Κεφάλαιο 19

Ανάλυση Επιβίωσης

Η ανάλυση επιβίωσης (survival analysis) αναφέρεται στην ανάλυση δεδομένων που αφορούν στο χρόνο που μεσολαβεί μέχρι κάποιο συγκεκριμένο συμβάν. Αρχικά η ανάλυση αναφερόταν στο χρόνο μεταξύ της θεραπείας μέχρι τον θάνατο και για αυτό το λόγο πήρε και το συγκεκριμένο όνομα. Η ανάλυση επιβίωσης όμως μπορεί να εφαρμοστεί σε αρκετές περιπτώσεις, όπως για παράδειγμα στη μηχανολογία, για την ανάλυση του χρόνου μέχρι την εμπλοκή ενός μηχανήματος ή τη γεωργία, για την ανάλυση του χρόνου μέχρι την στιγμή να βγάλει καρπό ένα δέντρο. Στην περίπτωση της μηχανολογίας η ανάλυση αναφέρεται και ως θεωρία αξιοπιστίας (reliability theory). Ο χρόνος επιβίωσης χρίζει ειδικής μεταχείρισης για το λόγο ότι είναι περιορισμένος στο να είναι πάντα θετικός, και γιατί τα δεδομένα περιέχουν λογοκριμένες (cencored) παρατηρήσεις. Τα λογοκριμένα δεδομένα είναι αυτά για τα οποία δεν είναι γνωστός ο χρόνος που συμβαίνει το γεγονός. Το μόνο που μπορεί να λεχθεί είναι ότι ο χρόνος επιβίωσής τους είναι μεγαλύτερος από την τιμή που έχει καταγραφεί.

Στην ανάλυση επιβίωσης είναι πολύ σημαντικές δύο συναρτήσεις οι οποίες περιγράφουν την κατανομή του χρόνου επιβίωσης: η συνάρτηση επιβίωσης και η συνάρτηση κινδύνου.

19.1 Συνάρτηση Επιβίωσης

Συμβολίζοντας το χρόνο επιβίωσης με T, η συνάρτηση επιβίωσης (survival function) $S(t)$ ορίζεται ως η πιθανότητα επιβίωσης ενός ατόμου πέραν τη χρονική

στιγμή t και δίνεται από τη σχέση:

$$S(t) = P(T > t) = 1 - F(t)$$

Η συνάρτηση επιβίωσης είναι μη αρνητική και μη αύξουσα συνάρτηση του t με $S(0) = 1$ και $S(\infty) = 0$. Η γραφική παράσταση της $S(t)$ συναρτήσει του t είναι γνωστή ως καμπύλη επιβίωσης και είναι πολύ σημαντική στην ανάλυση δεδομένων χρόνου επιβίωσης.

19.2 Συνάρτηση Κινδύνου

Η συνάρτηση κινδύνου, $h(t)$, ορίζεται ως η πιθανότητα αποβίωσης (ή πραγμάτωσης του γεγονότος που εξετάζεται) τη χρονική στιγμή t, δεδομένου ότι το άτομο έχει επιβιώσει μέχρι τη χρονική στιγμή t. Δηλαδή,

$$h(t) = \lim_{s \to 0} \frac{P(t \leq T \leq t + s \mid T \geq t)}{s}$$

Η συνάρτηση κινδύνου δίνει ένα μέτρο του πόσο πιθανό είναι ένα άτομο να αποβιώσει ως συνάρτηση της ηλικίας του ατόμου, για παράδειγμα ο κίνδυνος θανάτου ανάμεσα σε αυτούς που είναι ζωντανοί τη συγκεκριμένη στιγμή.

19.3 Μοντέλο αναλόγων συναρτήσεων κινδύνου

Στην ανάλυση επιβίωση παίζει μεγάλο ρόλο η εξεύρεση παραγόντων οι οποίοι να σχετίζονται με το χρόνο επιβίωσης. Αυτοί οι παράγοντες θα πρέπει να συμπεριληφθούν στο μοντέλο που θα χρησιμοποιηθεί για τη σχετική ανάλυση των δεδομένων. Αφού η συνάρτηση κινδύνου είναι μη αρνητική, ο λογάριθμός της μπορεί να εκφραστεί ως γραμμική συνάρτηση επεξηγηματικών μεταβλητών:

