ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΤΟΜΕΑΣ ΜΑΘΗΜΑΤΙΚΩΝ

Θέμα Εργασίας: Ανάλυση Παλινδρόμησης

Μάθημα: Ανάλυση Δεδομένων με Η/Υ

Διδάσκων: Δημήτρης Φουσκάκης

Φοιτήτρια: Ελένη Στυλιανού, ge21708

Email: elenistylianou03@live.com

Άσκηση 1:

Αρχικά, φορτώνω τα δεδομένα μου στην R,δημιουργώντας ένα dataframe που το ονομάζω data, χρησιμοποιώντας την εντολή

>data<na.omit(read.table("http://www.math.ntua.gr/~fouskakis/Data_Analysis/Exercises/pharmacy.txt", header=T, na.strings="$"))

Το τελευταίο όρισμα δηλώνει στην R ότι οι αγνοούμενες τιμές του dataframe είχαν συμβολιστεί με $ στο αρχείο των δεδομένων. Η R θα μετατρέψει το $ σε NA που είναι το δικό της σύμβολο για τις αγνοούμενες τιμές.

Με την εντολή:

>nrow(data)

Επιστράφηκε ο αριθμός των γραμμών του νέου πλαισίου δεδομένων data, δηλαδή ο αριθμός των πελατών που έκαναν μία μόνο αγορά των προηγούμενο μήνα για κάποιο μη φαρμακευτικό προϊόν. Παρατηρούμε ότι αφαιρέθηκε 1 πελάτης από τους 72 του αρχικού δείγματος.

i. Για να εκτιμήσουμε σημειακά τους συντελεστές συσχέτισης όλων των δυνατών συνδυασμών των μεταβλητών age, med, population και money και για να ελέγξουμε αν οι συντελεστές συσχέτισης είναι στατιστικά σημαντικοί για κάθε συνδυασμό δίνουμε τις εξής εντολές:

```
> cor.test(data$age,data$med)
        Pearson's product-moment correlation
data:  data$age and data$med
t = -1.4286, df = 69, p-value = 0.1576
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.38747304  0.06643985
sample estimates:
      cor
-0.1694911
> cor.test(data$age,data$population)
        Pearson's product-moment correlation
data:  data$age and data$population
t = 0.14302, df = 69, p-value = 0.8867
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.2169599  0.2495168
sample estimates:
      cor
0.01721519
> cor.test(data$age,data$money)
        Pearson's product-moment correlation
data:  data$age and data$money
t = 2.7499, df = 69, p-value = 0.007606
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.08737517 0.51017006
sample estimates:
     cor
0.3142725
```

```
> cor.test(data$med,data$population)
      Pearson's product-moment correlation
data:  data$med and data$population
t = -0.71177, df = 69, p-value = 0.479
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.3124542  0.1509361
sample estimates:
      cor
-0.08537395
> cor.test(data$med,data$money)
      Pearson's product-moment correlation
data:  data$med and data$money
t = 7.0581, df = 69, p-value = 1.044e-09
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.4879201 0.7652188
sample estimates:
    cor
0.6475149
> cor.test(data$population,data$money)
      Pearson's product-moment correlation
data:  data$population and data$money
t = 0.16581, df = 69, p-value = 0.8688
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.2143444  0.2520873
sample estimates:
    cor
0.0199573
```

Παρατηρούμε ότι οι συντελεστές συσχέτισης(κόκκινο χρώμα) των συνδυασμών age-money και med-money είναι στατιστικά σημαντικοί αφού η P-τιμή στους ελέγχους είναι 0.007606 και 1.044e-09 αντίστοιχα και είναι μικρότερες από 0.05 που είναι το επίπεδο σημαντικότητας. Οι υπόλοιποι συνδυασμοί δεν έχουν στατιστικά σημαντικούς συντελεστές συσχέτισης, εφόσον η P-τιμή στους ελέγχους είναι μεγαλύτερη από 0.05.

Παρατηρούμε πως ο συντελεστής συσχέτισης(κόκκινο χρώμα) των μεταβλητών age και med είναι αρνητικός και πολύ κοντά στο 0, άρα ο βαθμός γραμμικής συσχέτισης των δύο χαρακτηριστικών είναι πάρα πολύ μικρός, επομένως μπορούμε να πούμε ότι οι μεταβλητές είναι ασυσχέτιστες. Επίσης, ο συντελεστής συσχέτισης των μεταβλητών age και population είναι θετικός αλλά και αυτός πολύ κοντά στο 0, άρα η δύο μεταβλητές είναι ασυσχέτιστες. Όμοια και στους συνδυασμούς population-money και med-population που είναι και σε αυτούς ασυσχέτιστες οι μεταβλητές που τους αποτελούν, εφόσον ο συντελεστής συσχέτισης είναι πολύ κοντά στο 0. Για τις μεταβλητές age και money παρατηρούμε ότι ο συντελεστής συσχέτισης είναι λίγο μεγαλύτερος από τους προηγούμενους που μελετήθηκαν, αλλά είναι και αυτός πολύ κοντά στο 0, επομένως μπορούμε να αποφανθούμε ότι και αυτές οι μεταβλητές είναι ασυσχέτιστες. Τέλος, παρατηρούμε ότι ο συντελεστής συσχέτισης των μεταβλητών med και money είναι θετικός(δηλαδή αύξηση του ποσού των χρημάτων σε ευρώ που ξόδεψε στο φαρμακείο σε φαρμακευτικά προϊόντα τον προηγούμενο μήνα κατά μέσο όρο αυξάνει το ποσό των χρημάτων

σε ευρώ που ξόδεψε στην αγορά του μη φαρμακευτικού προϊόντος) και αρκετά κοντά στη μονάδα, άρα υπάρχει μία γραμμική συσχέτιση μεταξύ των δύο αυτών μεταβλητών σύμφωνα με το δείγμα.

ii. Μέσω του παρακάτω κώδικα θα προσαρμόσουμε το απλό γραμμικό μοντέλο, θα δημιουργήσουμε δηλαδή ένα μοντέλο της μορφής $Y=\alpha+bx+\varepsilon$ ($\varepsilon \sim N(0,\sigma^2)$), με μεταβλητή απόκρισης την τ.μ. money και επεξηγηματική μεταβλητή την τ.μ. med, με σκοπό να μελετήσουμε την επίδραση του ποσού των χρημάτων σε ευρώ που ξόδεψε στο φαρμακείο σε φαρμακευτικά προϊόντα τον προηγούμενο μήνα στο ποσό των χρημάτων σε ευρώ που ξόδεψε στην αγορά του μη φαρμακευτικού προϊόντος.

```
> attach(data)
> results<-lm(money~med)
> results
Call:
lm(formula = money ~ med)
Coefficients:
(Intercept)        med
   19.3225      0.4256
```

Από αυτό βλέπουμε πως η εκτιμώμενη ευθεία της απλής γραμμικής παλινδρόμησης είναι η Y=0,4256x+19,3225, όπου Y είναι το ποσό των χρημάτων σε ευρώ που ξόδεψε στην αγορά του μη φαρμακευτικού προϊόντος και x το ποσό των χρημάτων σε ευρώ που ξόδεψε στο φαρμακείο σε φαρμακευτικά προϊόντα τον προηγούμενο μήνα.

Από την εντολή:

```
> summary(results)
Call:
lm(formula = money ~ med)

Residuals:
   Min     1Q  Median     3Q    Max
-14.041  -3.306   0.273   3.173  11.790

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  19.3225     4.6425   4.162 8.95e-05 ***
med           0.4256     0.0603   7.058 1.04e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.606 on 69 degrees of freedom
Multiple R-squared:  0.4193,    Adjusted R-squared:  0.4109
F-statistic: 49.82 on 1 and 69 DF,  p-value: 1.044e-09
```

Παρατηρούμε ότι η P-τιμή=1.04e-09 *** στον έλεγχο για το b είναι πολύ μικρή. Αυτό σημαίνει πως όταν το ποσό των χρημάτων σε ευρώ που ξόδεψε στο φαρμακείο σε φαρμακευτικά προϊόντα τον προηγούμενο μήνα αυξηθεί κατά ένα ευρώ η αναμενόμενη αύξηση του ποσού των χρημάτων σε ευρώ που ξόδεψε στην αγορά του μη φαρμακευτικού προϊόντος είναι 0,43 ευρώ. Επίσης, παρατηρούμε ότι η P-τιμή= 8.95e-05 *** στον έλεγχο για το α είναι πολύ μικρή, οπότε όταν το ποσό των χρημάτων σε ευρώ που ξόδεψε στο φαρμακείο σε φαρμακευτικά προϊόντα

τον προηγούμενο μήνα είναι 0 ευρώ τότε το ποσό των χρημάτων σε ευρώ που ξόδεψε στην αγορά του μη φαρμακευτικού προϊόντος είναι 19,32 ευρώ.
Επιπρόσθετα παρατηρούμε ότι ο συντελεστής προσδιορισμού R^2 είναι 0.4193 και ο διορθωμένος συντελεστής προσδιορισμού είναι 0.4109. Ο συντελεστής προσδιορισμού εκφράζει το ποσοστό της διασποράς της τυχαίας μεταβλητής Y (money) που εξηγείται με βάση το πιο πάνω μοντέλο παλινδρόμησης. Όσο μεγαλύτερος είναι ο συντελεστής προσδιορισμού τόσο ισχυρότερη είναι η γραμμική σχέση εξάρτησης μεταξύ των τ.μ. Y(money) και X(med). Ο R^2 υποδηλώνει ότι το 41,09% της διακύμανσης στο ποσό των χρημάτων σε ευρώ που ξόδεψε στην αγορά του μη φαρμακευτικού προϊόντος εξηγείται από το ποσό των χρημάτων σε ευρώ που ξόδεψε στο φαρμακείο σε φαρμακευτικά προϊόντα τον προηγούμενο μήνα. Επίσης, παρατηρούμε πως ο συντελεστής προσδιορισμού είναι μεγαλύτερος από τον διορθωμένο συντελεστή προσδιορισμού και οι δύο είναι μικρότερη της μονάδας.

