(B) i) Αρχικά εισάγουμε τα δεδομένα μας στην R με τις εξής εντολές:

> x<-c(1.8,3.0,4.8,5.0,6.5,7.0,9.0,9.1)

> y<-c(20.0,30.5,40.0,55.1,60.3,74.9,88.4,95.2)

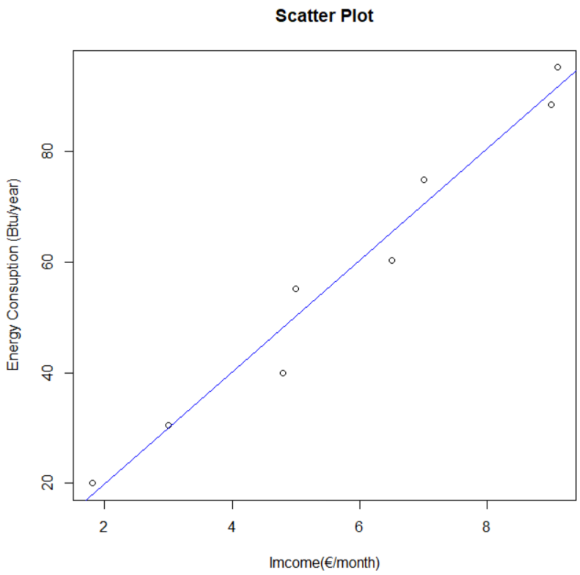
Στη συνέχεια για τη δημιουργία του διαγράμματος διασποράς των μεταβλητών y και x δίνουμε την εντολή:

> plot(x,y,main="Scatter Plot",xlab="Imcome(x)",ylab="Energy Consuption(y)")

Και τις επόμενες εντολές για προσαρμογή του μοντέλου παλινδρόμησης και εμφάνισης του στο διάγραμμα:

> model<-lm(y~x)

> abline(model,col="blue")



ii) Για την εύρεση του συντελεστή προσδιορισμού R^2 δίνουμε την εντολή:

> summary(model)$r.squared

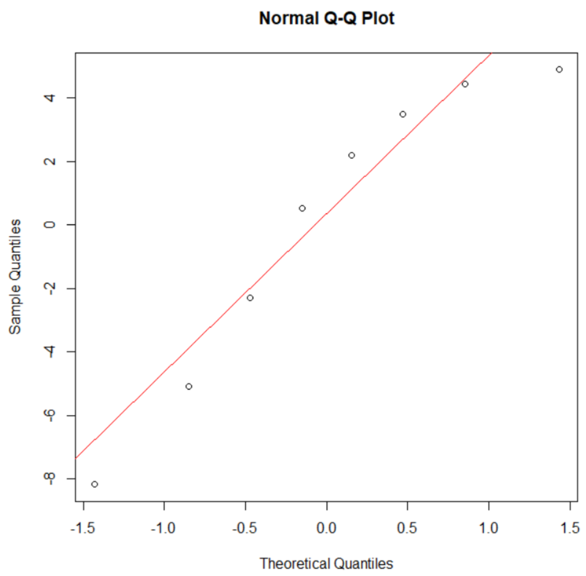
[1] 0.9689917

Παρατηρούμε πως ο συντελεστής προσδιορισμού R^2=0,9690. Επομένως, μπορούμε να συμπεράνουμε πως έχουμε ένα πολύ καλά προσαρμοσμένο μοντέλο αφού η τιμή του συντελεστή προσδιορισμού είναι αρκετά κοντά στη μονάδα. Επίσης, ο συντελεστής προσδιορισμού εκφράζει το ποσοστό της διασποράς της τυχαίας μεταβλητής y που εξηγείται με βάση το πιο πάνω μοντέλο παλινδρόμησης. Όσο πιο κοντά στη μονάδα είναι ο συντελεστής τόσο ισχυρότερη είναι η γραμμική σχέση μεταξύ των τ.μ. y και x. O R^2 υποδηλώνει ότι το 96,90% της διακύμανσης στην ενεργειακή κατανάλωση σε Btu/έτος εξηγείται από το εισόδημα σε 1000 ευρώ/μήνα.

iii) Για τον έλεγχο της κανονικότητας θα χρησιμοποιήσουμε 5 διαφορετικούς τρόπους:

