

Εργασία  
στα  
Γενικευμένα Γραμμικά Μοντέλα

---

Μ. Παρζακώνης

ΜΕΣ/ 06015

Ο παρακάτω πίνακας δίνει τα αποτελέσματα 800 αιτήσεων για δάνειο σε μία τράπεζα. Ο πίνακας παρουσιάζει τον αριθμό των δανείων που εγκρίθηκαν, ανάλογα με το εισόδημα του πελάτη (με τρεις κατηγορίες: χαμηλό, μέτριο, υψηλό) και μία βαθμολογία που χρησιμοποιεί η τράπεζα και προκύπτει από τα αρχεία της σε σχέση με την οικονομική φερεγγυότητα του πελάτη. Σε κάθε πελάτη δίνεται μία βαθμολογία A, B, ή Γ ανάλογα με το πόσο αξιόπιστος θεωρείται οικονομικά (ο βαθμός A αντιστοιχεί στη μεγαλύτερη αξιοπιστία).

Εισόδημα	Βαθμολογία πελάτη	Έγκριση δανείου	
		Ναι	Όχι
Χαμηλό	A	33	26
	B	36	48
	Γ	17	58
Μέτριο	A	51	45
	B	62	57
	Γ	39	65
Υψηλό	A	73	9
	B	55	28
	Γ	43	45

(α) Να εξεταστεί αν το εισόδημα και η βαθμολογία του πελάτη επηρεάζει την πιθανότητα έγκρισης του δανείου.

(β) Είναι η αλληλεπίδραση μεταξύ των δύο ερμηνευτικών μεταβλητών στο μοντέλο στατιστικά σημαντική;

(γ) Επιλέγοντας ένα κατάλληλο μοντέλο λογιστικής παλινδρόμησης, το οποίο θα αναφέρετε, να δοθούν οι εκτιμώμενες τιμές για το πλήθος των ατόμων που το δανειό τους αναμένεται να εγκριθεί ανάλογα με το εισόδημα και τη βαθμολογία τους.

(δ) Με βάση το ίδιο μοντέλο, να εκτιμηθεί σημειακά η σχετική πιθανότητα έγκρισης δανείου για κάποιον με μέτριο εισόδημα και βαθμολογία Γ.

(ε) Πόσο αυξάνεται (ως ποσοστό, %) η πιθανότητα έγκρισης δανείου σε κάποιον με υψηλό εισόδημα και βαθμολογία Γ σε σχέση με κάποιον με μέτριο εισόδημα και βαθμολογία Γ; Να εκτιμηθεί ο λόγος σχετικής πιθανότητας ανάμεσα στις δύο αυτές κατηγορίες πελατών.

(στ) Για κάποιον ο οποίος έχει χαμηλό εισόδημα, και για καθεμία από τις 3 δυνατές βαθμολογίες αξιολόγησης (A, B, Γ), να δοθεί ένα διάστημα εμπιστοσύνης 95% για την πιθανότητα έγκρισης του δανείου.

(ζ) Κάνοντας και όποιους άλλους ελέγχους θεωρείτε απαραίτητους, να συνοψίσετε τα αποτελέσματα από την παραπάνω ανάλυση για τη σχέση μεταξύ των μεταβλητών με τρόπο ώστε αυτά να είναι κατανοητά για κάποιον ο οποίος δεν είναι ειδικός στη στατιστική.

