ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

NATIONAL TECHNICAL UNIVERSITY OF ATHENS
SCHOOL OF APPLIED MATHEMATICAL AND PHYSICAL SCIENCES

Ανάλυση Παλινδρόμησης

Ονοματεπώνυμο φοιτητή: Γεώργιος Λεβής
Αριθμός μητρώου: ge19120
Έτος: 8ο
Εξάμηνο: 4ο
Email: ge19120@mail.ntua.gr / levgiorg@gmail.com

Άσκηση 1

Εισάγουμε τα δοθέντα στοιχεία στην R με την παρακάτω εντολή :

```
gifts <-
read.table("http://www.math.ntua.gr/~fouskakis/Data_Analysis/Exercises/gifts.txt", header = TRUE,na.strings="*")
```

Έτσι δηλώνουμε το path του αρχείου .txt, με το argument (header = TRUE) ενημερώνουμε ότι στην πρώτη γραμμή δηλώνονται ονόματα και με το argument (na.strings="*") δηλώνουμε ότι ο χαρακτήρας "*" είναι οι αγνοούμενες τιμές. Η R μας δίνει το πλαίσιο δεδομένων gifts το οποίο είναι ένα data frame. Συνεχίζουμε με τον υπόλοιπο κώδικα.

```
str(gifts)
gifts <- na.omit(gifts)
gifts <- subset(gifts, age >= 18)
```

Χρησιμοποιώ την εντολή (str(gift)) για να δω την δομή του 'gifts'. Έπειτα χρησιμοποιώ την εντολή (gifts <- na.omit(gifts)) για να αφαιρέσω οποιαδήποτε γραμμή περιέχει αγνοουμένη τιμή NA. Η εντολή (gifts <- subset(gifts, age >= 18)) με βοηθάει να αφαιρέσω οποιαδήποτε γραμμή που περιέχει τιμές από μη ενήλικα άτομα. Χρησιμοποιούμε την εντολή subset() γιατί μας επιτρέπει να πάρουμε ένα συγκεκριμένο subset από ένα data frame. Έτσι, στο πρώτο argument επιλέγουμε το data frame που θέλουμε να πάρουμε ('gifts') και στο δεύτερο argument γράφουμε τα conditions, στην περίπτωση μας είναι age >= 18 με το οποίο κρατάμε όλες τις σειρές που έχουν τιμή ίση ή πάνω από 18.

Αρχικά θα ορίσουμε τις 6 μεταβλητές από τις στήλες εξαρχής για δική μας ευκολία.

```
spend <- gifts$spend
age <- gifts$age
holiday <-gifts$holiday
sex <- gifts$sex
time <- gifts$time
salary <- gifts$salary
```

i)
Μας ζητείται να βρούμε τον συντελεστή συσχέτισης. Επειδή ανα δύο είναι συνεχείς τυχαίες μεταβλητές, μπορούμε να κάνουμε χρήση ενός Pearson correlation Coefficient test. Αντιπροσωπεύει την ισχύ μεταξύ των διαφορετικών μεταβλητών και των σχέσεων τους. Η τιμή μπορεί να κυμαίνεται από -1 έως 1, όπου το -1 υποδηλώνει ισχυρή αρνητική συσχέτιση, το 1 υποδηλώνει ισχυρή θετική συσχέτιση και το 0 υποδηλώνει μηδενική συσχέτιση. Θέτουμε ως μηδενική υπόθεση την H_0: p = 0 με εναλλακτική της την H_1: p≠0. Οπότε με τον παρακάτω κώδικα έχουμε:

Spend ~ Age

```
cor_spend_age <- cor(spend, age)
cat("Correlation between spend and age:", correlation, "\n")
```

με Output:

```
Correlation between spend and age: 0.2957294
```

Με τη εντολή «cor.test(spend, age)$estimate» αντλούμε μόνο τον συντελεστή συσχέτισης μεταξύ των δύο μεταβλητών spend & age. Το αποτέλεσμα που πήραμε δείχνει μια θετική συσχέτιση μεταξύ των δύο μεταβλητών, γεγονός που υποδηλώνει ότι όσο αυξάνεται η ηλικία, το ποσό που δαπανάται κατά τη διάρκεια της εορταστικής περιόδου τείνει επίσης να αυξάνεται.

Spend ~ Time

```
cor_spend_time <- cor(spend, time)
cat("Correlation between spend and time:", cor_spend_time, "\n")
```

με Output:

```
Correlation between spend and time: 0.3796202
```

Όμοια με παραπάνω βρίσκουμε ότι ο συντελεστής συσχέτισης μεταξύ spend & time είναι 0,3796202. Αυτό σημαίνει ότι, κατά μέσο όρο, καθώς αυξάνεται ο χρόνος που δαπανάται για τη λήψη διαφημιστικών προωθητικών ενεργειών, το ποσό που δαπανάται κατά τη διάρκεια της εορταστικής περιόδου τείνει επίσης να αυξάνεται.

Spend ~ Salary

```
cor_spend_salary <- cor(spend, salary)
cat("Correlation between spend and salary:", cor_spend_salary, "\n")
```

με Output:

```
Correlation between spend and salary: 0.7798325
```

Ο συντελεστής συσχέτισης 0,7798325 υποδηλώνει σημαντική θετική συσχέτιση μεταξύ των "δαπανών" και του "μισθού". Όσο υψηλότερος είναι ο συντελεστής συσχέτισης, τόσο ισχυρότερη είναι η θετική γραμμική σχέση μεταξύ των μεταβλητών.

Age ~ Time

```
cor_age_time <- cor(age, time)
cat("Correlation between age and time:", cor_age_time, "\n")
```

με Output:

Correlation between age and time: -0.08352963

Αντίστοιχα βρίσκουμε ότι αυτός ο συντελεστής συσχέτισης(-0.08352963) υποδηλώνει μια ασθενή αρνητική συσχέτιση μεταξύ των δύο μεταβλητών. Υποδηλώνει ότι υπάρχει μια ελαφρά τάση ο χρόνος που απαιτείται για να γίνει κάποιος αποδέκτης διαφημιστικών προωθητικών ενεργειών να μειώνεται ελαφρώς με την αύξηση της ηλικίας, αν και η συσχέτιση δεν είναι ιδιαίτερα ισχυρή.