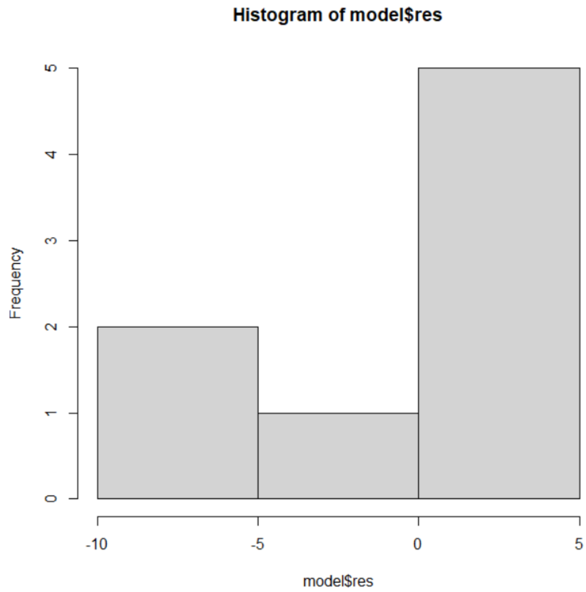
α) > qqnorm(residuals(model))

> qqline(residuals(model),col="red")



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

β) > hist(model$res)



Από το πιο πάνω σχήμα παρατηρούμε ότι υπάρχει κάποια συμμετρία, αλλά δεν φαίνεται να ακολουθεί την κλασική μορφή συμμετρίας της κανονικής κατανομής. Μια κανονική κατανομή θα είχε συμμετρική διάταξη γύρω από το μηδέν, με τις τιμές να είναι πιο συχνές κοντά στη μέση και να μειώνονται στα άκρα. Εδώ, υπάρχει συσσώρευση τιμών κοντά στο μηδέν είτε σε αρνητικές τιμές κοντά στο -10 και μια μεγαλύτερη κατανομή στα θετικά υπολείμματα, κάτι που υποδηλώνει πιθανή απόκλιση από την κανονικότητα ή την ύπαρξη σφαλμάτων ή επιδράσεων που δεν έχουν συμπεριληφθεί σωστά στο μοντέλο

Τα αποτελέσματα αυτού του ελέγχου δεν συμφωνούν με τα αποτελέσματα του προηγούμενου ελέγχου και των επόμενων που παρουσιάζονται αφού εδώ δεν φαίνεται η ύπαρξη της κανονικότητας όπως στους άλλους ελέγχους.

γ) > shapiro.test(residuals(model))

Shapiro-Wilk normality test

data: residuals(model)

W = 0.91031, p-value = 0.3562

Από τον συγκεκριμένο έλεγχο παρατηρούμε ότι η p-value = 0.3562>0.05 που είναι το επίπεδο σημαντικότητας και συνεπώς δεν απορρίπτουμε την αρχική υπόθεση της κανονικότητας. Επομένως, ισχύει η κανονικότητα.

δ) > ks.test(model$res,"pnorm",mean(model$res),sd(model$res))

Exact one-sample Kolmogorov-Smirnov test

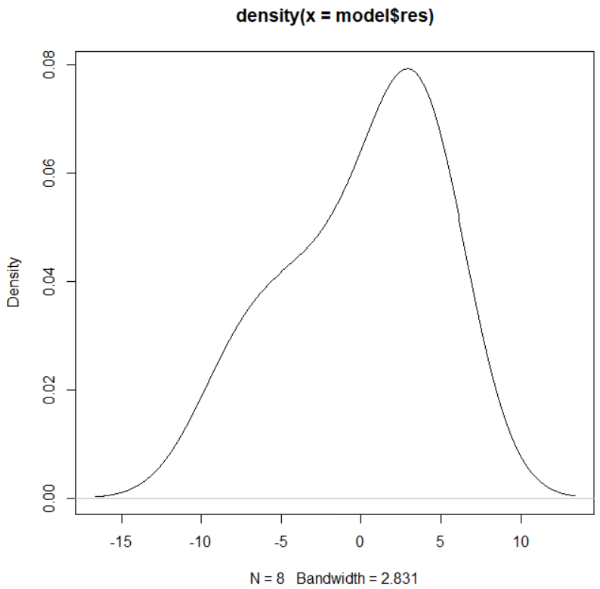
data: model$res

D = 0.17655, p-value = 0.9294

alternative hypothesis: two-sided

Και από αυτόν τον έλεγχο παρατηρούμε ότι η p-value = 0.9294>0.05 που είναι το επίπεδο σημαντικότητας επομένως αποδεχόμαστε την αρχική υπόθεση της κανονικότητας. Άρα, ισχύει η κανονικότητα.