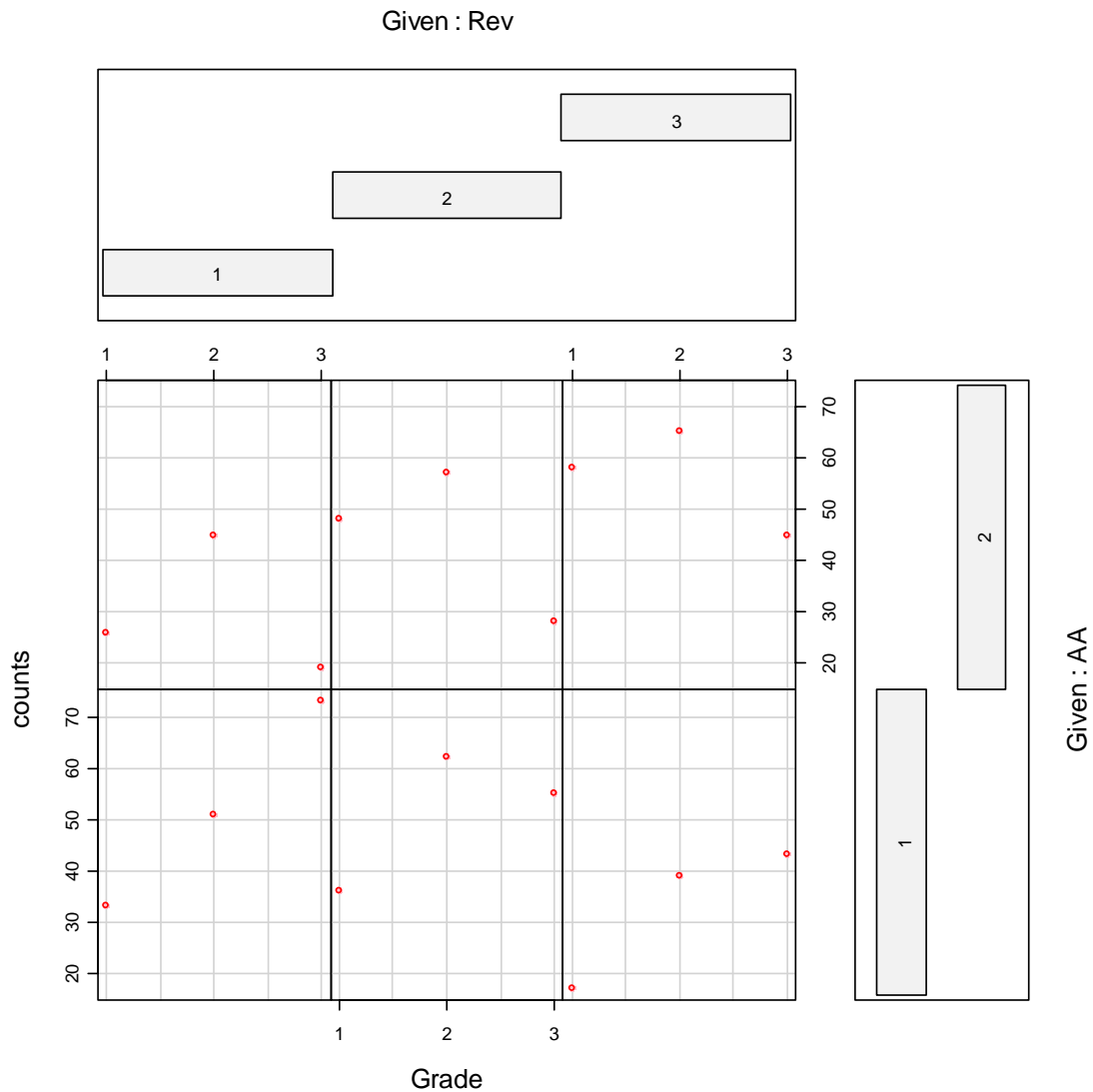
### Σημείωση

Προφανώς, έχει γίνει κάποιο τυπογραφικό λάθος στον αυθεντικό πίνακα της εργασίας

	rev	grade	apr	dapr
1	1	1	33	26
2	1	2	36	48
3	1	3	17	58
4	2	1	51	45
5	2	2	62	57
6	2	3	39	65
<b>7</b>	<b>3</b>	<b>1</b>	<b>73</b>	<b>9</b>
8	3	2	55	28
9	3	3	43	45

Καθώς το άθροισμα των παρατηρήσεων είναι 790 και όχι 800 που αναφέρεται στην εκφώνηση. Λογικά το **9** είναι **19**

Στο επόμενο γράφημα έχουμε την αναπαράσταση του παρατηρούμενο πίνακα



Παρατηρούμε ότι ανάμεσα στις κατηγορίες της βαθμολογίας των αιτούντων υπάρχουν μεγάλες διαφορές

Θέλουμε να ελέγξουμε εάν οι μεταβλητές εισόδημα (*rev*) και βαθμολογία(*grade*) επηρεάζουν την πιθανότητα έγκρισης της αίτησης δανείου. Αυτό μπορούμε να το υλοποιήσουμε «μέσα» από ένα μοντέλο λογιστικής παλινδρόμησης

Το μοντέλο θα είναι (στην γλώσσα του R)

```
apr.total~rev+grade
```

, όπου *apr.total* το ποσοστό επιτυχίας (έγκρισης της αίτησης))

Για να ελέγξουμε τη σημαντικότητα των παραγόντων χρησιμοποιούμε το πίνακα ανάλυσης απόκλισης.