Age ~ Salary

```
cor_age_salary <- cor(age, salary)
cat("Correlation between age and salary:", cor_age_salary, "\n")
```

με Output:

Correlation between age and salary: 0.0132712

Ο συντελεστής συσχέτισης 0,0132712 υποδηλώνει μια πολύ ασθενή θετική συσχέτιση μεταξύ της "ηλικίας" και του "μισθού". Το μικρό μέγεθος του συντελεστή συσχέτισης υποδηλώνει ότι δεν υπάρχει σχεδόν καμία συσχέτιση μεταξύ των μεταβλητών.

Time ~ Salary

```
cor_time_salary <- cor(time, salary)
cat("Correlation between time and salary:", cor_time_salary, "\n")
```

με Output:

Correlation between time and salary: 0.05841616

Όμοια ο συντελεστής συσχέτισης 0,05841616 υποδηλώνει μια πολύ ασθενή θετική συσχέτιση μεταξύ του "χρόνου" και του "μισθού". Το μικρό μέγεθος του συντελεστή συσχέτισης υποδηλώνει ότι δεν υπάρχει σχεδόν καμία συσχέτιση μεταξύ των μεταβλητών.

ii)

Για να προσαρμόσουμε ένα πολλαπλό γραμμικό μοντέλο με το "spend" ως μεταβλητή απόκρισης και το "age", το "holiday", το "sex", το "time" και το "salary" ως επεξηγηματικές μεταβλητές, θα χρησιμοποιήσουμε τη εντολή lm().

```
model <- lm(spend ~ age + holiday + sex + time + salary, data= gifts)
```

Με Output :

```
Call:
lm(formula = spend ~ age + holiday + sex + time + salary, data = gifts)

Coefficients:
  (Intercept)         age  holidayEaster   holidayOther       sexWoman           time         salary
     135.7154       2.0571       -39.8724       -80.9037         1.9825         2.5107         0.4918
```

Με βάση τα αποτελέσματα του μοντέλου πολλαπλής γραμμικής παλινδρόμησης τις μεταβλητές μας, ακολουθεί η ερμηνεία των εκτιμητών των συντελεστών:

Intercept: Η εκτιμώμενη τομή είναι 135,7154, που αντιπροσωπεύει την αναμενόμενη "δαπάνη" όταν όλες οι άλλες μεταβλητές είναι μηδενικές. Age: Μια αύξηση της ηλικίας κατά μία μονάδα συνδέεται με μια μέση αύξηση των "δαπανών" κατά 2,0571 μονάδες, διατηρώντας τις άλλες μεταβλητές σταθερές. HolidayEaster: Τα άτομα που πραγματοποίησαν αγορές κατά τη διάρκεια των διακοπών του Πάσχα τείνουν να ξοδεύουν, κατά μέσο όρο, 39,8724 μονάδες λιγότερες σε σύγκριση με άλλες περιόδους διακοπών, κρατώντας τις άλλες μεταβλητές σταθερές.
HolidayOther: Τα άτομα που πραγματοποίησαν αγορές κατά τη διάρκεια άλλων διακοπών εκτός των Χριστουγέννων ή του Πάσχα τείνουν να ξοδεύουν, κατά μέσο όρο, 80,9037 μονάδες λιγότερες σε σύγκριση με τα άτομα που πραγματοποίησαν αγορές κατά τη διάρκεια των Χριστουγέννων, κρατώντας τις άλλες μεταβλητές σταθερές.

SexWoman: Οι γυναίκες τείνουν να ξοδεύουν, κατά μέσο όρο, 1,9825 μονάδες περισσότερο από τους άνδρες, κρατώντας τις άλλες μεταβλητές σταθερές. Time: Μια αύξηση κατά μία μονάδα του χρόνου που απαιτείται για να γίνει κάποιος αποδέκτης διαφημιστικών προωθητικών ενεργειών συνδέεται με μια μέση αύξηση της "δαπάνης" κατά 2,5107 μονάδες, κρατώντας τις άλλες μεταβλητές σταθερές. Salary: Μια αύξηση του μισθού κατά μία μονάδα συνδέεται με μια μέση αύξηση της "δαπάνης" κατά 0,4918 μονάδες, κρατώντας τις άλλες μεταβλητές σταθερές.

Τα 95% διαστήματα εμπιστοσύνης για τους συντελεστές του εν λόγω μοντέλου θα δημιουργηθούν με το παρακάτω κώδικα.

```r
conf_intervals <- confint(model, level = 0.95)
print(conf_intervals)
```

Με Output :

```
                   2.5 %     97.5 %
(Intercept)   126.0543489 145.376465
age             1.9539036   2.160288
holidayEaster -42.4346892 -37.310205
holidayOther  -83.7153014 -78.092000
```

```
sexWoman      -0.2249080   4.189860
time          2.4177344    2.603690
salary        0.4822309    0.501431
```

Αυτά τα διαστήματα εμπιστοσύνης παρέχουν ένα εύρος πιθανών τιμών για τους συντελεστές, αντανακλώντας την αβεβαιότητα που συνδέεται με την εκτίμησή τους από τα δεδομένα του δείγματος.

iii)

```
summary(model)
```

Με Output :

```
Call:
lm(formula = spend ~ age + holiday + sex + time + salary, data = gifts)

Residuals:
    Min      1Q  Median      3Q     Max
-6.5651 -2.2959 -0.1514  1.7858  7.9214

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   135.715407   4.793695   28.31   <2e-16 ***
age           2.057096   0.051203   40.18   <2e-16 ***
holidayEaster -39.872447   1.271352  -31.36   <2e-16 ***
holidayOther  -80.903651   1.395106  -57.99   <2e-16 ***
sexWoman       1.982476   1.095276    1.81   0.0771 .
time          2.510712   0.046134   54.42   <2e-16 ***
salary        0.491831   0.004763  103.25   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.662 on 44 degrees of freedom
Multiple R-squared:  0.9978,    Adjusted R-squared:  0.9975
```