Με την εντολή:

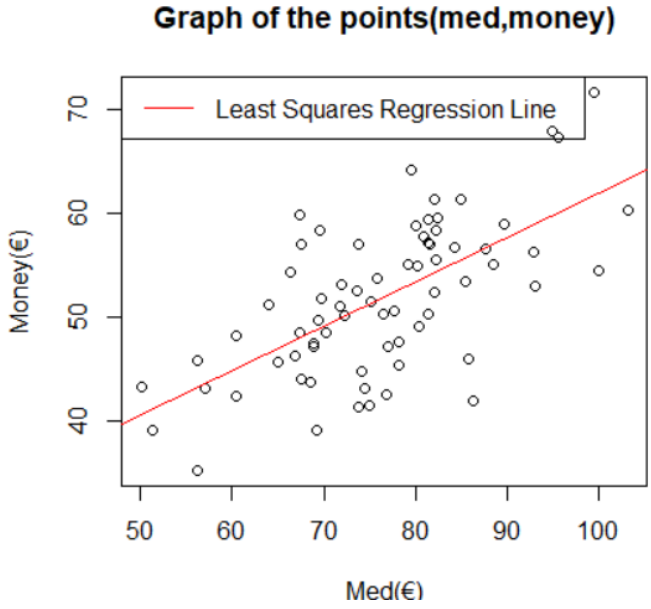
```
> confint(results)
                2.5 %    97.5 %
(Intercept) 10.0609511 28.5839838
med          0.3053246  0.5459268
```

βρίσκουμε τα εκτιμώμενα διαστήματα εμπιστοσύνης για τα εκτιμώμενα α και b. Τα συμμετρικά 95% διαστήματα εμπιστοσύνης για το α είναι (10.0610,28.5840) και για το b είναι (0.3053,0.5459).

Για το γράφημα διασποράς των παρατηρήσεων μαζί με την ευθεία ελαχίστων τετραγώνων δίνουμε τις εξής εντολές:

```
> plot(med,money,xlab='Med(€)',ylab='Money(€)',main='Graph of the points(med,money)')
> abline(results,col='red')
> legend('topleft',col='red',lty=1,legend='Least Squares Regression Line')
```

Και παίρνουμε το παρακάτω γράφημα:



Η θετική κλίση της ευθείας δείχνει ότι υπάρχει θετική συσχέτιση μεταξύ των ποσών που ξοδεύονται σε φαρμακευτικά και μη προϊόντα.

iii. Οι προϋποθέσεις που απαιτεί το απλό γραμμικό μοντέλο είναι οι εξής

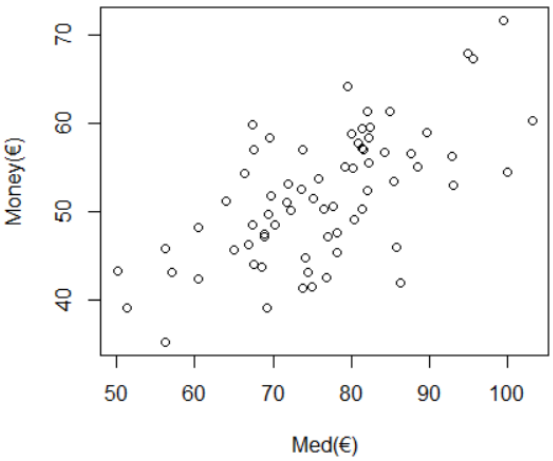
- Γραμμικότητα μεταξύ των μεταβλητών money και med
- Κανονικότητα σφαλμάτων
- Ομοσκεδαστικότητα

- Ανεξαρτησία σφαλμάτων

Γραμμικότητα

Απαιτείται οι μεταβλητές money και med να συνδέονται με κάποια γραμμική σχέση. Για να εξετάσουμε την γραμμικότητα δημιουργούμε ένα διάγραμμα διασποράς, μια απεικόνιση δηλαδή των σημείων (xi,yi), όπου x η τιμές της μεταβλητής med και y οι τιμές τις μεταβλητής money.

```
> x<-data$med
> y<-data$money
> plot(x,y,xlab='Med(€)',ylab='Money(€)')
```

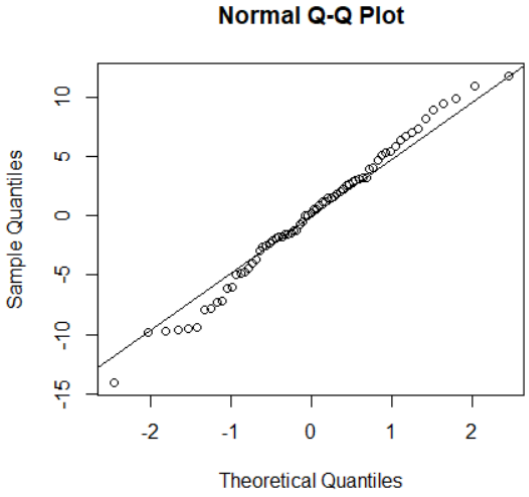


Από το γράφημα που παίρνουμε παρατηρούμε πως η υπόθεση της γραμμικότητας είναι λογική.

Κανονικότητα Σφαλμάτων

Απαιτείται τα σφάλματα να ακολουθούν κανονική κατανομή με μέση τιμή 0. Για τον έλεγχο της κανονικότητας των σφαλμάτων θα δημιουργήσουμε qqplots για τα υπόλοιπα.

```
> qqnorm(results$res)
> qqline(results$res)
```

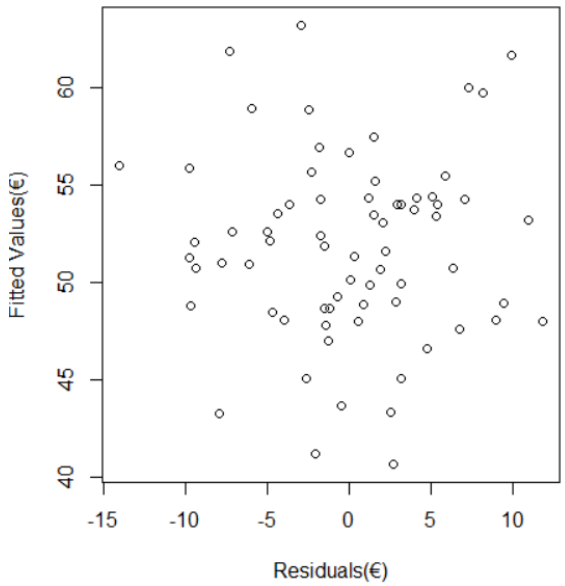


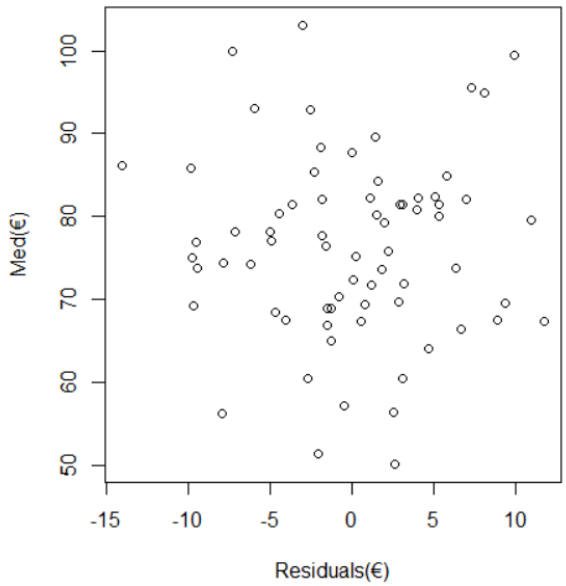
Από το διπλανό σχήμα μπορούμε εύκολα να συμπεράνουμε πως ισχύει και η προϋπόθεση της κανονικότητας σφαλμάτων, αφού παρατηρούμε πως τα υπόλοιπα ακολουθούν μια κανονική κατανομή.

