Ì ¢óèùôüò	1256	1,000	,000	
	2,00 Áõôïáðáó÷ïëïýìåíïò	1961	,000	1,000	
	3,00 ëëï	1568	,000	,000	
ASTIKOT ÁÓÔ£Ê Ï ÔÇÔÁ ÐÅÑ£Ï ×ÇÓ ÐÁÔÑ£ÊÇÓ ÊÁÔ Ï ¢Ê£ÁÓ	1 Áóô¢êÞ (10.000 êÜôï¢êï¢ êá¢ Üíù)	3241	1,000	,000	
	2 Çì¢áóô¢êÞ (2.000 Ýùò 9.999 êÜôï¢ê ï¢)	590	,000	1,000	
	3 Áãñïô¢êÞ (ë¢ãüôåñï¢ áðü 2.000 êÜôï¢êï¢)	954	,000	,000	

Στο μοντέλο εισήχθησαν 7 από τις 9 ανεξάρτητες μεταβλητές. Εξαιρέθηκαν οι:

2	Ekp_kat	Binary	ΕΚΠΑΙΔΕΥΣΗ-ΚΑΤΑΡΤΙΣΗ ΜΕΤΑ ΤΗΝ ΑΠΟΦΟΙΤΗΣΗ	0: ΟΧΙ 1: ΝΑΙ
9	Thesi_pa	Nominal	ΘΕΣΗ ΠΑΤΕΡΑ ΣΤΗΝ ΑΓΟΡΑ ΕΡΓΑΣΙΑΣ	1: Μισθωτός 2: Αυτοαπασχολούμενος 3: Άλλο

Στον Πινακα που ακολουθεί απεικονίζονται οι δείκτες καλής προσαρμογής του μοντέλου (αντίστοιχοι με το R^2 της πολλαπλής παλινδρόμησης)

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	3893,934	,150	,241
2	3853,698	,157	,252
3	3834,088	,161	,258
4	3826,171	,162	,260
5	3821,477	,163	,261
6	3812,729	,165	,264
7	3806,864	,166	,266

Ο δείκτης Nagelkerke R Square παίρνει τιμές από 0-1(πλήρης προσαρμογή). Εφόσον μετά την εισαγωγή και της $7^{ης}$ μεταβλητής ο δείκτης αυτός παραμένει χαμηλός, εκτιμούμε ότι το μοντέλο παρουσιάζει μικρή προσαρμογή στα δεδομένα άρα δεν είναι δυνατό να χρησιμοποιηθεί για πρόβλεψη. Είναι δυνατόν όμως να εκτιμήσουμε την επίδραση στη διαμόρφωση των τιμών της εξαρτημένης μεταβλητής καθεμίας από τις 7 ανεξάρτητες μεταβλητές που υπεισήλθαν στην εξίσωση. Η επίδραση αυτή απεικονίζεται στον παρακάτω πίνακα:

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 7[g]	TYPOS_SX (1)	,264	,196	1,808	1	,179	1,302
	TYPOS_SX (2)	-,081	,107	,570	1	,450	,922
	TYPOS_SX (3)	-,244	,109	5,030	1	,025	,783
	LANGUAGE	,455	,099	21,107	1	,000	1,577
	ERGAS_SX	,477	,117	16,605	1	,000	1,611
	SEX	2,187	,108	407,416	1	,000	8,910
	EKP_PAT	,079	,035	4,973	1	,026	1,082
	ASTIKOT			5,991	2	,050	
	ASTIKOT(1)	,002	,106	,000	1	,984	1,002
	ASTIKOT(2)	-,287	,142	4,080	1	,043	,751
	GRADE_G	,081	,028	8,551	1	,003	1,085
	Constant	-1,124	,462	5,930	1	,015	,325

a. Variable(s) entered on step 1: SEX.

b. Variable(s) entered on step 2: LANGUAGE.

c. Variable(s) entered on step 3: ERGAS_SX.

d. Variable(s) entered on step 4: GRADE_G.

e. Variable(s) entered on step 5: EKP_PAT.

f. Variable(s) entered on step 6: TYPOS_SX.

g. Variable(s) entered on step 7: ASTIKOT.

ΣΥΝΟΠΤΙΚΗ ΕΡΜΗΝΕΙΑ ΤΟΥ ΠΙΝΑΚΑ

Από όλες τις μεταβλητές που υπεισήλθαν στο μοντέλο και είναι στατιστικά σημαντικές το φύλο (sex) παρουσιάζεται ως ο σημαντικότερος παράγοντας προσδιορισμού της εργασιακής κατάστασης ενός ατόμου (B=2,187).

Ακολουθούν σε μεγάλη απόσταση όμως, οι μεταβλητές ergas_sx (εργασία κατά τη διάρκεια των σπουδών στο σχολείο) με B=0,477 και η language (γνώση ξένης γλώσσας) με B=0,445.

Από τις υπόλοιπες μεταβλητές άξια σχολιασμού είναι η μεταβλητή typos_sx. Στη μεταβλητή αυτή αντιστοιχούν 3 dummy μεταβλητές η καθεμία από τις οποίες αντιπροσωπεύει έναν τύπο σχολείου. Σχετικά υψηλός παρουσιάζεται ο συντελεστής της dummy: typos_sx(1) (B=0,264) που αντιπροσωπεύει τις ΤΕΣ. Ο συντελεστής αυτός, παρόλο που δεν είναι στατιστικά σημαντικός, υποδηλώνει μια σαφή κατεύθυνση, δηλαδή ότι οι απόφοιτοι των ΤΕΣ έχουν μεγαλύτερες πιθανότητες να βρίσκονται στην αγορά εργασίας από τους αποφοίτους των άλλων τύπων λυκείων.