F-statistic: 3318 on 6 and 44 DF, p-value: < 2.2e-16

Από το summary του μοντέλου πολλαπλής γραμμικής παλινδρόμησης, ακολουθεί η εξήγηση των αποτελεσμάτων:

- Residuals: τεταρτημόριο, διάμεσος, τρίτο τεταρτημόριο και μέγιστη τιμή των υπολοίπων. Τα residuals αντιπροσωπεύουν τις διαφορές μεταξύ των παρατηρούμενων τιμών και των προβλεπόμενων τιμών της μεταβλητής απόκρισης.

- Coefficients: Αυτή η ενότητα εμφανίζει τις εκτιμήσεις των συντελεστών για κάθε μεταβλητή του μοντέλου, μαζί με τα τυπικά σφάλματα, τις τιμές t και τις τιμές p.

- Intercept: Η εκτιμώμενη διατομή είναι 135,715407 και είναι στατιστικά σημαντική με πολύ χαμηλή τιμή p-value (< 2e-16). Αυτό δείχνει ότι υπάρχει ένας σημαντικός σταθερός όρος στο μοντέλο.

- Age: Η εκτίμηση του συντελεστή για την "ηλικία" είναι 2,057096 και είναι στατιστικά σημαντική με πολύ χαμηλή p-τιμή (< 2e-16). Αυτό υποδηλώνει ότι η ηλικία έχει σημαντική επίδραση στη "δαπάνη" όταν ελέγχεται για άλλες μεταβλητές στο μοντέλο.

- holidayEaster: Η εκτίμηση του συντελεστή για το "holidayEaster" είναι -39,872447 και είναι στατιστικά σημαντική με πολύ χαμηλή p-value (< 2e-16). Αυτό δείχνει ότι τα άτομα που πραγματοποίησαν αγορές κατά τη διάρκεια των διακοπών του Πάσχα τείνουν να ξοδεύουν σημαντικά λιγότερα σε σύγκριση με άλλες περιόδους διακοπών, ενώ οι άλλες μεταβλητές παραμένουν σταθερές.

- holidayOther: Η εκτίμηση του συντελεστή για την "holidayOther" είναι -80,903651 και είναι στατιστικά σημαντική με πολύ χαμηλή p-value (< 2e-16). Αυτό υποδηλώνει ότι τα άτομα που πραγματοποίησαν αγορές κατά τη διάρκεια άλλων εορτών εκτός των Χριστουγέννων ή του Πάσχα τείνουν να ξοδεύουν σημαντικά λιγότερα σε σύγκριση με εκείνα που πραγματοποίησαν αγορές κατά τη διάρκεια των Χριστουγέννων, κρατώντας τις άλλες μεταβλητές σταθερές.

- sexWoman: Η εκτίμηση του συντελεστή για το "sexWoman" είναι 1,982476 και έχει p-value 0,0771. Η τιμή p-value είναι σχετικά υψηλή (μεγαλύτερη από 0,05), υποδεικνύοντας ότι η επίδραση του φύλου (αν είσαι γυναίκα) στις "δαπάνες" μπορεί να μην είναι στατιστικά σημαντική σε επίπεδο σημαντικότητας 0,05.

- Time: Η εκτίμηση του συντελεστή για τον "χρόνο" είναι 2,510712 και είναι στατιστικά σημαντική με πολύ χαμηλή τιμή p-value (< 2e-16). Αυτό υποδηλώνει ότι η αύξηση του χρόνου που απαιτείται για να γίνει κάποιος αποδέκτης διαφημιστικών προωθητικών ενεργειών συνδέεται με σημαντική αύξηση της "δαπάνης", ενώ ελέγχεται για άλλες μεταβλητές στο μοντέλο.

- Salary: Η εκτίμηση του συντελεστή για τον "μισθό" είναι 0,491831, και είναι στατιστικά σημαντική με πολύ χαμηλή τιμή p-value (< 2e-16). Αυτό δείχνει ότι η αύξηση του μισθού συνδέεται με σημαντική αύξηση της "δαπάνης", ενώ οι άλλες μεταβλητές παραμένουν σταθερές.

Significance codes: Οι κωδικοί σημαντικότητας (0 '', 0,001 '', 0,01 '', 0,05 '.', 0,1 '', 1) παρέχουν μια γρήγορη ένδειξη της στατιστικής σημαντικότητας των συντελεστών. Για

παράδειγμα, το "***" υποδηλώνει πολύ χαμηλή τιμή p-value (< 0,001), υποδηλώνοντας υψηλή στατιστική σημαντικότητα, ενώ το "." υποδηλώνει τιμή p-value μεγαλύτερη από 0,05, υποδηλώνοντας ότι ο συντελεστής μπορεί να μην είναι στατιστικά σημαντικός.

Residual standard error: Το τυπικό σφάλμα υπολοίπου (3,662) αντιπροσωπεύει την εκτίμηση της τυπικής απόκλισης των υπολοίπων. Δείχνει τη μέση απόκλιση των παρατηρούμενων τιμών "δαπάνης" από τις προβλεπόμενες τιμές.

Multiple R-squared and Adjusted R-squared: Η τιμή του multiple R-squared (0,9978) δείχνει ότι το μοντέλο εξηγεί περίπου το 99,78% της διακύμανσης των "δαπανών". Η προσαρμοσμένη τιμή R-squared (0,9975) λαμβάνει υπόψη τον αριθμό των προβλεπτικών παραγόντων και παρέχει ένα ελαφρώς προσαρμοσμένο μέτρο της καλής προσαρμογής του μοντέλου.