Ομοσκεδαστικότητα

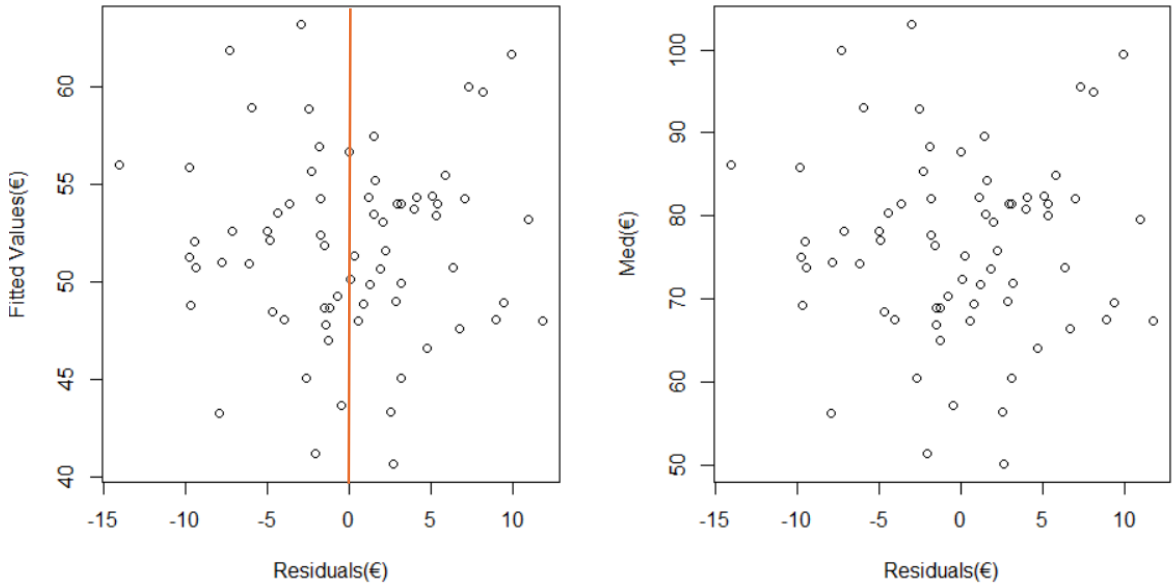
Απαιτείται η διασπορά των τυχαίων σφαλμάτων να παραμένει σταθερή για τις διάφορες τιμές της μεταβλητής med του τυχαίου διανύσματος. Για το έλεγχο της ομοσκεδαστικότητας θα δημιουργήσουμε την γραφική παράσταση των υπολοίπων συναρτήσει των προβλεπόμενων τιμών ή συναρτήσει των τιμών της μεταβλητής med. Πιο κάτω δίνονται και τα δύο διαγράμματα με τις εξής εντολές:

```
> par(mfrow=c(1,2))
> plot(results$res,results$fitted,xlab='Residuals(€)',ylab='Fitted Values(€)')
> plot(results$res,x,xlab='Residuals(€)',ylab='Med(€)')
```





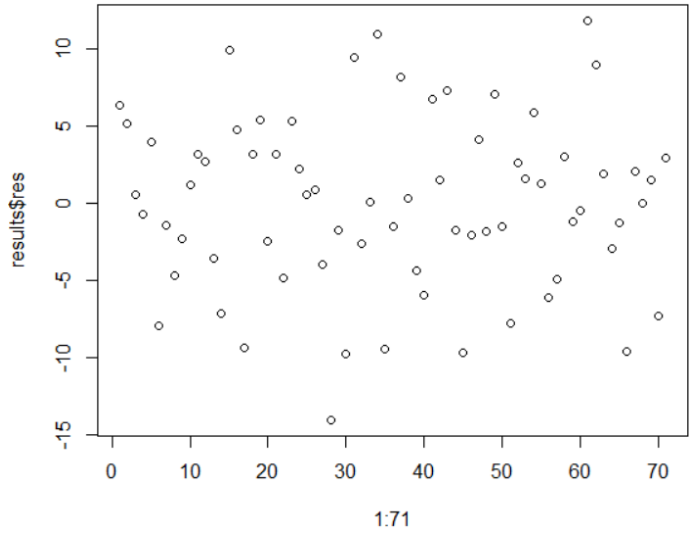
Από τα πιο πάνω γραφήματα δεν παρατηρούμε κάποιο συστηματικό τρόπο συμπεριφοράς των δεδομένων, επομένως ισχύει η υπόθεση της ομοσκεδαστικότητας.
Επίσης, αν φέρουμε μία κάθετη γραμμή στο 0 σε ένα από τα γραφήματα παρατηρούμε πως δεξιά και αριστερά τις γραμμής έχουμε περίπου ίσο αριθμό σημείων, γεγονός που μας δείχνει πως ισχύει η γραμμικότητα.



Ανεξαρτησία Σφαλμάτων

Απαιτείται τα σφάλματα να είναι ανεξάρτητα μεταξύ τους. Για να ελέγξουμε την ανεξαρτησία των σφαλμάτων κατασκευάζουμε ένα διάγραμμα υπολοίπων σε σχέση με τη σειρά των δεδομένων.

>plot(1:71,results$res)



Από το παραπάνω γράφημα παρατηρούμε πως δεν παρουσιάζεται κάποια σχέση και τα υπόλοιπα συμπεριφέρονται τυχαία. Επομένως, ισχύει και η προϋπόθεση της ανεξαρτησίας σφαλμάτων.

iv. Προσαρμόζουμε το πολλαπλό γραμμικό μοντέλο με μεταβλητή απόκρισης την τ.μ. money και επεξηγηματικές μεταβλητές τις age, category, sex και med μέσω του πιο κάτω κώδικα:

```
> results2<-lm(money~age+category+sex+med+population)
> results2
Call:
lm(formula = money ~ age + category + sex + med + population)
```

Coefficients:

(Intercept)	age	categoryhealthcare	categoryother
7.9553472	0.2902512	-4.3897690	-1.7055805
sexWoman	med	population	
1.9246377	0.4475493	0.0002518	

Το πολλαπλό γραμμικό μοντέλο είναι της μορφής $Y=\alpha+b_1x_1+b_2x_2+b_3x_3+b_4x_4+ b_5x_5+ b_6x_6+\varepsilon$ ($\varepsilon \sim N(0,\sigma^2)$), όπου Y είναι το ποσό χρημάτων σε ευρώ που ξόδεψε στην αγορά του μη φαρμακευτικού προϊόντος, x_1 είναι η ηλικία σε έτη, x_2 η κατηγορία healthcare, x_3 η κατηγορία other, x_4 το φύλο, x_5 το ποσό χρημάτων σε ευρώ που ξόδεψε στο φαρμακείο για φαρμακευτικά προϊόντα τον προηγούμενο μήνα και x_6 ο αριθμός των κατοίκων που έχει η περιοχή που βρίσκεται το κατάστημα που έγινε η αγορά.

Οι συντελεστές του πιο πάνω γραμμικού μοντέλου ερμηνεύονται ως εξής:

- α=7.96: Εκφράζει το ποσό χρημάτων σε ευρώ που ξόδεψε στην αγορά του μη φαρμακευτικού προϊόντος στην περίπτωση που η ηλικία είναι 0, η κατηγορία του μη φαρμακευτικού προϊόντος αγοράς είναι cosmetics, το φύλο είναι άνδρας, το ποσό χρημάτων σε ευρώ που ξόδεψε στο φαρμακείο για φαρμακευτικά προϊόντα τον προηγούμενο μήνα και ο πληθυσμός είναι 0.
- b_1=0.29: Το ποσό χρημάτων σε ευρώ που ξόδεψε ένα άτομο στην αγορά του μη φαρμακευτικού προϊόντος ανεξαρτήτως κατηγορίας, φύλου, πληθυσμού και του ποσού χρημάτων για φαρμακευτικά προϊόντα, αυξάνεται κατά 0.29 ευρώ αν η ηλικία του ατόμου αυξηθεί κατά 1 έτος.
- b_2=-4.39: Το ποσό χρημάτων σε ευρώ που ξόδεψε ένα άτομο στην αγορά του μη φαρμακευτικού προϊόντος που ανήκει στην κατηγορία healthcare είναι 4.39 ευρώ λιγότερο από το ποσό χρημάτων που ξόδεψε σε προϊόν της κατηγορίας cosmetics ένα άτομο, ίδιας ηλικίας, ιδίου φύλου, που ξόδεψε ίσο ποσό σε φαρμακευτικά προϊόντα και σε κατάστημα σε περιοχή με ίδιο πληθυσμό.
- b_3=-1.71: Το ποσό χρημάτων σε ευρώ που ξόδεψε ένα άτομο στην αγορά του μη φαρμακευτικού προϊόντος που ανήκει στην κατηγορία other είναι 1.71 ευρώ λιγότερο από το ποσό χρημάτων που ξόδεψε σε προϊόν της κατηγορίας cosmetics ένα άτομο, ίδιας ηλικίας, ιδίου φύλου, που ξόδεψε ίσο ποσό σε φαρμακευτικά προϊόντα και σε κατάστημα σε περιοχή με ίδιο πληθυσμό.
- b_4=1.92: Το ποσό χρημάτων σε ευρώ που ξόδεψε μια γυναίκα στην αγορά του μη φαρμακευτικού προϊόντος είναι 1.92 ευρώ μεγαλύτερο από αυτό που ξόδεψε ένας άνδρας, ανεξαρτήτως ηλικίας, κατηγορίας, πληθυσμού και ποσού που ξόδεψε σε φαρμακευτικά προϊόντα.
- b_5=0.45: Το ποσό χρημάτων σε ευρώ που ξόδεψε ένα άτομο στην αγορά του μη φαρμακευτικού προϊόντος, ανεξαρτήτως ηλικίας, κατηγορίας, πληθυσμού και φύλου, αυξάνεται κατά 0.45 ευρώ αν το ποσό που ξόδεψε σε φαρμακευτικά προϊόντα αυξηθεί κατά 1 ευρώ.
- b_6=0.00: Το ποσό χρημάτων σε ευρώ που ξόδεψε ένα άτομο στην αγορά του μη φαρμακευτικού προϊόντος, ανεξαρτήτως ηλικίας, κατηγορίας, φύλου και ποσού που ξοδεύτηκε σε φαρμακευτικά προϊόντα, δεν αλλάζει σε επίπεδο σημαντικότητα 2 δεκαδικών ψηφίων αν ο αριθμός των κατοίκων της περιοχής που βρίσκεται το φαρμακείο αυξηθεί κατά 1.