ε) > plot(density(model$res))



Από το παραπάνω διάγραμμα παρατηρούμε πως ισχύει η κανονικότητα, αφού η καμπύλη των καταλοίπων είναι παρόμοια με την καμπύλη της κανονικής κατανομής.

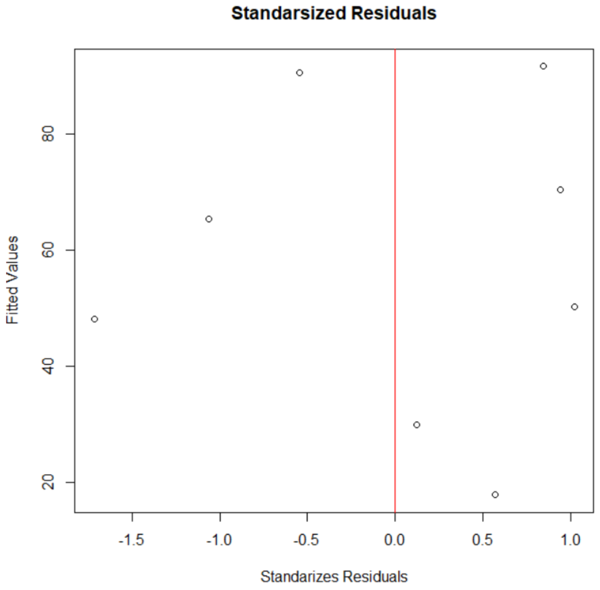
iv) Ο έλεγχος της προϋπόθεσης της κανονικότητας με χρήση των standardized residuals έχει παρουσιαστεί στο προηγούμενο ερώτημα. Για τον έλεγχο των υπόλοιπων προϋποθέσεων του μοντέλου χρησιμοποιώντας τα standardized residuals δίνουμε τις εξής εντολές:

> plot(rstandard(model),fitted(model), main="Standarsized Residuals", ylab="Fitted Values", xlab="Standarizes Residuals")

> abline(v=0,col="red")

Από το παρακάτω γράφημα, που μας έδωσαν οι εντολές, δεν παρατηρούμε κάποιο συστηματικό τρόπο συμπεριφοράς των δεδομένων, επομένως ισχύει η υπόθεση της ομοσκεδαστικότητας.

Επίσης, παρατηρούμε ότι δεξιά και αριστερά της κάθετης κόκκινης γραμμής στο 0 έχουμε περίπου ίσο αριθμό σημείων, γεγονός που μας δείχνει ότι ισχύει η γραμμικότητα.

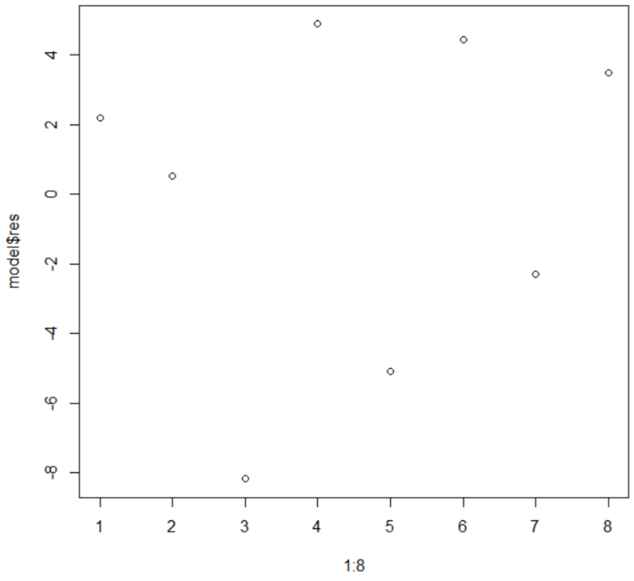


Επιπλέον, με την εντολή:

> plot(1:8,model$res)

μπορούμε να εξετάσουμε την προϋπόθεση της ανεξαρτησίας των σφαλμάτων.