F-statistic και p-value: Το F-statistic (3318) ελέγχει τη συνολική σημαντικότητα του μοντέλου συγκρίνοντας τη διακύμανση που εξηγείται από το μοντέλο με την ανεξήγητη διακύμανση. Η τιμή p-value (< 2,2e-16) δείχνει ότι το συνολικό μοντέλο είναι εξαιρετικά σημαντικό.

iv)

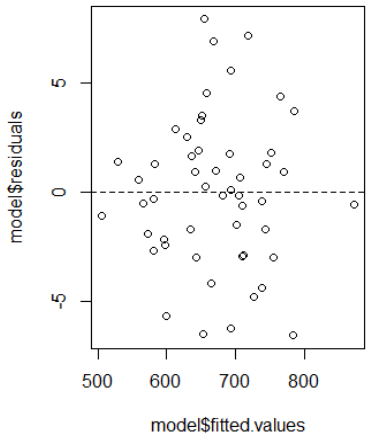
Μπορούμε να το ελέγξουμε από τα παρακάτω μέσω γραφημάτων:

Γραμμικότητα, Ανεξαρτησία (τα residuals είναι ανεξάρτητα), Ομοσκεδαστικότητα(η διακύμανση των residuals είναι σταθερή), Κανονικότητα(τα residuals κατανέμονται κανονικά).

Ελέγχουμε τα υπόλοιπα με τα fitted values (Ομοσκεδαστικότητα)

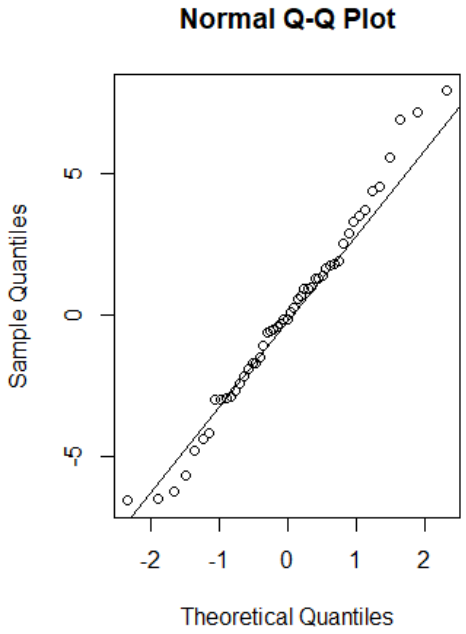
Για να ελέγξουμε την ομοσκεδαστικότητα πρέπει να κάνουμε τη γραφική παράσταση των υπολοίπων συναρτήσει των προβλεπόμενων τιμών, fitted values, και να μην παρατηρηθεί κάποιο μοτίβο στα σημεία που σχηματίζονται.

```
plot(model$residuals ~ model$fitted.values)
abline(h = 0, lty = 2)
```



ΚΑΝΟΝΙΚΟΤΗΤΑ ΣΦΑΛΜΑΤΩΝ, πρέπει τα υπόλοιπα, residuals, να ακολουθούν κανονική κατανομή. Αυτό ελέγχεται με τις εντολές:

```
qqnorm(model$residuals)
qqline(model$residuals)
```



Όπως παρατηρούμε η κατανομή των υπολοίπων συμφωνεί με την κανονική κατανομή και δεν υπάρχουν σημεία που να μας κάνουν να αμφισβητήσουμε την κανονικότητα των υπολοίπων.

Δυστυχώς λόγω χώρου δεν μπορώ να κατασκευάσω όλες τις γραφικές παραστάσεις και να γινώ όσο περιγραφικός θέλω.

v)

Για να το πετύχουμε αυτό θα φτιάξουμε ένα νέο data.frame με τα στοιχεία που μας δίνονται.

```
newdata <- data.frame(
  age = 25,
  holiday = "Easter",
  sex = "Woman",
  time = 13,
  salary = 575
)
```

Θα χρησιμοποιησουμε την συνάρτηση predict, που χρησιμοποιεί τους συντελεστές από το μοντέλο παλινδρόμησής για να εκτιμήσει τις δαπάνες με βάση τα νέα δεδομένα που

παρέχονται. Το όρισμα interval καθορίζει ότι θέλουμε ένα διάστημα εμπιστοσύνης και το όρισμα level καθορίζει το επίπεδο εμπιστοσύνης 90%.

```
estimate <- predict(model, newdata = newdata, interval = "confidence",
level = 0.90)
print(estimate)
```

Με Output :

```
     fit      lwr      upr
1 464.6949 461.4693 467.9205
```

Αυτό το αποτέλεσμα δείχνει ότι η σημειακή εκτίμηση της αναμενόμενης δαπάνης για μια γυναίκα 25 ετών, η οποία παρακολουθεί 13 λεπτά διαφημιστικών μηνυμάτων την ημέρα και αμείβεται με 575 ευρώ το μήνα κατά τη διάρκεια των διακοπών του Πάσχα, είναι περίπου 464,69 ευρώ. Το 90% διάστημα εμπιστοσύνης για την εκτίμηση αυτή κυμαίνεται από περίπου 461,47 ευρώ έως 467,92 ευρώ. Αυτό σημαίνει ότι μπορούμε να είμαστε 90% σίγουροι ότι η πραγματική αναμενόμενη δαπάνη για ένα άτομο με αυτά τα χαρακτηριστικά βρίσκεται εντός αυτού του διαστήματος.

vii)