#### Analysis of Deviance Table

Model 1: apr.total ~ 1  
Model 2: apr.total ~ rev + grade

	Resid.	Df	Resid. Dev	Df	Deviance	P(> Chi )
1		8	75.808			
2		6	41.627	2	34.181	3.782e-08

Βλέπουμε ότι σε ε.σ 5% το εισόδημα και η βαθμολογία του συστήματος που εφαρμόζει η τράπεζα είναι σημαντική (από κοινού).

Για να ελέγξουμε την σημαντικότητα κάθε παράγοντα ξεχωριστά έχουμε

#### Analysis of Deviance Table

Model: binomial, link: logit

Response: apr.total

Terms added sequentially (first to last)

	Df	Deviance	Resid.	Df	Resid. Dev	P(> Chi )
NULL				8	75.808	
rev	2	34.181		6	41.627	3.782e-08
grade	2	36.965		4	4.663	9.403e-09

Το συμπέρασμα είναι ότι και οι δύο παράγοντες είναι σημαντικές προσθήκες στο μοντέλο.

Για να ελέγξουμε ένα υπάρχει μια σημαντική αλληλεπίδραση μεταξύ των παραγόντων εισόδημα και βαθμολογία απλά συγκρίνουμε τα δύο μοντέλα

```
apr.total~rev+grade
apr.total~rev*grade
```

Και πάλι μέσω του πίνακα ανάλυσης απόκλισης

#### Analysis of Deviance Table

Model 1: apr.total ~ rev + grade  
Model 2: apr.total ~ rev \* grade

	Resid.	Df	Resid. Dev	Df	Deviance	P(> Chi )
1		4	4.6629			
2		0	4.219e-15	4	4.6629	0.3237

Η αλληλεπίδραση των δύο μεταβλητών δεν είναι σημαντική (αφού η διαφορά τους είναι στατιστικά μηδενική, τα δύο μοντέλα είναι ισοδύναμα)

Για να βρούμε ένα κατάλληλο μοντέλο λογιστικής παλινδρόμησης μπορούμε να αρχίσουμε από το μηδενικό μοντέλο και να προσθέτουμε σε κάθε βήμα και ένα παράγοντα

Δηλαδή

```

apr.total~1
apr.total~rev
apr.total~rev+grade
apr.total~rev*grade

```

και τελικά να συγκρίνουμε το κατά πόσο διαφέρει το ένα από το άλλο (διαδοχικά). Όταν βρεθούν δύο μοντέλα με (στατιστικά) μηδενική διαφορά τότε έχουμε βρει το επιθυμητό

#### Analysis of Deviance Table

```

Model 1: apr.total ~ 1
Model 2: apr.total ~ rev
Model 3: apr.total ~ rev + grade
Model 4: apr.total ~ rev * grade
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1         8      75.808
2         6      41.627  2   34.181 3.782e-08
3         4       4.663  2   36.965 9.403e-09
4         0  4.219e-15  4     4.663  0.324

```

Βλέπουμε ότι το μοντέλο των κύριων επιδράσεων είναι το μοντέλο το οποίο επιλέγεται με αυτή τη μέθοδο (το μοντέλο 3 και 4 είναι ισοδύναμα με την έννοια ότι δεν προσφέρει κάτι η εισαγωγή της αλληλεπίδρασης)

Συνεπώς, το μοντέλο με το οποίο δουλεύουμε από εδώ και πέρα είναι το

```
apr.total ~ rev + grade
```