Από το παρακάτω γράφημα, που μας δίνουν οι πιο πάνω εντολές, παρατηρούμε πως δεν παρουσιάζεται κάποια σχέση και τα υπόλοιπα συμπεριφέρονται τυχαία. Επομένως, ισχύει και η προϋπόθεση της ανεξαρτησίας των σφαλμάτων.



v) Για την κατασκευή ενός 95% δ.ε. για τον συντελεστή $β_1$ δίνουμε την εντολή:

```
> confint(model)
                2.5 %   97.5 %
(Intercept) -11.75914  10.95272
x             8.31302  11.93052
```

και παρατηρούμε πως το 95% δ.ε. του $β_1$ είναι [8.3130,11.9305]

Ερμηνεία του συντελεστή $β_1$: είναι η αναμενόμενη μεταβολή της ενεργειακής κατανάλωσης y (Btu/έτος) όταν το εισόδημα x μεταβληθεί κατά 1000 ευρώ/έτος (δηλαδή κατά 1 μονάδα).

Από την εντολή summary(model) παρατηρούμε ότι $β_1$=10.1218. Επομένως, η ενεργειακή κατανάλωση y (Btu/έτος) αυξάνεται κατά 10.1218 αν το εισόδημα x μεταβληθεί κατά 1*1000=1000 ευρώ/έτος.

iv) Για την σημειακή εκτίμηση της ενεργειακής κατανάλωσης όταν x_0=8 και για το 95% δ.ε. ατομικής πρόβλεψης για την αντίστοιχη παρατήρηση δίνουμε την εντολή:

```
> predict(model,list(x=c(8)),int="p")
       fit      lwr      upr
1 80.57093 66.61278 94.52909
```

Παρατηρούμε ότι y_{x0}=80.5709 και το 95% δ.ε. ατομικής πρόβλεψης είναι [66.6128, 94.5291]

Για το 95% δ.ε. μέσης πρόβλεψης δίνουμε την εντολή:

```
> predict(model,list(x=c(8)),int="c")
       fit      lwr      upr
1 80.57093 74.56723 86.57464
```

Παρατηρούμε το 95% δ.ε. μέσης πρόβλεψης είναι [74.5672, 86.5746]

Το διάστημα εμπιστοσύνης ατομικής πρόβλεψης είναι πιο ευρύ από το διάστημα εμπιστοσύνης μέσης πρόβλεψης.

(Γ) i) Αρχικά εισάγουμε τα δεδομένα μας στην R με τις εξής εντολές:

```
> x<-c(2,4,6,12,18,24)
> y<-c(1.07,1.88,2.26,2.78,2.97,2.99)
```

Στη συνέχεια για την προσαρμογή του μοντέλου κάνουμε τον εξής μετασχηματισμό:

1. Ορίζουμε το y'=3−y, έτσι ώστε:

   $y'=ae^{βx}$

2. Παίρνουμε τον φυσικό λογάριθμο των δύο μελών:

   $\ln(y')=\ln(a)+βx$

Έτσι, το πρόβλημα μετατρέπεται σε μια γραμμική σχέση μεταξύ ln(y') και x.

Για την προσαρμογή στη R δίνουμε τις εντολές:

> logty<-log(3-y)

> model<-lm(logty~x)

Και από την εντολή:

> summary(model)

παρατηρούμε τα αποτελέσματα του μοντέλου.

Για τις γραφικές παραστάσεις πριν και μετά τον μετασχηματισμό δίνουμε τις εντολές:

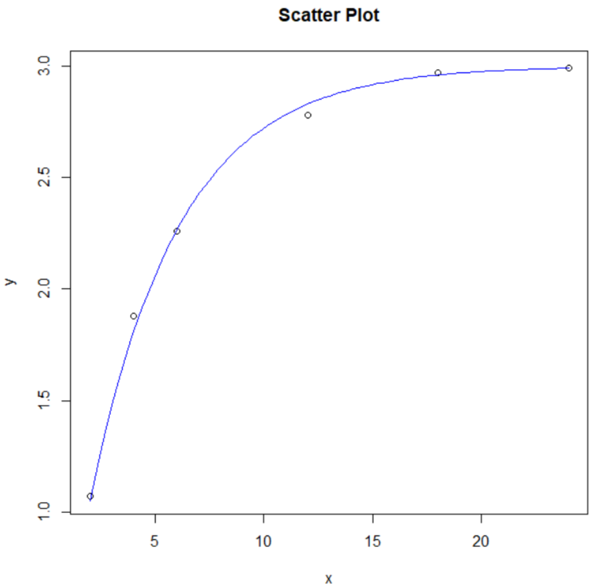
ΠΡΙΝ

> a<-exp(coef(model)[1])

> b<-coef(model)[2]

> plot(x,y,main="Scatter Plot", xlab="x", ylab="y")