Για να προσαρμόσουμε ένα πολλαπλασιαστικό μοντέλο, μπορούμε να πάρουμε το φυσικό λογάριθμο της μεταβλητής απόκρισης (spend) και των επεξηγηματικών μεταβλητών (age, time και salary- αυτές είναι οι ποσοτικές μεταβλητές στο σύνολο δεδομένων μας). Οι κατηγορικές μεταβλητές (holiday και sex) θα παραμείνουν αμετάβλητες. Επίσης θα προσθέσω τις μετασχηματισμένες μεταβλητές πίσω στο αρχικό dataframe.Παρακάτω δίνω τον κώδικα.

```
log_spend <- log(gifts$spend)
log_age <- log(gifts$age)
log_time <- log(gifts$time)
log_salary <- log(gifts$salary)

gifts$log_spend <- log(gifts$spend)
gifts$log_age <- log(gifts$age)
gifts$log_time <- log(gifts$time)
gifts$log_salary <- log(gifts$salary)

gifts$holiday <- as.factor(gifts$holiday)
gifts$sex <- as.factor(gifts$sex)

log_model <- lm(log_spend ~ log_age + holiday + sex + log_time +
log_salary, data = gifts)
```

```
summary(log_model)
```

με output :

```
Call:
lm(formula = log_spend ~ log_age + holiday + sex + log_time +
    log_salary, data = gifts)

Residuals:
      Min        1Q    Median        3Q       Max
-0.0228267 -0.0064113 -0.0008046  0.0084871  0.0230075

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    1.497466   0.085725  17.468   <2e-16 ***
log_age        0.107531   0.005341  20.132   <2e-16 ***
holidayEaster -0.063940   0.003862 -16.557   <2e-16 ***
holidayOther  -0.133901   0.004230 -31.652   <2e-16 ***
sexWoman       0.002693   0.003316   0.812    0.421
log_time       0.119935   0.004346  27.595   <2e-16 ***
log_salary     0.635040   0.012261  51.792   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.01109 on 44 degrees of freedom
Multiple R-squared:  0.9911,        Adjusted R-squared:  0.9899
F-statistic: 818.1 on 6 and 44 DF,  p-value: < 2.2e-16
```

log_age: Μια αύξηση 1% στην ηλικία προβλέπει αύξηση 0,107531% στις δαπάνες, όλα τα άλλα ίσα.

holidayEaster : Οι δαπάνες κατά τη διάρκεια του Πάσχα προβλέπεται να είναι 6,394% λιγότερες σε σχέση με τα Χριστούγεννα, διατηρώντας σταθερούς όλους τους άλλους παράγοντες.

holidayOther: Οι δαπάνες κατά τη διάρκεια άλλων εορτών προβλέπεται να είναι 13,3901% λιγότερες σε σχέση με τα Χριστούγεννα, όλοι οι άλλοι παράγοντες σταθεροί.

sexWoman: Οι δαπάνες των γυναικών προβλέπεται να είναι 0,2693% υψηλότερες από τις δαπάνες των ανδρών, όλα τα άλλα σταθερά, αλλά αυτή η επίδραση δεν είναι στατιστικά σημαντική.

log_time: Μια αύξηση 1% στον χρόνο παρακολούθησης διαφημίσεων προβλέπει αύξηση 0,119935% στις δαπάνες, ενώ όλες οι άλλες μεταβλητές παραμένουν σταθερές.

log_salary: Μια αύξηση 1% στον μισθό προβλέπει αύξηση 0,63504% στις δαπάνες, όλα τα άλλα ίσα. Ο μισθός φαίνεται να έχει τον ισχυρότερο αντίκτυπο στις δαπάνες σε αυτό το μοντέλο.

Αυτές οι ερμηνείες αντικατοπτρίζουν πώς η ηλικία, ο χρόνος που αφιερώνεται στην παρακολούθηση διαφημίσεων, ο μισθός και ο περίοδος διακοπών μπορούν να επηρεάσουν τις δαπάνες κατά τη διάρκεια των εορταστικών περιόδων.

viii)

Για να ελέγξουμε τις συνθήκες του πολλαπλασιαστικού μοντέλου, συνήθως εξετάζουμε διάφορα διαγνωστικά διαγράμματα:

Residuals vs Fitted values: Αυτό το διάγραμμα χρησιμοποιείται για τον έλεγχο της παραδοχής της γραμμικότητας και της ίσης διακύμανσης (ομοσκεδαστικότητα). Ιδανικά, θα θέλαμε να βλέπουμε μια οριζόντια γραμμή σημείων, τυχαία διασκορπισμένων, χωρίς μοτίβα ή τάσεις.

Normal Q-Q plot: Αυτό το διάγραμμα χρησιμοποιείται για τον έλεγχο της υπόθεσης κανονικότητας των καταλοίπων. Εάν τα σημεία πέφτουν περίπου κατά μήκος της γραμμής, αυτό υποδηλώνει ότι τα υπολοιπα είναι κανονικά κατανεμημένα.

Scale-Location plot: Αυτό είναι ένα άλλο διάγραμμα για τον έλεγχο της παραδοχής της ομοσκεδαστικότητας. Παρόμοια με το διάγραμμα Residuals vs Fitted plot, θα θέλαμε να δούμε μια οριζόντια γραμμή με ισόποσα κατανεμημένα σημεία.