$$\ln h(t) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$$

Το μοντέλο αυτό όμως είναι πολύ περιοριστικό αφού υποθέτει ότι η συνάρτηση κινδύνου δεν εξαρτάται από το χρόνο. Υπάρχουν διάφορες μέθοδοι με τις οποίες το μοντέλο θα μπορούσε να υιοθετήσει την εξάρτηση του χρόνου, με την πιο γνωστή να είναι το μοντέλο αναλόγων συναρτήσεων κινδύνου (Cox 1972). Το μοντελό αυτό δίνεται από

$$\ln h(t) = \alpha(t) + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$$

όπου $\alpha(t)$ είναι οποιαδήποτε συνάρτηση του χρόνου. Ο όρος ¨αναλόγων συναρτήσεων κινδύνου¨ είναι λόγω του γεγονότος ότι για οποιαδήποτε άτομα για οποιδήποτε σημείο του χρόνου, ο λόγος των συναρτήσεων κινδύνου είναι σταθερός. Εξαιτίας του ότι η συνάρτηση κινδύνου $\alpha(t)$ δεν είναι ανάγκη να οριστεί εξ ολοκλήρου, το μοντέλο αναλόγων συναρτήσεων κινδύνου θεωρείται ως ημιπαραμετρικό.

Ο Cox εισηγήθηκε μια μέθοδο δεσμευμένης πιθανοφάνειας για εκτίμηση των παραμέτρων. Το σημαντικό στοιχείο αυτής της μεθόδου είναι οι εκτιμήσεις εξαρτώνται από τη σειρά με την οποία συμβαίνει το γεγονός και όχι τον ακριβή χρόνο.

19.4 Παράδειγμα

Τα δεδομένα του παραδείγματος, τα οποία δίνονται στο παράρτημα στο τέλος του κεφαλαίου, αναφέρονται σε 51 ασθενείς οι οποίοι πάσχουν από οξεία μυελοβλαστική λευχαιμία και που δεν έχουν μέχρι τώρα δεχθεί οποιαδήποτε θεραπεία. Οι ασθενείς αυτοί υποβάλλονται σε μια σειρά θεραπείας, στο τέλος της οποίας έχουν εξετασθεί αν έχουν ανταποκριθεί ή όχι. Έχουν καταγραφεί πριν τη θεραπεία έξι μεταβλητές :

1. η ηλικία διάγνωσης, `Age`,
2. το ποσοστό επίστρωσης των βλαστοκυττάρων, `Smear`,
3. το ποσοστό των κυττάρων από τη λευχαιμία που εισήλθαν στο μυελό των οστών, `Infil`,
4. το ποσοστό των κυττάρων που προήλθαν από το μυελό των οστών, `Index`,
5. τα απόλυτα βλαστοκύτταρα, `Blasts`, και
6. η ψηλότερη θερμοκρασία σώματος πριν τη θεραπεία, `Temp`.

Επίσης, καταγράφεται ο χρόνος επιβίωσης του ατόμου, `Time`, και η ανταπόκρισή του στη θεραπεία, `Resp`. Τέλος, η μεταβλητή `Status` δείχνει αν οι παρατηρήσεις ενός ατόμου είναι λογοκριμένες ή όχι. Η προς εξέταση ερώτηση είναι το κατά πόσο σημαντικές είναι στη πρόβλεψη του χρόνου επιβίωσης οι έξι μεταβλητές που καταγράφηκαν πριν από τη θεραπεία. Για να απαντηθεί η ερώτηση αυτή θα χρησιμοποιηθεί το μοντέλο αναλόγων συναρτήσεων κινδύνου.

Στην R οι αναγκαίες συναρτήσεις για την ανάλυση επιβίωσης βρίσκονται στη βιβλιοθήκη `survival`. Το μοντέλο αναλόγων συναρτήσεων κινδύνου εφαρμόζεται χρησιμοποιώντας την εντολή `coxph`. Η περιγραφή του μοντέλου στην `coxph`