Οι πιο πάνω τιμές αποτελούν εκτιμήσεις των συντελεστών του πολλαπλού γραμμικού μοντέλου.

Με τον πιο κάτω κώδικα παίρνουμε 95% διαστήματα εμπιστοσύνης για τις πραγματικές τιμές των συντελεστών, δηλαδή τα επόμενα διαστήματα περιέχουν, με πιθανότητα 95% τους συντελεστές του πολλαπλού γραμμικού μοντέλου.

```
> confint(results2)
                                2.5 %          97.5 %
(Intercept)            -1.365942164 17.27663663
age                     0.186635726  0.39386663
categoryhealthcare -7.173697545 -1.60584055
categoryother          -4.272342781  0.86118172
sexWoman               -0.285631012  4.13490637
med                     0.347232135  0.54786640
population             -0.001206825  0.00171052
```

Με την εντολή:

```
> summary(results2)

Call:
lm(formula = money ~ age + category + sex + med + population)

Residuals:
   Min     1Q  Median     3Q    Max
-9.5399 -3.2746  0.1195  2.5428  9.2008

Coefficients:
                       Estimate Std.        Error  t value    Pr(>|t|)
(Intercept)             7.9553472  4.6659413   1.705  0.09305 .
age                     0.2902512  0.0518666   5.596 4.91e-07 ***
categoryhealthcare -4.3897690  1.3935462  -3.150  0.00248 **
categoryother          -1.7055805  1.2848396  -1.327  0.18907
sexWoman                1.9246377  1.1063903   1.740  0.08674 .
med                     0.4475493  0.0502156   8.913 8.04e-13 ***
population              0.0002518  0.0007302   0.345  0.73129
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.357 on 64 degrees of freedom
Multiple R-squared:  0.6746,    Adjusted R-squared:  0.6441
```

F-statistic: 22.11 on 6 and 64 DF, p-value: 6.561e-14

θα μελετήσουμε τις P-τιμές(αυτές που είναι σημειωμένες με κόκκινο χρώμα πιο πάνω) της κάθε επεξηγηματικής μεταβλητής για να καταλήξουμε αν είναι στατιστικά σημαντικές ή όχι σε επίπεδο σημαντικότητας 5%.

Παρατηρούμε πως, οι μεταβλητές age,categoryhealthcare και med είναι στατιστικά σημαντικές, εφόσον οι P-τιμές τους είναι 4.91e-07 ***, 0.00248 ** , 8.04e-13 *** αντίστοιχα και είναι μικρότερες του 0.05 που είναι το επίπεδο σημαντικότητας. Οι υπόλοιπες επεξηγηματικές μεταβλητές δεν είναι στατιστικά σημαντικές εφόσον η P-τιμές τους είναι μεγαλύτερες από 0.05.
Για το συγκεκριμένο μοντέλο ο συντελεστής προσδιορισμού R^2 είναι 0.6746 και ο διορθωμένος συντελεστής προσδιορισμού είναι 0.6441. Ο συντελεστής προσδιορισμού εκφράζει το ποσοστό της διασποράς της τυχαίας μεταβλητής Y (money) που εξηγείται με βάση το πιο πάνω μοντέλο παλινδρόμησης. Ο R^2 υποδηλώνει ότι το 67.46% της διακύμανσης στο ποσό των χρημάτων σε ευρώ που ξόδεψε στην αγορά του μη φαρμακευτικού προϊόντος εξηγείται από το φύλο την κατηγορία του προϊόντος, την ηλικία του αγοραστή, τον πληθυσμό στην περιοχή που βρίσκεται το φαρμακείο και το ποσό των χρημάτων σε ευρώ που ξόδεψε στο φαρμακείο σε φαρμακευτικά προϊόντα τον προηγούμενο μήνα. Στο διορθωμένο συντελεστή προσδιορισμού λαμβάνεται υπόψη και το μέγεθος του δείγματος(71) και το πλήθος των επεξηγηματικών μεταβλητών(5) Επίσης, παρατηρούμε πως ο συντελεστής προσδιορισμού είναι μεγαλύτερος από τον διορθωμένο συντελεστή προσδιορισμού και οι δύο είναι μικρότερη της μονάδας.

v. Δίνοντας την επόμενη εντολή:

```
> summary(results2)
Call:

lm(formula = money ~ age + category + sex + med + population)

Residuals:

   Min     1Q  Median     3Q    Max

-9.5399 -3.2746  0.1195  2.5428  9.2008

Coefficients:
                        Estimate Std.        Error  t value   Pr(>|t|)
(Intercept)              7.9553472  4.6659413   1.705  0.09305 .

age                      0.2902512  0.0518666   5.596 4.91e-07 ***

categoryhealthcare -4.3897690  1.3935462  -3.150  0.00248 **

categoryother         -1.7055805  1.2848396  -1.327  0.18907

sexWoman                 1.9246377  1.1063903   1.740  0.08674 .

med                      0.4475493  0.0502156   8.913 8.04e-13 ***

population               0.0002518  0.0007302   0.345  0.73129

---

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.357 on 64 degrees of freedom
```

Multiple R-squared: 0.6746, Adjusted R-squared: 0.6441

F-statistic: 22.11 on 6 and 64 DF, p-value: 6.561e-14

Λαμβάνουμε τα εξής αποτελέσματα:

- Αρχικά παίρνουμε κάποια ποσοστιαία σημεία των υπολοίπων. Η μικρότερη και μεγαλύτερη τιμή που παίρνουν είναι -9.54 (ευρώ) και 9.20 (ευρώ) αντίστοιχα. Το 25% των υπολοίπων είναι μικρότερο από -3.27 (ευρώ), το 50% μικρότερο από 0.12(ευρώ) και το 75% μικρότερο από 2.54 (ευρώ).
- Όσο αφορά τους συντελεστές του μοντέλου μας, παίρνουμε αρχικά τις εκτιμήσεις τους, όπως και στο ερώτημα (iv), τα τυπικά τους σφάλματα από τη στήλη Std. Error, την t value του ελέγχου που έχει ως μηδενική υπόθεση ότι ο αντίστοιχος συντελεστής είναι 0, με εναλλακτική ότι δεν είναι 0 και το p value του ελέγχου. Στις μεταβλητές population categoryother και sexWoman το p value είναι μεγαλύτερο από 5%, συνεπώς σε ε.σ. 5% δεν υπάρχουν ισχυρές ενδείξεις για απόρριψη της υπόθεσης ότι ο αντίστοιχος συντελεστής είναι 0. Αντιθέτως, για τις μεταβλητές age,categoryhealthcare και med έχουμε σημαντικές ενδείξεις να απορρίψουμε την υπόθεση ότι ο αντίστοιχος συντελεστής είναι 0. Συνεπώς, όπως είδαμε και στο ερώτημα (iv) οι μεταβλητές age,categoryhealthcare και med είναι στατιστικά σημαντικές στο μοντέλο μας.
- Έπειτα έχουμε μια εκτίμηση της τυπικής απόκλισης των σφαλμάτων (τυπικό σφάλμα παλινδρόμησης) με 64 βαθμούς ελευθερίας που ισούται με 4.36 ευρώ.
- Το 67,46% της μεταβλητότητας του ποσού των χρημάτων σε ευρώ που ξόδεψε στην αγορά του μη φαρμακευτικού προϊόντος εξηγείται από το φύλο την κατηγορία του προϊόντος, την ηλικία του αγοραστή, τον πληθυσμό στην περιοχή που βρίσκεται το φαρμακείο και το ποσό των χρημάτων σε ευρώ που ξόδεψε στο φαρμακείο σε φαρμακευτικά προϊόντα τον προηγούμενο μήνα, με βάση το πολλαπλό γραμμικό μοντέλο παλινδρόμησης.
- Το 64,41% της μεταβλητότητας του ποσού των χρημάτων σε ευρώ που ξόδεψε στην αγορά του μη φαρμακευτικού προϊόντος εξηγείται από το φύλο την κατηγορία του προϊόντος, την ηλικία του αγοραστή, τον πληθυσμό στην περιοχή που βρίσκεται το φαρμακείο και το ποσό των χρημάτων σε ευρώ που ξόδεψε στο φαρμακείο σε φαρμακευτικά προϊόντα τον προηγούμενο μήνα, με βάση το πολλαπλό γραμμικό μοντέλο παλινδρόμησης, λαμβάνοντας υπόψη το μέγεθος του δείγματος (71) και το πλήθος των επεξηγηματικών μεταβλητών (5).
- Τέλος, παίρνουμε την τιμή του στατιστικού ελέγχου για την υπόθεση ότι $b_1=b_2=b_3=b_4=b_5=b_6=0$, με εναλλακτική ότι τουλάχιστον ένα από αυτά είναι διαφορετικό του 0. Το p value είναι 6.561e-14 που είναι πολύ μικρότερο του 5%, συνεπώς έχουμε πολύ ισχυρές ενδείξεις εναντίον της υπόθεσης ότι όλοι οι συντελεστές είναι 0. Δηλαδή, σε ε.σ. 5% υπάρχει τουλάχιστον μια στατιστικά σημαντική μεταβλητή στο μοντέλο μας.