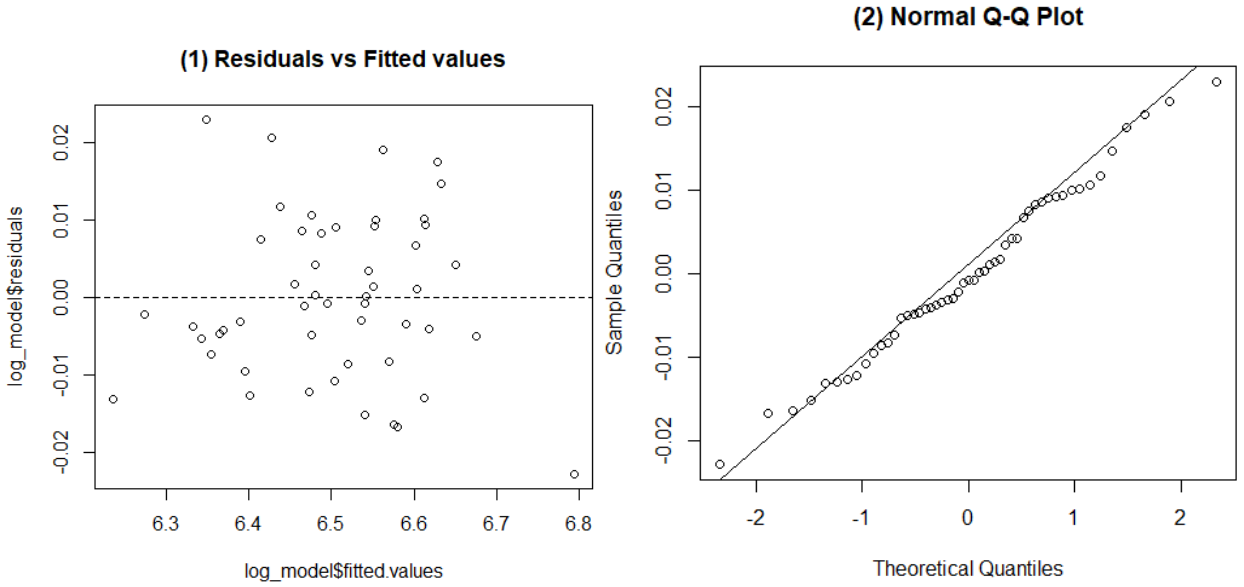
Residuals vs Leverage: Αυτό το διάγραμμα μας βοηθά να εντοπίσουμε τις περιπτώσεις με επιρροή, εάν υπάρχουν.

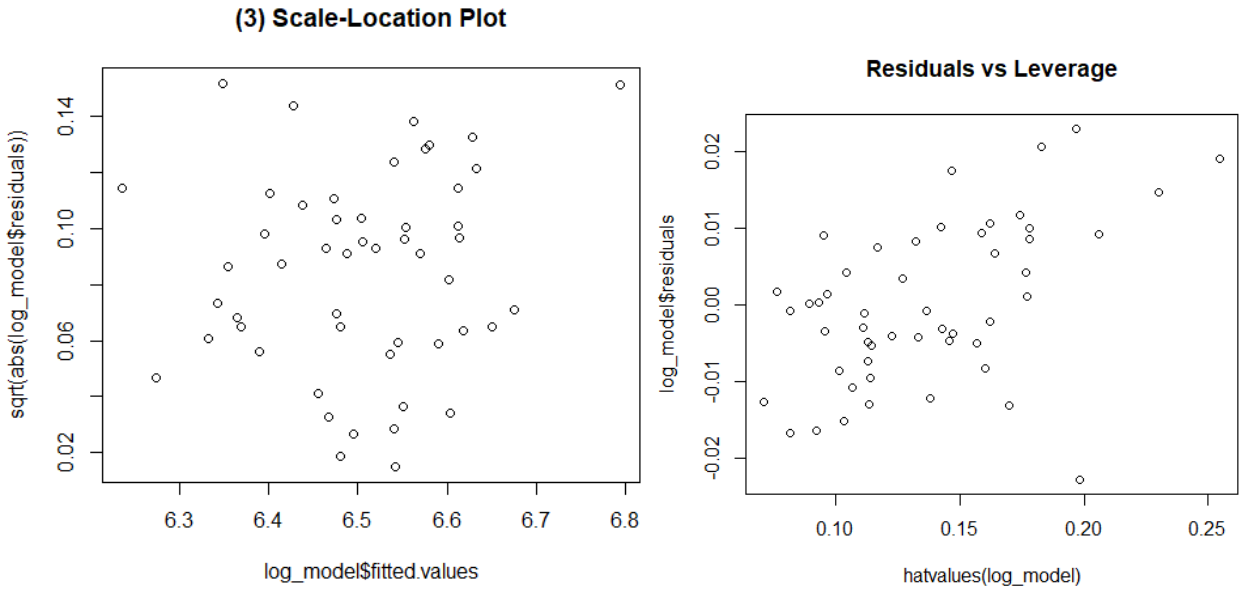
```
plot(log_model$fitted.values, log_model$residuals, main = "(1) Residuals
vs Fitted values")
abline(h = 0, lty = 2)

qqnorm(log_model$residuals, main = "(2) Normal Q-Q Plot")
qqline(log_model$residuals)


plot(log_model$fitted.values, sqrt(abs(log_model$residuals)), main =
"(3) Scale-Location Plot")

plot(hatvalues(log_model), log_model$residuals, main = "Residuals vs
Leverage")
```





Στο πρώτο διάγραμμα βλέπουμε μια οριζόντια γραμμή σημείων, τυχαία διασκορπισμένων, χωρίς μοτίβα ή τάσεις. Οπότε αυτό σημαίνει ότι οι παράγοντες που περιλαμβάνονται στο μοντέλο (age, holiday, sex, time, salay) είναι κατάλληλοι για την πρόβλεψη των δαπανών (spend).

Στο δεύτερο διάγραμμα βλέπουμε ότι τα σημεία πέφτουν περίπου κατά μήκος της γραμμής, αυτό υποδηλώνει ότι τα υπόλοιπα είναι κανονικά κατανεμημένα. Αυτό υποδηλώνει ότι οι προβλέψεις του μοντέλου είναι αξιόπιστες και τα συμπεράσματα που βγάζουμε από αυτό το μοντέλο είναι έγκυρα.