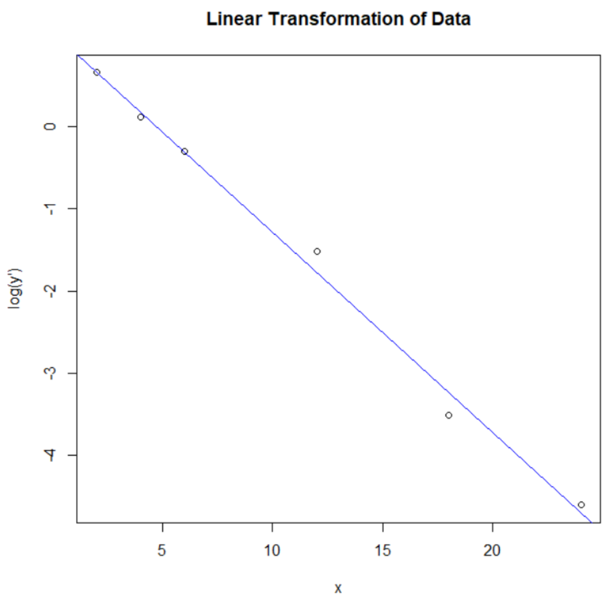
> curve(3-a*exp(b*x),add=TRUE,col="blue")



ΜΕΤΑ

> plot(x, logty, main="Linear Transformation of Data", xlab="x", ylab="log(y')")

> abline(model,col="blue")



ii) Έλεγχος προϋποθέσεων πριν και μετά τον μετασχηματισμό:

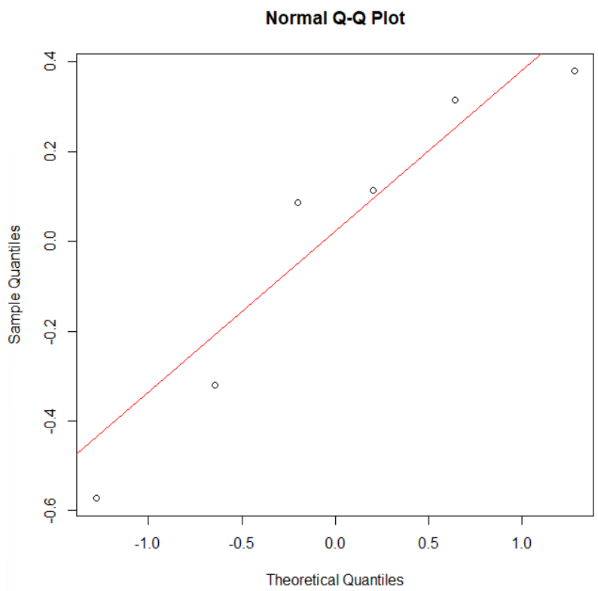
ΠΡΙΝ

Πριν από τον μετασχηματισμό δεν έχουμε γραμμική σχέση γι' αυτό και χρειάζεται να κάνουμε τον μετασχηματισμό για να προσαρμόσουμε το μοντέλο. Έστω όμως ότι προσαρμόζουμε το μοντέλο χωρίς τον μετασχηματισμό με την εντολή:

```
> model2<-lm(y~x)
```

Για να ελέγξουμε τις προϋποθέσεις δίνουμε τις εξής εντολές:

```
> qqnorm(residuals(model2))
> qqline(residuals(model2),col="red")
```



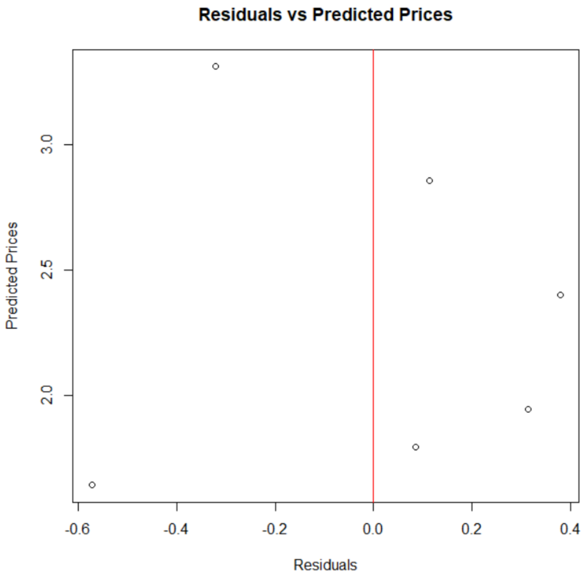
Από το παραπάνω σχήμα μπορούμε να συμπεράνουμε πως ισχύει η προϋπόθεση της κανονικότητας σφαλμάτων, αφού παρατηρούμε πως τα υπόλοιπα ακολουθούν μια κανονική κατανομή.