Η κατηγορία αναφοράς της κατηγορικής μεταβλητής category είναι η κατηγορία καλλυντικών (cosmetics), ενώ της sex ο άνδρας (Man).

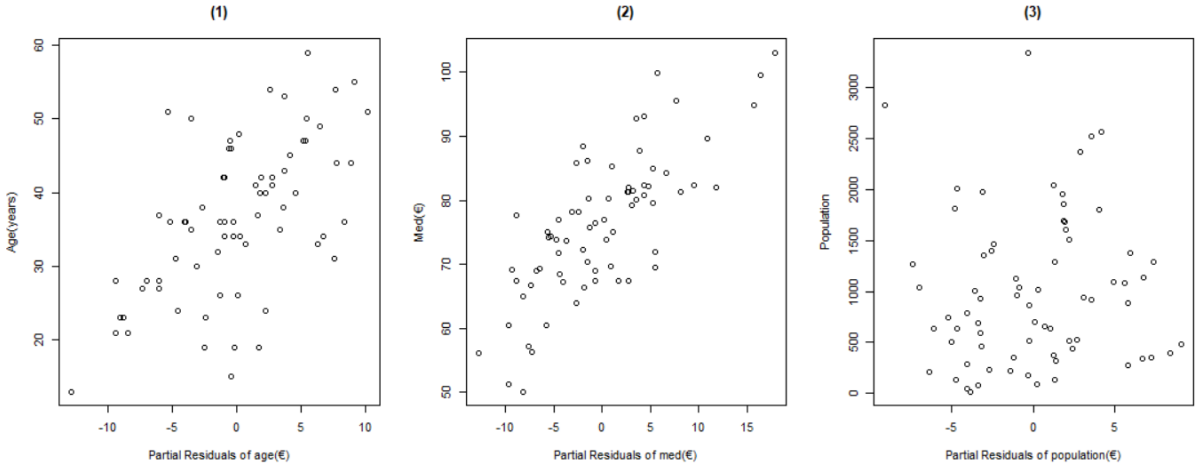
vi. Για να ελέγξουμε τις προϋποθέσεις του πολλαπλού γραμμικού μοντέλου εκτελούμε τα πιο κάτω:

<u>Γραμμικότητα</u>

Για τον έλεγχο της γραμμικότητας θα εξετάσουμε να η μεταβλητή money έχει γραμμική σχέση με τις ποσοτικές επεξηγηματικές μεταβλητές age, med και population, μέσω των μερικών υπολοίπων. Η υπόθεση αυτή θα ελεγχθεί ως εξής:

```
> Page<-residuals(results2,'partial')[,1]
> Pmed<-residuals(results2,'partial')[,4]
> Ppopulation<-residuals(results2,'partial')[,5]
> par(mfrow=c(1,3))
> plot(Page,age,xlab='Partial Residuals of age(€)',ylab='Age(years)',main="(1)")
```

> plot(Pmed,med,xlab='Partial Residuals of med(€)',ylab='Med(€)',main="(2)")
> plot(Ppopulation,population,xlab='Partial Residuals of population(€)', ylab='Population', main="(3)")

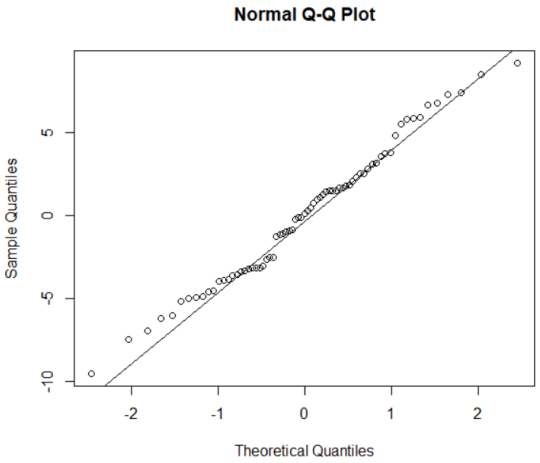


Στο γράφημα (1) παρατηρούμε πως υπάρχει μία γραμμική σχέση ανάμεσα στην μεταβλητή age και τα μερικά υπόλοιπα που της αντιστοιχούν. Στο γράφημα (2) παρατηρούμε πως υπάρχει μια ακόμη πιο γραμμική σχέση ανάμεσα στην μεταβλητή med και τα μερικά υπόλοιπα που της αντιστοιχούν. Στο γράφημα (3) μπορούμε και εδώ να υποθέσουμε πως υπάρχει μία γραμμικότητα, αφού δεν υπάρχουν πολύ σημαντικές αποκλίσεις. Επομένως, ικανοποιείται η υπόθεση της γραμμικότητας, σε μεγάλο βαθμό σε αυτό το πολλαπλό μοντέλο.

<u>Κανονικότητα Σφαλμάτων</u>

Για τον έλεγχο της κανονικότητας των σφαλμάτων θα δημιουργήσουμε qqplots για τα υπόλοιπα.

> qqnorm(results2$res)
> qqline(results2$res)

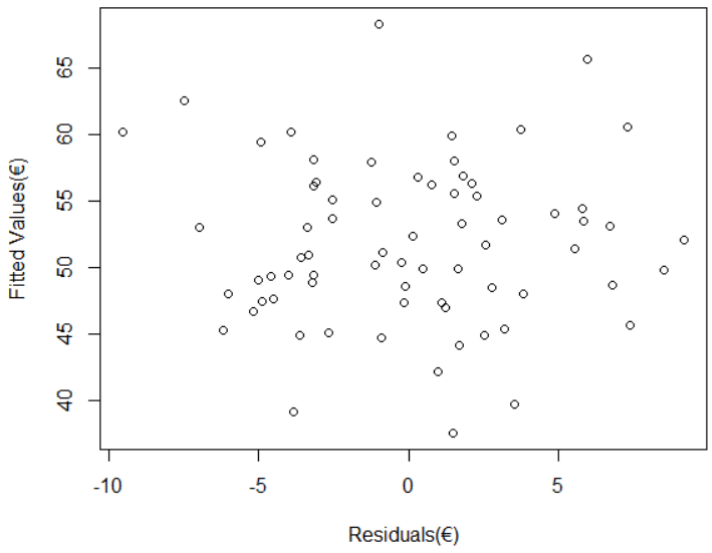


Από το διπλανό σχήμα μπορούμε εύκολα να συμπεράνουμε πως ισχύει και η προϋπόθεση της κανονικότητας σφαλμάτων, αφού παρατηρούμε πως τα υπόλοιπα ακολουθούν μια κανονική κατανομή.

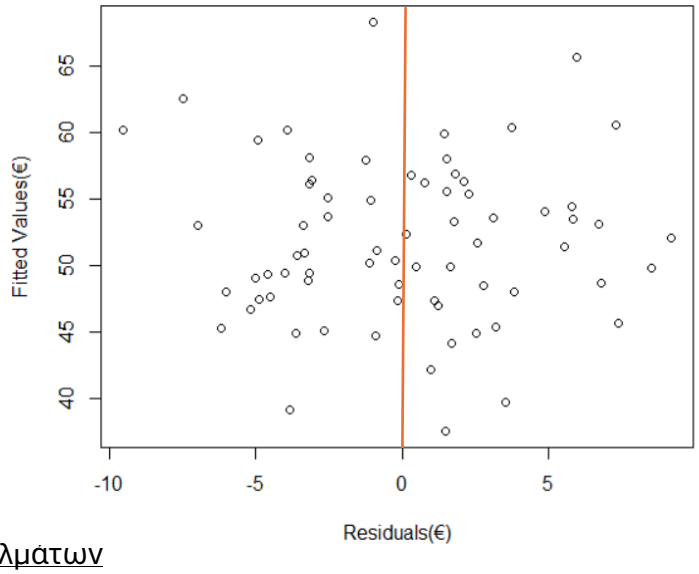
<u>Ομοσκεδαστικότητα</u>

Για το έλεγχο της ομοσκεδαστικότητας θα δημιουργήσουμε την γραφική παράσταση των υπολοίπων συναρτήσει των προβλεπόμενων τιμών.