Στο τρίτο διάγραμμα δεν παρατηρείται κάποιος συστηματικός τρόπος συμπεριφοράς των ζευγών που εξετάσαμε. Επομένως, δεν μπορούμε να απορρίψουμε ούτε την

ομοσκεδαστικότητα. Προχωράμε, λοιπόν, στον τελευταίο μας έλεγχο, δηλαδή στον έλεγχο για ανεξαρτησία των σφαλμάτων. Για τον έλεγχο αυτό, κατασκευάζουμε ένα διάγραμμα υπολοίπων σε σχέση με τη σειρά των δεδομένων.
Στο τέταρτο διάγραμμα δεν παρατηρούμε περιπτώσεις με επιρροή. Αυτό είναι ένα καλό σημάδι, καθώς δείχνει ότι οι προβλέψεις του μοντέλου δεν επηρεάζονται υπερβολικά από μερικές ακραίες τιμές, αλλά βασίζονται στο συνολικό μοτίβο των δεδομένων.

ix)

Για να υπολογίσουμε το ποσό σε ευρώ που αναμένεται να δαπανήσει κατά τις διακοπές του Πάσχα μια 25χρονη που παρακολουθεί 13 λεπτά διαφημιστικά μηνύματα την ημέρα και αμείβεται με 575 ευρώ το μήνα, θα χρησιμοποιήσουμε την ίδια λογική, απλά η R θα μας δώσει σαν αποτέλεσμα την λογαριθμισμένη τιμή του ποσού, οπότε χρειάζεται να την υψώσουμε σε εκθετικό.

```
newdata1 <- data.frame(
  log_age = log(25),
  holiday = factor("Easter", levels = levels(gifts$holiday)),
  sex = factor("Woman", levels = levels(gifts$sex)),
  log_time = log(13),
  log_salary = log(575)
)
estimate <- exp(predict(log_model, newdata = newdata1, interval =
"confidence", level = 0.90))
estimate
```

με output :

```
     fit     lwr     upr
1 457.2593 451.9755 462.6049
```

Οπότε, το εκτιμώμενο ποσό που θα ξοδέψει μια 25χρονη στις γιορτές του Πάσχα, που παρακολουθεί 13 λεπτά διαφημιστικά μηνύματα την ημέρα και αμείβεται με 575 ευρώ τον μήνα, είναι περίπου 457,26 ευρώ. Το διάστημα εμπιστοσύνης 90% για αυτήν την εκτίμηση κυμαίνεται από περίπου 451,98 ευρώ έως 462,60 ευρώ. Αυτό σημαίνει ότι είμαστε 90% σίγουροι ότι οι πραγματικές δαπάνες θα εμπίπτουν σε αυτό το εύρος.

x)

Αρχικά θα δημιουργήσουμε μια νεα μεταβλητή timeq στο dataframe μας και μετα θα προσαρμόσουμε το πολλαπλό γραμμικό μοντέλο.

```
gifts$timeq <- gifts$time^2
model_sq <- lm(spend ~ age + holiday + sex + time + salary + timeq,
data = gifts)
summary(model_sq)$coefficients
```

(Intercept): Η βασική δαπάνη αναμένεται να είναι περίπου 128,61 ευρώ όταν όλες οι άλλες μεταβλητές (age, holiday, sex, time, salary και $time^2$) είναι 0. Στην πραγματικότητα, αυτό

δεν έχει ουσιαστική ερμηνεία, καθώς οι τιμές των επεξηγηματικών μεταβλητών δεν μπορούν να είναι όλες 0.
Age: Για κάθε επιπλέον έτος ηλικίας, οι δαπάνες αναμένεται να αυξηθούν κατά περίπου 2,05 ευρώ, με την προϋπόθεση ότι όλες οι άλλες μεταβλητές παραμένουν σταθερές.

holidayEaster: Η δαπάνη αναμένεται να μειωθεί κατά περίπου 40,24 ευρώ εάν η περίοδος των εορτών είναι το Πάσχα, σε σύγκριση με τα Χριστούγεννα, υποθέτοντας ότι όλες οι άλλες μεταβλητές παραμένουν σταθερές.

holidayOther: Η δαπάνη αναμένεται να μειωθεί κατά περίπου 81,40 ευρώ εάν η περίοδος των εορτών χαρακτηρίζεται ως "Άλλη", σε σύγκριση με τα Χριστούγεννα (επίπεδο αναφοράς), υποθέτοντας ότι όλες οι άλλες μεταβλητές παραμένουν σταθερές.

sexWoman: Η δαπάνη αναμένεται να αυξηθεί κατά περίπου 1,99 ευρώ εάν το άτομο είναι γυναίκα, σε σύγκριση με το εάν το άτομο είναι άνδρας (επίπεδο αναφοράς), με την προϋπόθεση ότι όλες οι άλλες μεταβλητές παραμένουν σταθερές. Αυτό δεν είναι στατιστικά σημαντικό σε επίπεδο 5%, δεδομένου ότι η τιμή p-value είναι μεγαλύτερη από 0,05.

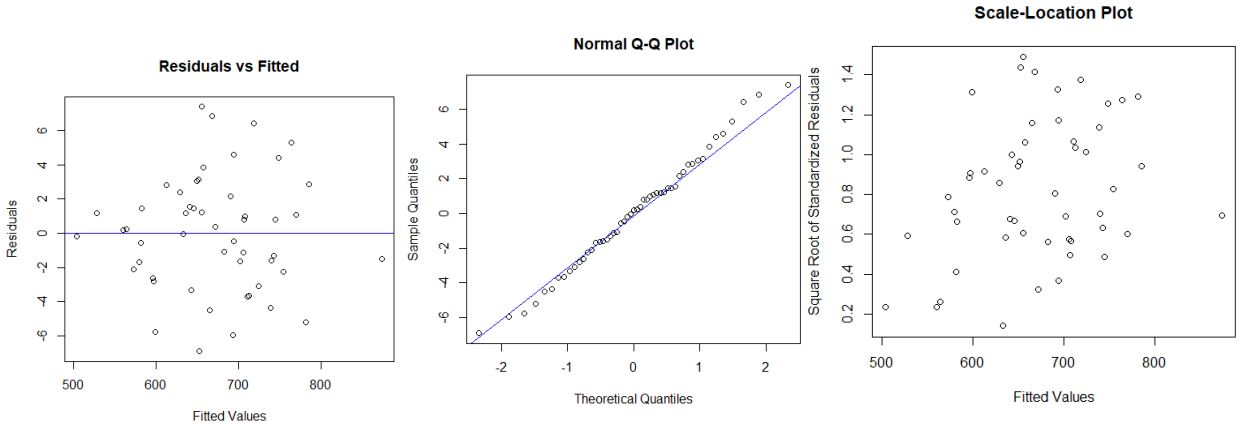
time: Για κάθε επιπλέον λεπτό παρακολούθησης διαφήμισης ανά ημέρα, η δαπάνη αναμένεται να αυξηθεί κατά περίπου 2,90 ευρώ, με την προϋπόθεση ότι όλες οι άλλες μεταβλητές παραμένουν σταθερές.

salary: Για κάθε πρόσθετο ευρώ μισθού, η δαπάνη αναμένεται να αυξηθεί κατά περίπου 0,49 ευρώ, υποθέτοντας ότι όλες οι άλλες μεταβλητές παραμένουν σταθερές.