γίνεται με παρόμοιο τρόπος όπως και στην περίπτωση των γραμμικών μοντέλων με την lm, με τη διαφορά ότι το αριστερό μέρος της είναι αντικείμενο επιβίωσης που δημιουργείται από τη συνάρτηση Surv. Στην περίπτωση που τα δεδομένα είναι δεξιά-λογοκριμένα, η συνάρτηση Surv έχει τη μορφή Surv(time,event), με time να είναι είτε ο χρόνος μέχρι το γεγονός ή ο χρόνος λογοκρισίας, και event να είναι μια δείκτρια μεταβλητή με τιμή ίση με 1 αν το γεγονός παρατηρείται ή ίση με 0 αν αν η παρατήρηση είναι λογοκριμένη. Εφαρμόζοντας το μοντέλο αναλόγων συναρτήσεων κινδύνου με τις 6 επεξηγηματικές μεταβλητές παίρνονται τα ακόλουθα αποτελέσματα:

```
> library(survival)
> cancer.dat <- read.table("cancer.dat", col.names=c("Age","Smear",
+ "Infil","Index","Blasts","Temp","Resp","Time","Status"))
> time<-cancer.dat[,"Time"]
> status<-1-cancer.dat[,"Status"]
> attach(cancer.dat)
> cancer.cox<-coxph(Surv(time,status)~Age+Smear+Infil+Index+Blasts)
> summary(cancer.cox)
Call:
coxph(formula = Surv(time, status) ~ Age + Smear + Infil + Index +
    Blasts)

  n= 51
          coef exp(coef) se(coef)      z       p
Age     0.03536     1.036  0.01018  3.473 0.00052
Smear   0.00915     1.009  0.01451  0.631 0.53000
Infil  -0.01835     0.982  0.01247 -1.472 0.14000
Index  -0.08955     0.914  0.04482 -1.998 0.04600
Blasts  0.00285     1.003  0.00973  0.293 0.77000

       exp(coef) exp(-coef) lower .95 upper .95
Age        1.036      0.965     1.016     1.057
Smear      1.009      0.991     0.981     1.038
Infil      0.982      1.019     0.958     1.006
Index      0.914      1.094     0.837     0.998
Blasts     1.003      0.997     0.984     1.022
```

```
Rsquare= 0.312   (max possible= 0.996 )
Likelihood ratio test= 19.1  on 5 df,   p=0.00188
Wald test            = 17.6  on 5 df,   p=0.00351
Score (logrank) test = 18.9  on 5 df,   p=0.00197
```

Είναι φανερό ότι η ηλικία διάγνωσης, Age, είναι η πιο σημαντική μεταβλητή για την πρόβλεψη του χρόνου επιβίωσης, αφού έχει p-value πολύ κοντά στο 0. Για το λόγο αυτό εφαρμόζεται ένα νέο μοντέλο με μόνο αυτή τη μεταβλητή ως επεξηγηματική και δίνει τα ακόλουθα αποτελέσματα:

```
> cancer.cox1<-coxph(Surv(time,status)~Age)
> cancer.cox1
Call:
coxph(formula = Surv(time, status) ~ Age)


      coef exp(coef) se(coef)   z       p
Age 0.0324      1.03  0.00952 3.4 0.00067

Likelihood ratio test=11.8  on 1 df, p=0.000577  n= 51
```

Η ερμηνεία του εκτιμημένου συντελεστή του μοντέλου είναι ότι κάθε επιπρόσθετος χρόνος ζωής αυξάνει το λογάριθμο της συνάρτησης κινδύνου κατά 0.0324. Μια πιο σωστή προσέγγιση της ερμηνείας μπορεί να γίνει αφού πρώτα βρεθεί η εκθετική συνάρτηση του συντελεστή. Έπειτα για κάθε αύξηση κατά μία μονάδα της επεξηγηματική μεταβλητής, η συνάρτηση κινδύνου πολλαπλασιάζεται με τον εκθετικό συντελεστή. Η τιμή της

$$100(\exp(coefficient) - 1)$$

δίνει την ποσοστιαία αλλαγή στη συνάρτηση κινδύνου για κάθε μοναδιαία αύξηση στην επεξηγηματική μεταβλητή. Συνεπώς, αύξηση ενό χρόνου στην ηλικία διάγνωσης οδηγεί σε αύξηση 3% της συνάρτησης κινδύνου.