> plot(results2$res,results2$fitted,xlab='Residuals(€)',ylab='Fitted Values(€)')



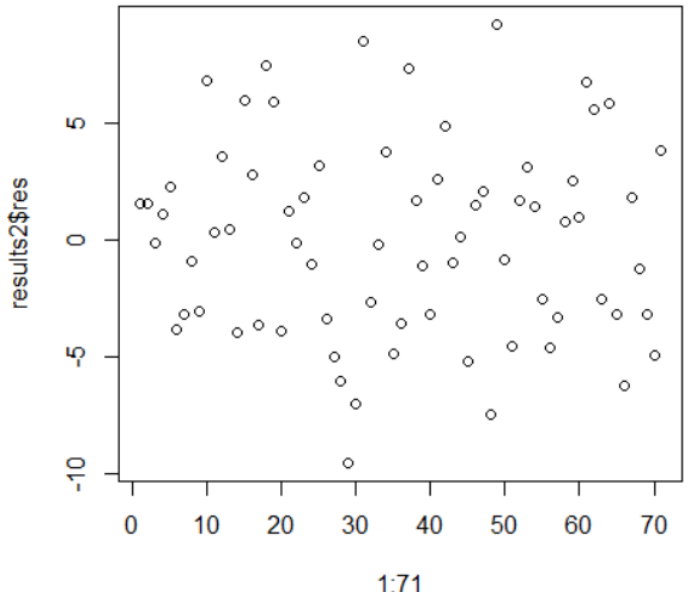
Από το πιο πάνω γραφήματα δεν παρατηρούμε κάποιο συστηματικό τρόπο συμπεριφοράς των δεδομένων, επομένως ισχύει η υπόθεση της ομοσκεδαστικότητας.
Επίσης, αν φέρουμε μία κάθετη γραμμή στο 0 σε ένα από τα γραφήματα παρατηρούμε πως δεξιά και αριστερά τις γραμμής έχουμε περίπου ίσο αριθμό σημείων, γεγονός που μας δείχνει πως ισχύει η γραμμικότητα.



Ανεξαρτησία Σφαλμάτων

Για να ελέγξουμε την ανεξαρτησία των σφαλμάτων κατασκευάζουμε ένα διάγραμμα υπολοίπων σε σχέση με τη σειρά των δεδομένων.

>plot(1:71,results2$res)



Από το παραπάνω γράφημα παρατηρούμε πως δεν παρουσιάζεται κάποια σχέση και τα υπόλοιπα συμπεριφέρονται τυχαία. Επομένως, ισχύει και η προϋπόθεση της ανεξαρτησίας σφαλμάτων.

vii. Προκειμένου να λάβουμε μια σημειακή εκτίμηση και ένα 95% Δ.Ε. με βάση το πιο πάνω πολλαπλό γραμμικό μοντέλο, για το αναμενόμενο ποσό που ξοδεύει σε μη φαρμακευτικά προϊόντα της κατηγορίας των καλλυντικών ένας 20 χρονών άντρας, ο οποίος έχει αγοράσει τον προηγούμενο μήνα φαρμακευτικά προϊόντα αξίας 60 ευρώ και κάνει τις αγορές του σε μια περιοχή με πληθυσμό 1500 κατοίκων, δίνουμε την εξής εντολή:

> predict(results2,int='c',list(category='cosmetics', age=20,sex='Man', med=60, population=1500))

	fit	lwr	upr
1	40.9911	37.89254	44.08966

Η σημειακή εκτίμηση για το αναμενόμενο ποσό που ξοδεύει σε μη φαρμακευτικά προϊόντα είναι 40.99 ευρώ, ενώ το διάστημα (37.89,44.09) περιέχει, με πιθανότητα 95%, το αναμενόμενο ποσό που ξοδεύει σε μη φαρμακευτικά προϊόντα κάποιο άτομο με τα προαναφερθέντα χαρακτηριστικά.

viii. Θεωρώντας ως κατηγορία αναφοράς της κατηγορικής μεταβλητής category τα υγειονομικά προϊόντα(healthcare) θα παρατηρήσουμε τα εξής σε σχέση με το μοντέλο του ερωτήματος (iv):

- Οι συντελεστές των μεταβλητών age, med, population και της εικονικής μεταβλητής sex δεν θα αλλάξουν σε σχέση με το προηγούμενο πολλαπλό γραμμικό μοντέλο
- Θα αλλάξει ο σταθερός όρος από 7.96 θα γίνει 3.57(7.96-4.39) εφόσον σε αυτό το μοντέλο θα εκφράζει το ποσό χρημάτων σε ευρώ που ξόδεψε στην αγορά του μη φαρμακευτικού προϊόντος στην περίπτωση που η ηλικία είναι 0, η κατηγορία του μη φαρμακευτικού προϊόντος αγοράς είναι healthcare και όχι cosmetics, το φύλο είναι άνδρας, το ποσό χρημάτων σε ευρώ που ξόδεψε στο φαρμακείο για φαρμακευτικά προϊόντα τον προηγούμενο μήνα και ο πληθυσμός είναι 0.
- Ο συντελεστής που αφορά την εικονική μεταβλητή categorycosmetics θα είναι ο αντίθετος από τον συντελεστή της categoryhealthcare του προηγούμενου μοντέλου αφού εκφράζουν αντίθετες ποσότητες. Σε αυτό το μοντέλο το ποσό χρημάτων σε ευρώ που ξόδεψε ένα άτομο στην αγορά του μη φαρμακευτικού προϊόντος που ανήκει στην κατηγορία cosmetics είναι 4.39 ευρώ μεγαλύτερο από το ποσό χρημάτων που ξόδεψε σε προϊόν της κατηγορίας healthcare ένα άτομο, ίδιας ηλικίας, ιδίου φύλου, που ξόδεψε ίσο ποσό σε φαρμακευτικά προϊόντα και σε κατάστημα σε περιοχή με ίδιο πληθυσμό.
- Ο συντελεστής που αφορά την εικονική μεταβλητή categoryother θα είναι 2.68(-1.71-(-4.39)) και δείχνει πως το ποσό χρημάτων σε ευρώ που ξόδεψε ένα άτομο στην αγορά του μη φαρμακευτικού προϊόντος που ανήκει στην κατηγορία other είναι 2.68 ευρώ μεγαλύτερο από το ποσό χρημάτων που ξόδεψε σε προϊόν της κατηγορίας healthcare ένα άτομο, ίδιας ηλικίας, ιδίου φύλου, που ξόδεψε ίσο ποσό σε φαρμακευτικά προϊόντα και σε κατάστημα σε περιοχή με ίδιο πληθυσμό.

ix. Για να κεντράρουμε τις τιμές των ποσοτικών επεξηγηματικών μεταβλητών age, med και population χρησιμοποιούμε τις εξής εντολές:

```
> mean(age)
[1] 36.40845
> mean(med)
[1] 76.19155
> mean(population)
[1] 994.8451
> agecentered<-age-mean(age)
> medcentered<-med-mean(med)
> populationcentered<-population-mean(population)
```

Αν στο μοντέλο του ερωτήματος (iv) είχαμε αυτές τις κεντραρισμένες τιμές δεν θα άλλαζε κάτι όσο αφορά τη στατιστική συμπερασματολογία. Αυτό που θα άλλαζε είναι ότι θα είχαμε νέες επεξηγηματικές μεταβλητές και για αυτό ο σταθερός όρος θα γινόταν:

(mean(age)*0.2902512+mean(med)*0.4475493+mean(population)*0.0002518)+7.9553472=52.8729643

Επομένως, το ποσό χρημάτων σε ευρώ που ξόδεψε ένα άτομο στην αγορά του μη φαρμακευτικού προϊόντος στην περίπτωση που η ηλικία είναι 36.41, η κατηγορία του μη φαρμακευτικού προϊόντος αγοράς είναι cosmetics, το φύλο είναι άνδρας, το ποσό χρημάτων σε ευρώ που ξόδεψε στο φαρμακείο για φαρμακευτικά προϊόντα τον προηγούμενο μήνα είναι 76.19 και ο πληθυσμός είναι 994.85 κάτοικοι, είναι 52.87 ευρώ.

x. Για να προσαρμόσουμε το πολλαπλό πολλαπλασιαστικό μοντέλο με χρήση του νεπέριου λογαρίθμου ορίζουμε τις βοηθητικές μεταβλητές:

```
> logMoney=log(money)
> logAge=log(age)
> logMed=log(med)
> logPopulation=log(population)
```