Timeq: Ο αρνητικός συντελεστής υποδηλώνει ότι υπάρχει φθίνουσα απόδοση στη δαπάνη καθώς αυξάνεται ο χρόνος παρακολούθησης διαφημιστικών μηνυμάτων. Για κάθε πρόσθετο λεπτό, η δαπάνη αναμένεται να μειωθεί κατά περίπου 0,006 φορές το τετράγωνο του χρόνου σε λεπτά, με την προϋπόθεση ότι όλες οι άλλες μεταβλητές παραμένουν σταθερές. Το αποτέλεσμα αυτό δεν είναι στατιστικά σημαντικό σε επίπεδο 5%, δεδομένου ότι η τιμή p είναι μεγαλύτερη από 0,05. Αυτό υποδηλώνει ότι ο τετραγωνικός όρος μπορεί να μην είναι απαραίτητος στο μοντέλο.

xi)

Η διαδικασία έχει γίνει άλλες 2 φορές οπότε θα παραθέσω τα γραφήματα. Δεν υπάρχει χορός δυστυχώς να γινώ όσο περιγραφικός θα ήθελα.



xii)

Θα χρησιμοποιήσουμε την ίδια λογική απλά στο νέο data.frame μας θα προσθέσουμε και την timeq.

```
newdata2 <- data.frame(
  age = 25,
  holiday = factor("Easter", levels = levels(gifts$holiday)),
  sex = factor("Woman", levels = levels(gifts$sex)),
  time = 13,
  salary = 575,
  timesq = 13^2
)
estimate2 <- predict(model_sq, newdata = newdata2, interval =
"confidence", level = 0.90)
estimate2
```

με Output:

```
     fit     lwr     upr
1 462.2328 458.1405 466.325
```

Σύμφωνα με το μοντέλο μας, μια 25χρονη γυναίκα που παρακολουθεί 13 λεπτά διαφημίσεις την ημέρα και κερδίζει 575 ευρώ το μήνα αναμένεται να ξοδέψει περίπου 462,23 ευρώ κατά τις διακοπές του Πάσχα. Το διάστημα εμπιστοσύνης 90% για αυτή την πρόβλεψη κυμαίνεται από περίπου 458,14 ευρώ έως 466,33 ευρώ. Αυτό σημαίνει ότι, σύμφωνα με το μοντέλο σας και δεδομένων των δεδομένων που έχετε συλλέξει, μπορείτε να είστε 90% σίγουροι ότι το πραγματικό ποσό δαπανών θα εμπίπτει σε αυτό το εύρος.

xiii)

Η πρόβλεψη που υπολογίζει ότι για μια 25χρονη γυναίκα που παρακολουθεί 13 λεπτά διαφημίσεις την ημέρα και κερδίζει 575 ευρώ το μήνα αναμένεται να ξοδέψει περίπου 462,23 ευρώ κατά τις διακοπές του Πάσχα. Αυτή η πρόβλεψη έρχεται με ένα διάστημα εμπιστοσύνης 90%, που κυμαίνεται από περίπου 458,14 € έως 466,33 €. Τα residuals του μοντέλου κατανέμονται κανονικά, κάτι που είναι καλό σημάδι. Αυτό σημαίνει ότι οι προβλέψεις του μοντέλου είναι πιθανό να είναι ακριβείς. Οι παράγοντες που περιλαμβάνονται στο μοντέλο (age, holiday, sex, time, salary) κρίνονται κατάλληλοι για την πρόβλεψη δαπανών. Αυτοί οι παράγοντες είναι πιθανό να έχουν σημαντικό αντίκτυπο στη συμπεριφορά των δαπανών, καθιστώντας τους σχετικούς για την πρόβλεψη. Ο τετραγωνικός όρος (timeq) στο μοντέλο δεν είναι στατιστικά σημαντικός στο επίπεδο του 5%. Αυτό υποδηλώνει ότι οι προβλέψεις του μοντέλου μπορεί να είναι πιο ακριβείς χωρίς αυτόν τον όρο. Η πρόβλεψη που επέλεξα δεν βασίζεται σε μεγάλο βαθμό σε αυτόν τον όρο.Η πρόβλεψη συνοδεύεται από ένα διάστημα εμπιστοσύνης 90%, που κυμαίνεται από περίπου 458,14 € έως 466,33 €. Αυτό σημαίνει ότι μπορούμε να είμαστε 90% σίγουροι ότι οι πραγματικές δαπάνες θα εμπίπτουν σε αυτό το εύρος. Αυτό μας δίνει ένα μέτρο της αξιοπιστίας της πρόβλεψης. Δεδομένων αυτών των παραγόντων, αν έπρεπε να διαλέξω, θα εμπιστευόμουν αυτήν την πρόβλεψη. Ωστόσο, είναι σημαντικό να θυμάστε ότι όλα τα μοντέλα έχουν περιορισμούς και υποθέσεις και η πραγματική ακρίβεια αυτής της πρόβλεψης θα εξαρτηθεί από διάφορους παράγοντες που δεν περιλαμβάνονται στη λειτουργία