Διεξάγοντας και τον έλεγχο:

```
> shapiro.test(residuals(model2))

	Shapiro-Wilk normality test

data:  residuals(model2)

W = 0.90608, p-value = 0.4111
```

Διαπιστώνουμε πως ισχύει η κανονικότητα αφού η , p-value = 0.4111> 0.05 που είναι το επίπεδο σημαντικότητας και συνεπώς δεν απορρίπτουμε την αρχική υπόθεση της κανονικότητας.

```
> plot(residuals(model2),fitted(model2), main="Residuals vs Predicted Prices", ylab="Predicted Prices",xlab="Residuals")

> abline(v=0,col="red")
```

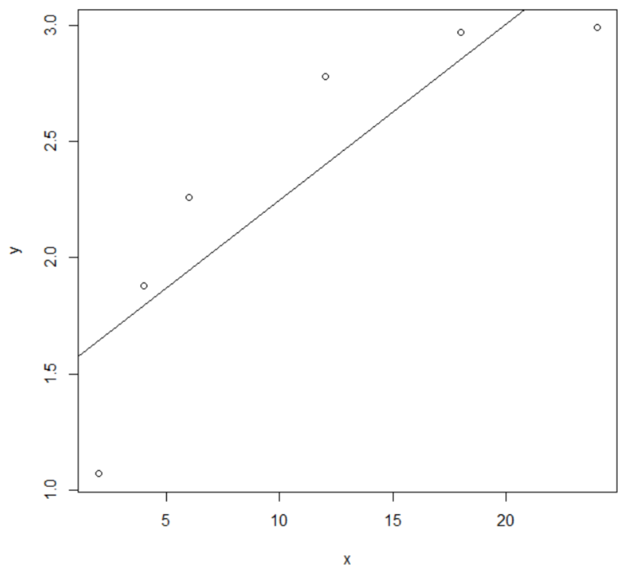


Από το παραπάνω γράφημα, παρατηρούμε πως τα δεδομένα σχηματίζουν μια καμπύλη, επομένως δεν ισχύει η υπόθεση της ομοσκεδαστικότητας.

Επίσης, παρατηρούμε ότι δεξιά και αριστερά της κάθετης κόκκινης γραμμής στο 0 δεν έχουμε ίσο αριθμό σημείων, γεγονός που μας δείχνει ότι δεν ισχύει η γραμμικότητα.

Την γραμμικότητα μπορούμε να την εξετάσουμε και από τις εντολές:

```
> plot(x,y)

> abline(model2)
```

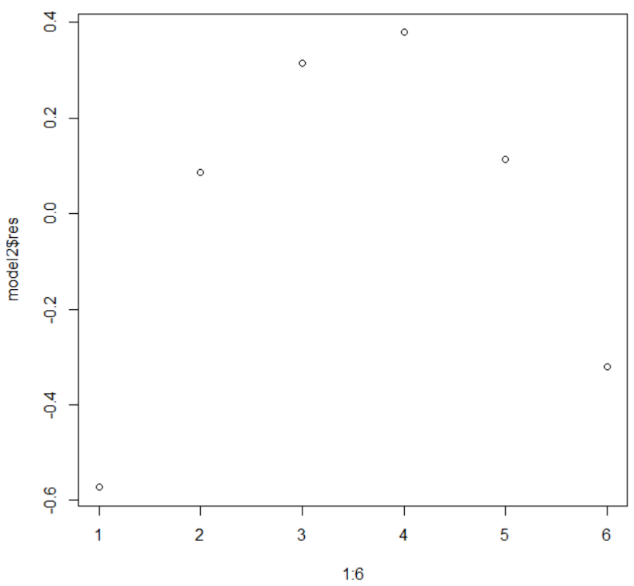


Από το παραπάνω γράφημα παρατηρούμε πως δεν ισχύει η προϋπόθεση της γραμμικότητας, αφού τα σημεία δημιουργούν μία καμπύλη.

Επιπλέον, με την εντολή:

```
> plot(1:6,model2$res)
```

μπορούμε να εξετάσουμε την προϋπόθεση της ανεξαρτησίας των σφαλμάτων.