Οι εκτιμήσεις του μοντέλου είναι

```

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.06047    0.18173   0.333  0.7393
rev2         0.32543    0.18234   1.785  0.0743 .
rev3         1.06498    0.19470   5.470 4.51e-08 ***
grade2      -0.35528    0.18149  -1.958  0.0503 .
grade3      -1.08886    0.18734  -5.812 6.16e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Εκτίμηση της πιθανότητας έγκρισης της αίτησης ενός ατόμου με μέτριο εισόδημα και βαθμολογία Γ είναι η 0.3311575

Για να συγκρίνουμε την ποσοστιαία μεταβολή της πιθανότητας έγκρισης μιας αίτησης όταν μεταβάλλεται η κατηγορία εισοδήματος από μέτριο σε υψηλό και παραμένουμε στη βαθμολογική κλίμακα Γ απλά

```

fit[9,1]/fit[6,1]-1
> 0.5374742

```

Για να εκτιμήσουμε το λόγο σχετικής πιθανότητας υπολογίζουμε τα `logit` των δύο κατηγοριών μέσω της εντολής

```
logits <-predict.glm(g2,newdata=pi2,se.fit=T,type="link")
```

```
      logit      s.e
1  0.06047229 0.04539160
2 -0.29480486 0.04108542
3 -1.02838796 0.03506380
4  0.38590584 0.03868171
5  0.03062869 0.03697878
6 -0.70295442 0.03526597
7  1.12544895 0.03229119
8  0.77017180 0.03660381
9  0.03658870 0.04108872
```

Η διαφορά των `logit` των κατηγοριών θα είναι ο λογάριθμος του λόγου σχετικών πιθανοτήτων

```
logits[9,1] -logits[6,1]
> 0.73955
```

ή αλλιώς (χρησιμοποιώντας μόνο τις εκτιμήσεις των παραμέτρων)

```
logOR.3.2<-1.06498-0.06047+0.32543-0.06047
> 0.73955
```

Και τελικά

```
OR.3.2<-exp(logOR.3.2)
> 2.094993
```

Συνεπώς, ο λόγος πιθανοτήτων έγκρισης της αίτησης ενός ατόμου με μέτριο εισόδημα είναι 2.095 φορές μεγαλύτερος όταν αυτός βρίσκεται στην 3<sup>η</sup> βαθμολογική κλίμακα σε σχέση με το λόγο πιθανοτήτων όταν βρίσκεται στην 2<sup>η</sup> βαθμολογική κλίμακα.

Για να δημιουργούμε τα δ.ε θα χρειαστούμε τα τυπικά σφάλματα των εκτιμήσεων. Με την παρακάτω εντολή δημιουργούμε ένα πίνακα με τις εκτιμήσεις και τα τυπικά σφάλματα τους

```
fit<-
cbind(predict.glm(g2,newdata=pi2,se.fit=T,type="response")$"fit"
      ,predict.glm(g2,newdata=pi2,se.fit=T,type="response")$"se.fit"
      )
```

```
      fit      se.fit
1  0.5151135 0.04539160
2  0.4268280 0.04108542
3  0.2633968 0.03506380
4  0.5952967 0.03868171
5  0.5076566 0.03697878
6  0.3311575 0.03526597
7  0.7549980 0.03229119
8  0.6835581 0.03660381
9  0.5091462 0.04108872
```

Τώρα μπορούμε να βρούμε τα (ασυμπτωτικά) δ.ε

```
fit[1,1]+c(-1,1)*1.96*fit[1,2]
> 0.4261459 0.6040810
fit[2,1]+c(-1,1)*1.96*fit[2,2]
> 0.3463006 0.5073554
fit[3,1]+c(-1,1)*1.96*fit[3,2]
> 0.1946717 0.3321218
```

Τελικά, έχουμε τα επόμενα σημειακά και διαστημικά αποτελέσματα της διαδικασίας εκτίμησης.

(rev, grade)	LB	Estimate	UB
A, A	0.4261459	0.5151135	0.6040810
A, B	0.3463006	0.4268280	0.5073554
A, Γ	0.1946717	0.2633968	0.3321218

για τις κατηγορίες Α,Β,Γ βαθμολογίας και σταθερή την κατηγορία χαμηλού εισοδήματος.

Σε αυτό το σημείο θα ελέγξουμε την καλή προσαρμογή του μοντέλου

```
goft <- function(fit) {
  pv<-pchisq(deviance(fit), df.residual(fit), lower.tail=F)
  cat('\n Pr(>Dev) =', pv, '\n')
}
```

Τελικά έχουμε