Το επόμενο στάδιο της ανάλυσης η ανάλυση των υπολοίπων από το μοντέλο με τη βοήθεια διαγνωστικών γραφημάτων για την εξεύρεση απομακρυσμένων τιμών, παρατηρήσεων επιρροής κ.α. Θα χρησιμοποιηθούν τρία διαφορετικά είδη υπολοίπων:

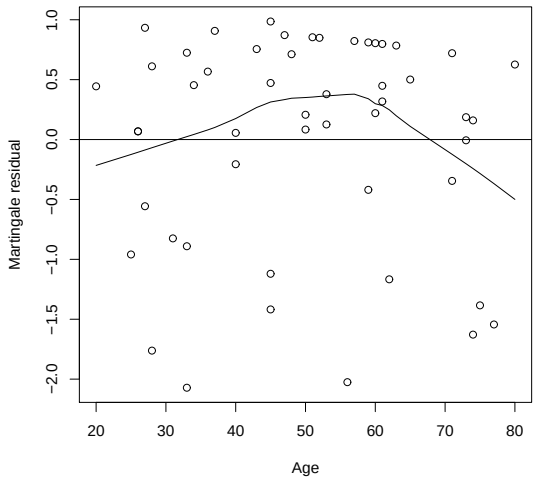
- martingale: χρήσιμα στο να αποκαλύπτουν τη συναρτησιακή μορφή των επεξηγηματικών μεταβλητών,

- deviance: χρήσιμα στην αναγνώριση των λιγότερο καλών εκτιμώμενων παρατηρήσεων,
- Schoenfeld: χρήσιμα στο να υποδεικνύουν αν το μοντέλο αναλόγων συναρτήσεων κινδύνου είναι κατάλληλο ή όχι.

Στην αρχή κατασκευάζεται το γράφημα των υπολοίπων martingale συναρτήσει της ηλικίας διάγνωσης χρησιμοποιώντας τις πιο κάτω εντολές,

```
> cancer.cox1.mart<-residuals(cancer.cox1,type="martingale")
> plot(Age,cancer.cox1.mart,ylab="Martingale residual")
> abline(h=0)
> lines(lowess(Age,cancer.cox1.mart))
```

Το γράφημα παρουσιάζεται στο Σχήμα 19.1. Η τελευταία εντολή δίνει τη γραμμή εξομάλυνσης lowess, η οποία υποδεικνύει ότι ίσως να πρέπει να θεωρηθεί ο τετραγωνικός παράγοντας της ηλικίας διάγνωσης στο μοντέλο, ή εναλλακτικά, να θεωρηθούν ξεχωριστά οι ηλικίες μικρότερες και μεγαλύτερες του 45.

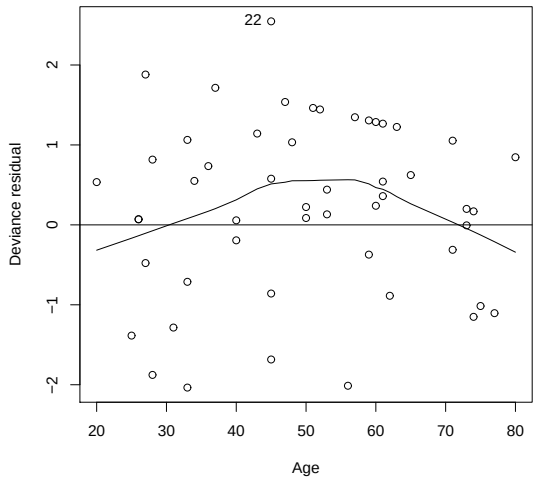


Σχήμα 19.1: Ηλικία διάγνωσης συναρτήσει των υπολοίπων martingale μαζί με τη γραμμή εξομάλυνσης lowess.

Στη συνέχεια, κατασκευάζεται με παρόμοιο τρόπο το γράφημα των υπολοίπων deviance συναρτήσει της ηλικίας διάγνωσης και παρουσιάζεται στο Σχήμα 19.2.