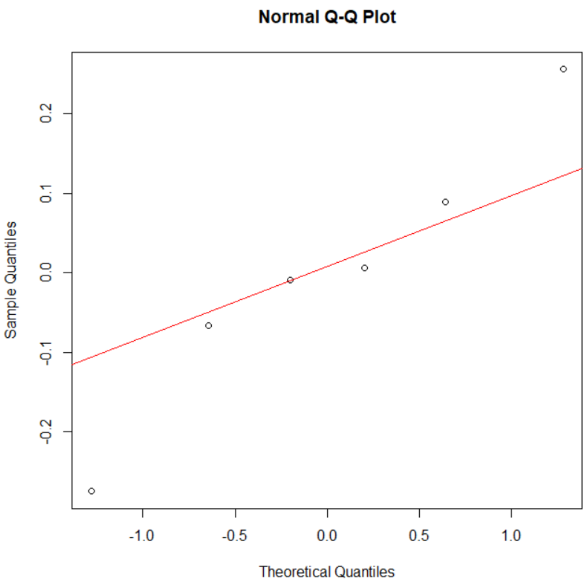


Από το παραπάνω γράφημα παρατηρούμε πως παρουσιάζεται κάποια σχέση και τα υπόλοιπα δημιουργούν μία καμπύλη. Επομένως, δεν ισχύει η προϋπόθεση της ανεξαρτησίας των σφαλμάτων.

META

> qqnorm(residuals(model))

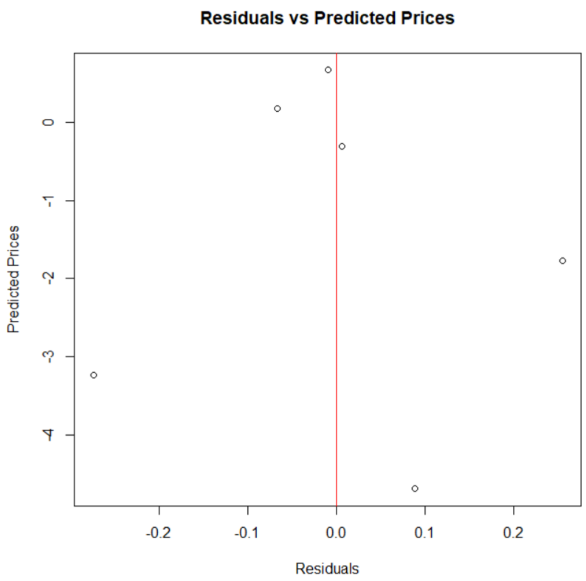
> qqline(residuals(model),col="red")



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

> plot(residuals(model),fitted(model), main="Residuals vs Predicted Prices", ylab="Predicted Prices",xlab="Residuals")

> abline(v=0,col="red")



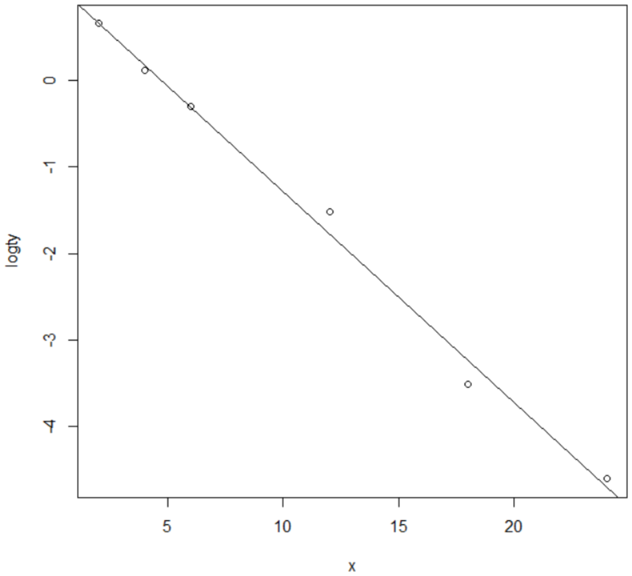
Από το παραπάνω γράφημα, δεν παρατηρούμε κάποιο συστηματικό τρόπο συμπεριφοράς των δεδομένων, επομένως ισχύει η υπόθεση της ομοσκεδαστικότητας.

Επίσης, παρατηρούμε ότι δεξιά και αριστερά της κάθετης κόκκινης γραμμής στο 0 έχουμε ίσο αριθμό σημείων, γεγονός που μας δείχνει ότι ισχύει η γραμμικότητα.