Και προσαρμόζουμε το μοντέλο με την εντολή:

```
> results3<-lm(logMoney~logAge+category+sex+logMed+logPopulation)
> results3
Call:
lm(formula = logMoney ~ logAge + category + sex + logMed + logPopulation)
Coefficients:
    (Intercept)          logAge  categoryhealthcare
        0.41592         0.17764         -0.07607
 categoryother       sexWoman          logMed
       -0.03327         0.04200         0.65087
 logPopulation
        0.01438
```

Ερμηνεία των εκτιμητών των συντελεστών του παραπάνω μοντέλου:

- Το $e^{0.42} = 1.52$ εκφράζει τη διάμεσο του ποσού χρημάτων σε ευρώ που ξόδεψε ένα άτομο στην αγορά του μη φαρμακευτικού προϊόντος στην περίπτωση που η ηλικία είναι 1 έτος, η κατηγορία του μη φαρμακευτικού προϊόντος αγοράς είναι cosmetics, το φύλο είναι άνδρας, το ποσό χρημάτων σε ευρώ που ξόδεψε στο φαρμακείο για φαρμακευτικά προϊόντα τον προηγούμενο μήνα είναι 1 και ο πληθυσμός είναι 1 κάτοικος.
- $e^{0.18} = 1.20$. Η διάμεσος του ποσού χρημάτων σε ευρώ που ξόδεψε ένα άτομο στην αγορά του μη φαρμακευτικού προϊόντος ανεξαρτήτως κατηγορίας, φύλου, πληθυσμού και του ποσού χρημάτων για φαρμακευτικά προϊόντα, αυξάνεται κατά 19.72% αν η ηλικία του ατόμου αυξηθεί κατά 1 έτος.
- $e^{-0.08} = 0.92$: Το ποσό χρημάτων σε ευρώ που ξόδεψε ένα άτομο στην αγορά του μη φαρμακευτικού προϊόντος που ανήκει στην κατηγορία healthcare μειώνεται κατά 8% συγκριτικά με το ποσό χρημάτων που ξόδεψε σε προϊόν της κατηγορίας cosmetics ένα άτομο, ίδιας ηλικίας, ιδίου φύλου, που ξόδεψε ίσο ποσό σε φαρμακευτικά προϊόντα και σε κατάστημα σε περιοχή με ίδιο πληθυσμό.
- $e^{-0.03} = 0.97$: Το ποσό χρημάτων σε ευρώ που ξόδεψε ένα άτομο στην αγορά του μη φαρμακευτικού προϊόντος που ανήκει στην κατηγορία other μειώνεται κατά 3% συγκριτικά με το ποσό χρημάτων που ξόδεψε σε προϊόν της κατηγορίας cosmetics ένα άτομο, ίδιας ηλικίας, ιδίου φύλου, που ξόδεψε ίσο ποσό σε φαρμακευτικά προϊόντα και σε κατάστημα σε περιοχή με ίδιο πληθυσμό.
- $e^{0.04} = 1.04$: Το ποσό χρημάτων σε ευρώ που ξόδεψε μια γυναίκα στην αγορά του μη φαρμακευτικού προϊόντος αυξάνεται κατά 4% συγκριτικά με αυτό που ξόδεψε ένας άνδρας, ανεξαρτήτως ηλικίας, κατηγορίας, πληθυσμού και ποσού που ξόδεψε σε φαρμακευτικά προϊόντα.
- $e^{0.65} = 1.92$: Η διάμεσος του ποσού χρημάτων σε ευρώ που ξόδεψε ένα άτομο στην αγορά του μη φαρμακευτικού προϊόντος, ανεξαρτήτως ηλικίας, κατηγορίας, πληθυσμού και φύλου, αυξάνεται κατά 91.55% αν το ποσό που ξόδεψε σε φαρμακευτικά προϊόντα αυξηθεί κατά 1 ευρώ.

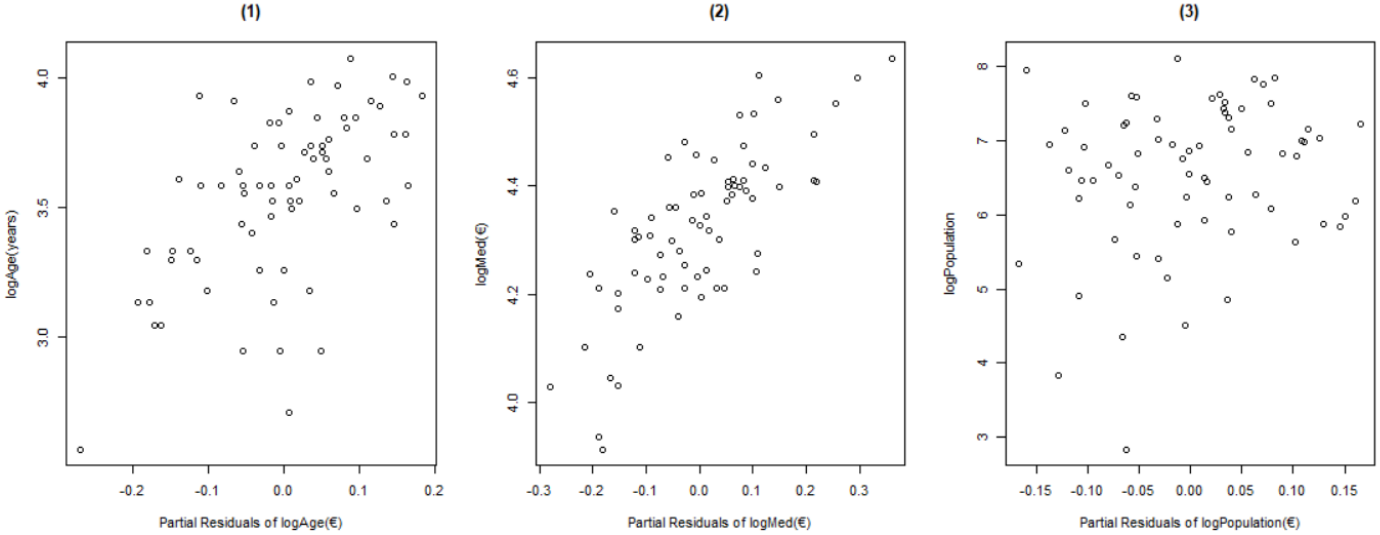
- $e^{0.01}$ = 1.01: Το ποσό χρημάτων σε ευρώ που ξόδεψε ένα άτομο στην αγορά του μη φαρμακευτικού προϊόντος, ανεξαρτήτως ηλικίας, κατηγορίας, φύλου και ποσού που ξοδεύτηκε σε φαρμακευτικά προϊόντα, αυξάνεται κατά 1% αν ο αριθμός των κατοίκων της περιοχής που βρίσκεται το φαρμακείο αυξηθεί κατά 1.

xi. Για να ελέγξουμε τις προϋποθέσεις του πολλαπλού πολλαπλασιαστικού μοντέλου εκτελούμε τα πιο κάτω:

Γραμμικότητα

Για τον έλεγχο της γραμμικότητας θα εξετάσουμε να η μεταβλητή logMoney έχει γραμμική σχέση με τις επεξηγηματικές μεταβλητές logAge, logMed και logPopulation, μέσω των μερικών υπολοίπων. Η υπόθεση αυτή θα ελεγχθεί ως εξής:

```
> Plage<-residuals(results3,'partial')[,1]
> Plmed<-residuals(results3,'partial')[,4]
> Plpopulation<-residuals(results3,'partial')[,5]
> par(mfrow=c(1,3))
> plot(Plage,logAge,xlab='Partial Residuals of logAge(€)',ylab='logAge(years)',main="(1)")
> plot(Plmed,logMed,xlab='Partial Residuals of logMed(€)',ylab='logMed(€)',main="(2)")
> plot(Plpopulation,logPopulation,xlab='Partial Residuals of logPopulation(€)',
ylab='logPopulation', main="(3)")
```

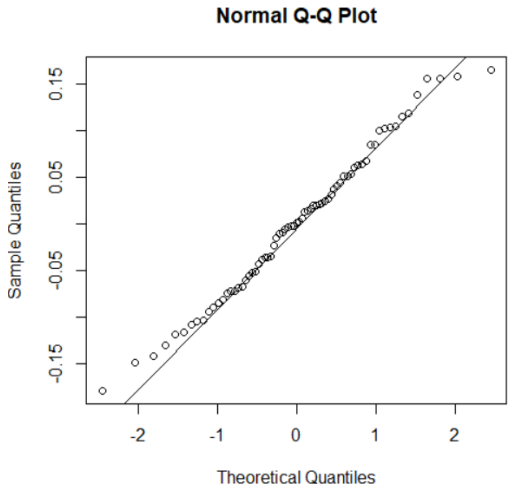


Στο γράφημα (1) παρατηρούμε πως υπάρχει μία γραμμική σχέση ανάμεσα στην μεταβλητή logAge και τα μερικά υπόλοιπα που της αντιστοιχούν. Στο γράφημα (2) παρατηρούμε πως υπάρχει μια ακόμη πιο γραμμική σχέση ανάμεσα στην μεταβλητή logMed και τα μερικά υπόλοιπα που της αντιστοιχούν. Στο γράφημα (3) μπορούμε και εδώ να υποθέσουμε πως υπάρχει μία γραμμικότητα, αφού δεν υπάρχουν πολύ σημαντικές αποκλίσεις. Επομένως, ικανοποιείται η υπόθεση της γραμμικότητας, σε μεγάλο βαθμό σε αυτό το πολλαπλό μοντέλο εφόσον δεν υπάρχουν σημαντικές ενδείξεις εναντίον αυτής.