Μόνο το υπόλοιπο το οποίο αντιστοιχεί στην 22η παρατήρηση παρουσιάζεται ως απομακρυσμένη τιμή. Αυτό αντιστοιχεί σε άτομο με μηδενικό χρόνο επιβίωσης.

```
> cancer.cox1.dev<-residuals(cancer.cox1,type="deviance")
> plot(Age,cancer.cox1.dev,ylab="Deviance residual")
> abline(h=0)
> lines(lowess(Age,cancer.cox1.dev))
> identify(Age,cancer.cox1.dev,n=1)
[1] 22
```

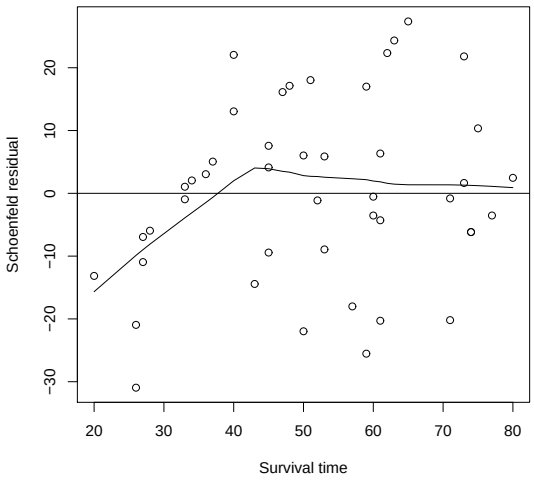


Σχήμα 19.2: Ηλικία διάγνωσης συναρτήσει των υπολοίπων deviance μαζί με τη γραμμή εξομάλυνσης lowess.

Τέλος, δίνεται το γράφημα των υπολοίπων Schoenfeld συναρτήσει του χρόνου επιβίωσης με την προσθήκη της γραμμής εξομάλυνσης lowess (Σχήμα 19.3).

```
> cancer.cox1.shoen<-residuals(cancer.cox1,type="schoenfeld")
> plot(Age[status==1],cancer.cox1.shoen,xlab="Survival time",
+ ylab="Schoenfeld residual")
> abline(h=0)
> lines(lowess(Age[status==1],cancer.cox1.shoen))
```

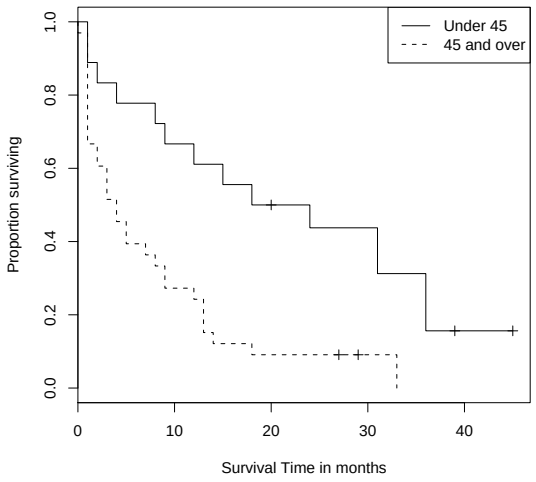
Το πιο σημαντικό συμπέρασμα είναι ότι όσο μικραίνει ο χρόνος επιβίωσης οι τιμές των υπολοίπων τείνουν να είναι αρνητικές. Αυτό αντιστοιχεί στους νεαρότερους ασθενείς και συνεπώς, υπάρχει και εδώ ένδειξη ότι ίσως να είναι καλύτερα να γίνει διαχωρισμός των ασθενών κατά ηλικία.



Σχήμα 19.3: Χρόνος επιβίωσης συναρτήσει των υπολοίπων Schoenfeld μαζί με τη γραμμή εξομάλυνσης lowess.

Λαμβάνοντας υπόψη την ένδειξη από την ανάλυση υπολοίπων, θα γίνει προσπάθεια ανάλυσης των δεδομένων εξετάζοντας τους ασθενείς βάσει της ηλικίας τους, διαχωρίζοντάς τους σε δύο κατηγορίες, μικρότερους και μεγαλύτερους από 45 ετών. Με τη βοήθεια των ακόλουθων εντολών κατασκευάζεται το γράφημα των καμπυλών διαβίωσης των δύο ηλικιακών κατηγοριών στο ίδιο γράφημα (Σχήμα 19.4).

```
> agroup<-cancer.dat[,"Age"]-45
> agroup[agroup>=0]<-1
> agroup[agroup<0]<-0
> plot(survfit(Surv(time,status)~agroup),xlab="Survival Time in months",
+ ylab="Proportion surviving",lty=1:2)
> legend("topright",c("Under 45","45 and over"),lty=1:2)
```



Σχήμα 19.4: Καμπύλη επιβίωσης για τα άτομα με ηλικία διάγνωσης κάτω και πάνω από 45 χρονών.