Την γραμμικότητα μπορούμε να την εξετάσουμε και από τις εντολές:

> plot(x,logty)

> abline(model)

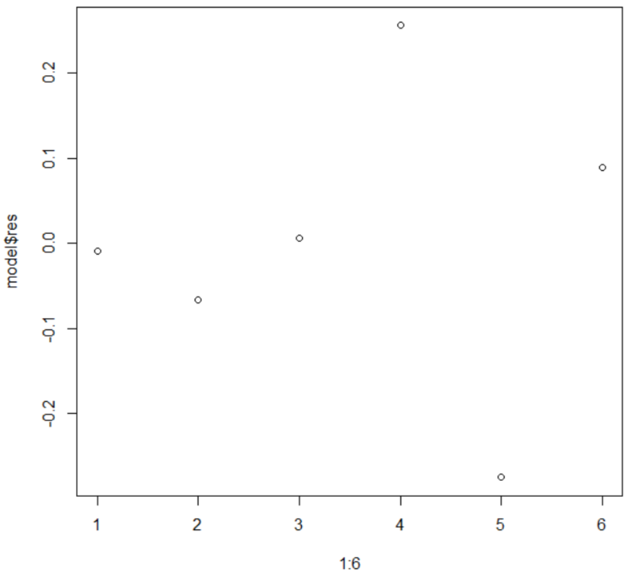


Από το παραπάνω γράφημα παρατηρούμε πως ισχύει η προϋπόθεση της γραμμικότητας.

Επιπλέον, με την εντολή:

> plot(1:6,model$res)

μπορούμε να εξετάσουμε την προϋπόθεση της ανεξαρτησίας των σφαλμάτων.



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

iii) Για την κατασκευή ενός 95% δ.ε. για τον συντελεστή β δίνουμε την εντολή:

```
> confint(model)
                 2.5 %    97.5 %
(Intercept)  0.7738880   1.534822
x           -0.2717697 -0.215571
```

και παρατηρούμε πως το 95% δ.ε. του β είναι [-0.2718,-0.2156]

Ερμηνεία του συντελεστή β: δείχνει τον ρυθμό με τον οποίο η συνάρτηση y πλησιάζει την ασυμπτωτική της τιμή (3) όσο το x αυξάνεται. Η αρνητική του τιμή υποδηλώνει ότι η προσέγγιση είναι εκθετική με μείωση της συμβολής του όρου ae^{β}

Από την εντολή summary(model) παρατηρούμε ότι β=-0.2437. Επομένως, $e^{\beta} \approx e^{-0.2437} \approx 0.7835$. Αυτό σημαίνει ότι για κάθε μονάδα αύξησης στο x, η συνεισφορά του όρου ae^{β} στο y μειώνεται περίπου στο 78.35% της προηγούμενης τιμής του.

iv) Για την σημειακή εκτίμηση της ενεργειακής κατανάλωσης όταν $x_0=9$ και για το 95% δ.ε. ατομικής πρόβλεψης για την αντίστοιχη παρατήρηση δίνουμε την εντολή:

```
> plogyp<-predict(model, list(x=c(9)),int="p")
> pyp<-3-exp(plogyp)
> pyp
       fit      lwr      upr
1 2.646078 2.803741 2.361758
```

Παρατηρούμε ότι y=2.6461 και το 95% δ.ε. ατομικής πρόβλεψης είναι [2.3618, 2.8037]

Αντιστρέφουμε το κάτω με το άνω όριο από αυτά που μας δίνει η R διότι εφαρμόσαμε τον μετασχηματισμό $3\text{-}e^{y}$

Για το 95% δ.ε. μέσης πρόβλεψης δίνουμε την εντολή:

```
> plogyc<-predict(model, list(x=c(9)),int="c")
> pyc<-3-exp(plogyc)
> pyc
       fit      lwr       upr
1 2.646078  2.718475  2.555063
```

Παρατηρούμε το 95% δ.ε. μέσης πρόβλεψης είναι [2.5551, 2.7185]

Αντιστρέφουμε και εδώ το κάτω με το άνω όριο από αυτά που μας δίνει η R διότι εφαρμόσαμε τον μετασχηματισμό $3\text{-}e^{y}$

Το διάστημα εμπιστοσύνης ατομικής πρόβλεψης είναι πιο ευρύ από το διάστημα εμπιστοσύνης μέσης πρόβλεψης.