Κανονικότητα Σφαλμάτων

Για τον έλεγχο της κανονικότητας των σφαλμάτων θα δημιουργήσουμε qqplots για τα υπόλοιπα.

```
> qqnorm(results3$res)
> qqline(results3$res)
```

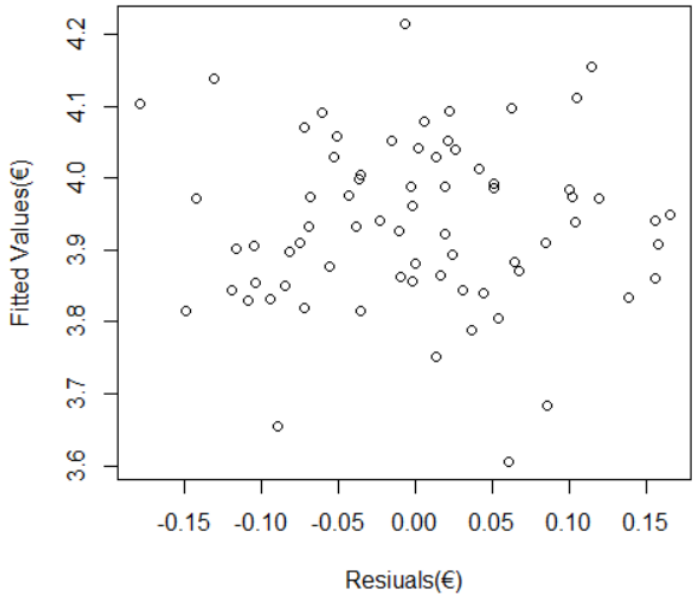


Από το διπλανό σχήμα μπορούμε εύκολα να συμπεράνουμε πως ισχύει και η προϋπόθεση της κανονικότητας σφαλμάτων, αφού παρατηρούμε πως τα υπόλοιπα ακολουθούν μια κανονική κατανομή.

Ομοσκεδαστικότητα

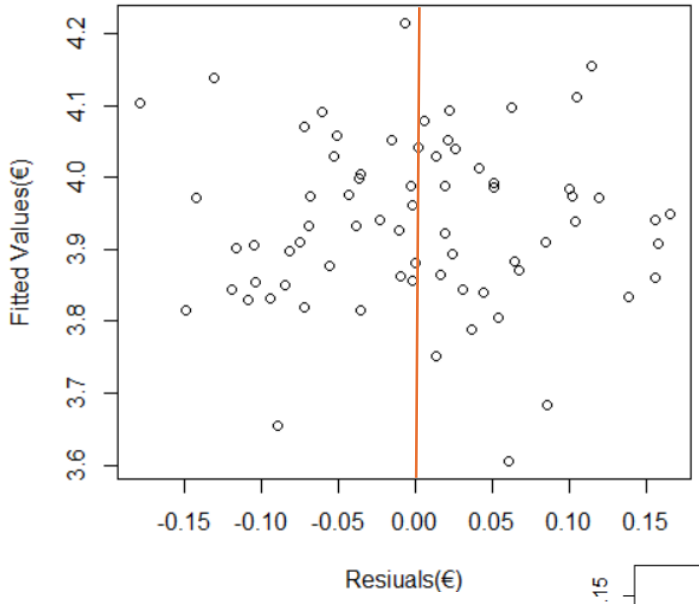
Για το έλεγχο της ομοσκεδαστικότητας θα δημιουργήσουμε την γραφική παράσταση των υπολοίπων συναρτήσει των προβλεπόμενων τιμών.

> plot(results3$res,results3$fitted,xlab='Residuals(€)',ylab='Fitted Values(€)')



Από το πιο πάνω γραφήματα δεν παρατηρούμε κάποιο συστηματικό τρόπο συμπεριφοράς των δεδομένων, επομένως ισχύει η υπόθεση της ομοσκεδαστικότητας.

Επίσης, αν φέρουμε μία κάθετη γραμμή στο 0 σε ένα από τα γραφήματα παρατηρούμε πως δεξιά και αριστερά τις γραμμής έχουμε περίπου ίσο αριθμό σημείων, γεγονός που μας δείχνει πως ισχύει η γραμμικότητα.

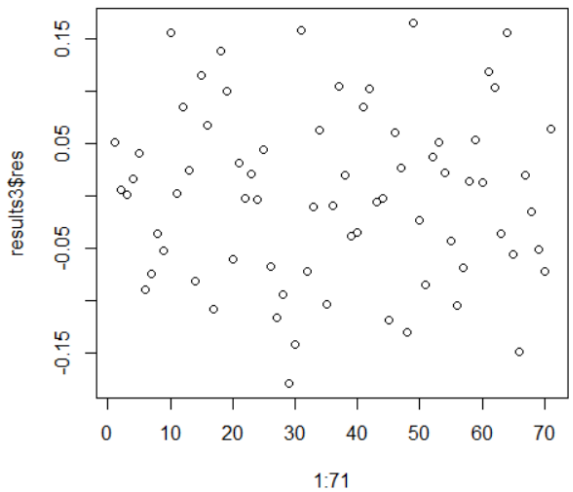


Ανεξαρτησία Σφαλμάτων

Για να ελέγξουμε την ανεξαρτησία των σφαλμάτων κατασκευάζουμε ένα διάγραμμα υπολοίπων σε σχέση με τη σειρά των δεδομένων.

>plot(1:71,results3$res)

Από το διπλανό γράφημα παρατηρούμε πως δεν παρουσιάζεται κάποια σχέση και τα υπόλοιπα συμπεριφέρονται τυχαία. Επομένως, ισχύει και η προϋπόθεση της ανεξαρτησίας σφαλμάτων.



xii. Προκειμένου να λάβουμε μια σημειακή εκτίμηση και ένα 95% Δ.Ε. με βάση το πιο πάνω πολλαπλό πολλαπλασιαστικό μοντέλο, για το αναμενόμενο ποσό που ξοδεύει σε μη φαρμακευτικά προϊόντα της κατηγορίας των καλλυντικών ένας 20 χρονών άντρας, ο οποίος έχει αγοράσει τον προηγούμενο μήνα φαρμακευτικά προϊόντα αξίας 60 ευρώ και κάνει τις αγορές του σε μια περιοχή με πληθυσμό 1500 κατοίκων, δίνουμε την εξής εντολή:

```
> predict(results3,int='c',list(category='cosmetics', logAge=log(20),sex='Man', logMed=log(60), logPopulation=log(1500)))

        fit      lwr      upr

1 3.718146 3.652237 3.784054
```

Η σημειακή εκτίμηση για το αναμενόμενο ποσό που ξοδεύει σε μη φαρμακευτικά προϊόντα είναι $e^{3.72}$=41.26 ευρώ, ενώ το διάστημα ($e^{3.65}$=38.47, $e^{3.78}$= 43.82) περιέχει, με πιθανότητα 95%, το αναμενόμενο ποσό που ξοδεύει σε μη φαρμακευτικά προϊόντα κάποιο άτομο με τα προαναφερθέντα χαρακτηριστικά.

xiii. Θα συγκρίνω μεταξύ τους το πολλαπλό γραμμικό μοντέλο και το πολλαπλό πολλαπλασιαστικό μοντέλο. Με βάση τον έλεγχο προϋποθέσεων που έγινε και για τα δύο μοντέλα δεν παρατηρείται κάποια έντονη διαφορά, ώστε να επιλέξουμε ως βέλτιστο ένα από τα δύο. Για περεταίρω σύγκριση χρησιμοποιώ το κριτήριο AIC, το οποίο εκτός από την καλή προσαρμογή λαμβάνει υπόψη και την πολυπλοκότητα των μοντέλων.

```
> AIC(results2)
[1] 419.1145
> AIC(results3)+2*sum(log(money))
[1] 419.9061
```

Με βάση το κριτήριο αυτό βέλτιστο μοντέλο είναι αυτό με το μικρότερο AIC, δηλαδή το πολλαπλό γραμμικό μοντέλο. Παρ' όλα αυτά και τα δύο μοντέλα παρουσιάζουν πολύ μικρή διαφορά και στις τιμές των AIC και στους ελέγχους προϋποθέσεων, επομένως είναι και τα δύο εξ' ίσου καλά προσαρμοσμένα και κατ' επέκταση και οι δύο προβλέψεις των ερωτημάτων (iv) και (x) εξ' ίσου έμπιστες.