Όπως διαφαίνεται από το γράφημα, υπάρχει σημαντική διαφορά στις καμπύλες διαβίωσης των δύο κατηγοριών.Αυτό εξετάζεται και με τη βοήθεια του ελέγχου log-rank, ό οποίος είναι έλεγχος X^2 και εφαρμόζεται στην R με την εντολή `survdiff`.

```
> survdiff(Surv(time,status)~agroup)
Call:
survdiff(formula = Surv(time, status) ~ agroup)

          N Observed Expected (O-E)^2/E (O-E)^2/V
agroup=0 18       14     23.8      4.05      11.0
agroup=1 33       31     21.2      4.55      11.0

 Chisq= 11  on 1 degrees of freedom, p= 0.000926
```

Ο έλεγχος δίνει τιμή για την ελεγχοσυνάρτηση ίση με 11 με 1 βαθμό ελευθερίας. Το p-value του ελέγχου είναι πολύ κοντά στο 0 με αποτέλεσμα να απορρίπτεται η υπόθεση ότι οι δύο ηλικιακές κατηγορίες έχουν την ίδια συνάρτηση διαβίωσης.

Παράρτημα

Τα δεδομένα που χρησιμοποιούνται σε αυτό το κεφάλαιο για εφαρμογή της ανάλυσης διαβίωσης.

```
> cancer.dat
   Age Smear Infil Index Blasts Temp Resp Time Status
1   20    78    39     7    0.6  990    1   18      0
2   25    64    61    16   35.0 1030    1   31      1
3   26    61    55    12    7.5  982    1   31      0
4   26    64    64    16   21.0 1000    1   31      0
5   27    95    95     6    7.5  980    1   36      0
6   27    80    64     8    0.6 1010    0    1      0
7   28    88    88    10    4.8  986    1    9      0
8   28    70    70    14   10.0 1010    1   39      1
9   31    72    72     5    2.3  988    1   20      1
10  33    58    58     7    5.7  986    0    4      0
11  33    92    92     5    2.6  980    1   45      1
12  33    42    38    12    2.5  984    1   36      0
13  34    26    26     7    7.0  982    0   12      0
14  36    55    55    14    4.5  986    1    8      0
15  37    71    71    15    4.4 1020    0    1      0
16  40    91    91     9   35.0  986    1   15      0
17  40    52    49    12    2.1  988    1   24      0
18  43    74    63     4    0.1  986    0    2      0
19  45    78    47    14    4.2  980    1   33      0
20  45    60    36    10    0.6  992    1   29      1
21  45    82    32    10   28.1 1016    0    7      0
22  45    79    79     4    1.1 1030    0    0      0
23  47    56    28     2    0.9  990    0    1      0
24  48    60    54    10    2.2 1002    0    2      0
25  50    83    66    19   11.6  996    1   12      0
26  50    36    32    14    4.5  992    1    9      0
27  51    88    70     8    0.5  982    0    1      0
28  52    87    87     7   10.3  986    0    1      0
29  53    75    68    13    2.3  980    1    9      0
30  53    65    65     6    2.3  982    0    5      0
```

31	56	97	92	10	16.0	992	1	27	1
32	57	87	83	19	21.6	1020	0	1	0
33	59	45	45	8	1.1	999	0	13	0
34	59	36	34	5	0.0	1038	0	1	0
35	60	39	33	7	0.9	988	0	5	0
36	60	76	53	12	0.4	982	0	1	0
37	61	46	37	4	1.4	1006	0	3	0
38	61	39	8	8	0.3	990	0	4	0
39	61	90	90	1	9.9	990	0	1	0
40	62	84	84	19	115.0	1020	1	18	0
41	63	42	27	5	0.3	1014	0	1	0
42	65	75	75	10	20.0	1004	0	2	0
43	71	44	22	6	0.3	990	0	1	0
44	71	63	63	11	10.0	986	1	8	0
45	73	33	33	4	0.5	1010	0	3	0
46	73	93	84	6	38.0	1020	0	4	0
47	74	58	58	10	2.4	1002	1	14	0
48	74	32	30	16	6.7	988	0	3	0
49	75	60	60	17	8.2	990	1	13	0
50	77	69	69	9	1.5	986	1	13	0
51	80	73	73	7	1.5	986	0	1	0