```
Pr(>Dev) = 0.3236705
```

Και δεν μπορούμε να απορρίψουμε την απόθεση της καλής προσαρμογής.

Συνεπώς, ο επόμενος πίνακας έχει νόημα να χρησιμοποιηθεί

	rev	grade	apr	dapr	fit.apr	fit.dapr	Total
1	1	1	33	26	30.39169	28.60831	59
2	1	2	36	48	35.85355	48.14645	84
3	1	3	17	58	19.75476	55.24524	75
4	2	1	51	45	57.14849	38.85151	96
5	2	2	62	57	60.41113	58.58887	119
6	2	3	39	65	34.44038	69.55962	104
7	3	1	73	19	69.45982	22.54018	92
8	3	2	55	28	56.73532	26.26468	83
9	3	3	43	45	44.80486	43.19514	88

Επανερχόμενοι στις εκτιμήσεις των logits βλέπουμε ότι οι 6 από τις 9 εκτιμήσεις είναι στατιστικά σημαντική, δηλ η πιθανότητα έγκρισης σε σχέση με την πιθανότητα απόρριψης είναι διαφορετική του 0.5

	logit.fit	logit.se.fit	LB	UB
1	0.06047229	0.04539160	-0.02849525	0.1494398
2	<b>-0.29480486</b>	0.04108542	-0.37533228	-0.2142774
3	<b>-1.02838796</b>	0.03506380	-1.09711302	-0.9596629
4	<b>0.38590584</b>	0.03868171	0.31008968	0.4617220
5	0.03062869	0.03697878	-0.04184973	0.1031071

6	<b>-0.70295442</b>	0.03526597	-0.77207572	-0.6338331
7	<b>1.12544895</b>	0.03229119	1.06215822	1.1887397
8	<b>0.77017180</b>	0.03660381	0.69842833	0.8419153
9	0.03658870	0.04108872	-0.04394519	0.1171226

Από το παραπάνω πίνακα βλέπουμε ότι η μετακίνηση προς τα ανώτερα εισοδηματικά κλιμάκια ( $A \rightarrow B \rightarrow \Gamma$ ) βελτιώνει τις σχετικές πιθανότητες (και των τριών κατηγοριών βαθμολογίας, αυτό είναι το νόημα του μοντέλου των κυριών επιδράσεων).

Επίσης, παρατηρούμε ότι παράλληλα υπάρχει μια σχέση διάταξη ανάμεσα στις βαθμολογικές κατηγορίες. Είναι ξεκάθαρο ότι οι σχετικές πιθανότητες ευνοούν κάποιον όσο πιο «ψηλά» κατατάσσετε στην κλίμακα  $A > B > \Gamma$  (αν με  $>$  συμβολίσουμε το «καλύτερο»)

[Τα ίδια συμπεράσματα προκύπτουν και χρησιμοποιώντας τις πιθανότητες αντί για τα logits]

Τέλος, ελέγχουμε το μοντέλο για παραβιάσεις από τις υποθέσεις.

Προσοχή χρειάζεται στις παρατηρήσεις με επιρροή (influential)

Potentially influential observations of  
 glm(formula = apr.total ~ rev + grade, family = "binomial", data =  
 pi2, weights = total) :

	dfb.1_	dfb.rev2	dfb.rev3	dfb.grd2	dfb.grd3	dffit	cov.r	cook.d	hat
2	0.02_	-0.03	-0.03	0.02	0.00	0.05	10.00_*	0.00	0.58
4	-2.25_*	-2.28_*	0.02	3.25_*	3.03_*	-5.14_*	0.00_	1.01	0.60
5	-0.13_	0.28_	0.03	0.30	0.01_	0.56_	9.24_*	0.08	0.65
6	-0.32	0.84	-0.10	-0.02	1.09_*	1.88	0.48_	0.51	0.58
8	0.10	-0.01	-0.29	-0.25	0.02_	-0.50	5.93_*	0.06	0.51
9	0.10	0.00	-0.31	-0.02	-0.31	-0.61	6.91_*	0.09	0.59

Τέλος, μερικά γραφήματα τα οποία δείχνουν ότι θα χρειαστεί να έχουμε στο μυαλό μας ότι οι υποθέσεις του μοντέλου επιδέχονται κριτική

