Άσκηση 1:

Στον πίνακα παρακάτω θα βρείτε τον αριθμό αυτοκτονιών στις ΗΠΑ μεταξύ 1968 και 1970.

Έτος	Ιαν	Φεβ	Μαρ	Απρ	Μαι	Ιουν
1968	1720	1712	1924	1882	1870	1680
1969	1831	1609	1973	1944	2003	1774
1970	1867	1789	1944	2094	2097	1981
	Ιουλ	Αυγ	Σεπτ	Οκτ	Νοεμ	Δεκ
1968	1868	1801	1756	1760	1666	1733
1969	1811	1873	1862	1897	1866	1921
1970	1887	2024	1928	2032	1978	1859

Απαντήστε στα παρακάτω ερωτήματα αφού λάβετε υπ'όψην ότι κάθε μήνας έχει διαφορετικό αριθμο ημερών (30 ή 31) και ο Φεβρουάριος έχει 28 ημέρες. Μας ενδιαφέρει ο αναμενόμενος αριθμός αυτοκτονιών. Για τα ερωτήματα $(A.) - (E.)$ θεωρείστε ότι ο αναμενόμενος αριθμός αυτοκτονιών είναι ανάλογος του αριθμού των ημερών του μήνα παρατήρησης.

Α. Υπάρχουν διαφορές μεταξύ των τριών ετών στον αναμενόμενο αριθμό αυτοκτονιών, αγνοώντας την επίδραση των μηνών;

Β. Υπάρχουν διαφορές μεταξύ των τεσσάρων εποχών θεωρώντας ότι οι διαφορές αυτές είναι ίδιες για όλα τα έτη;

Γ. Οι διαφορές των τεσσάρων εποχών φαίνεται να είναι διαφορετικές μεταξύ των τριών ετών;

Δ. Για το μοντέλο του $(B.)$ εκτιμήστε το λόγο του ρυθμού αυτοκτονιών μεταξύ Άνοιξης και καλοκαιριού για το 1969.

Ε. Για το μοντέλο του $(B.)$ ελέγξτε αν δικαιολογείται η υπόθεση ο αναμενόμενος αριθμός αυτοκτονιών είναι ανάλογος του αριθμού των ημερών του μήνα παρατήρησης. Περιγράψτε τι σημαίνει η μη ισχύ της υπόθεσης αυτής για το πρόβλημα που εξετάζουμε.

ΛΥΣΗ:

Το μοντέλο αυτό περιλαμβάνει μετρήσεις, οπότε για να το λύσουμε θα σκεφτόμασταν αρχικά να χρησιμοποιήσουμε την κατανομή *Poisson*. Όμως, εαν δούμε την καλή προσαρμογή του μοντέλου, βλέπουμε ότι δεν είναι καλό αυτό το μοντέλο. Αφού εισάγουμε λοιπόν τα δεδομένα, θα ελέγξουμε αν οι παρατηρήσεις μας προέρχονται από κανονική κατανομή.

Παρατηρούμε ότι οι μετρήσεις έγιναν σε διαφορετικό αριθμό ημερών οπότε μέσα στην παλιδρόμισή μας πρέπει να υπάρχει ένας αντισταθμιστής (*offset*) για να βγάλουν νόημα οι εκτιμήσεις μας.

Ξεκινώντας απο τον Ιανουάριο του 1968, πάμε λοιπόν πρώτα να εισάγουμε τα δεδομένα μας στην *R*. Ο αριθμός αυτοκτονιών θα είναι:

```
n.suicides <- c(1720, 1712, 1924, 1882, 1870, 1680, 1868, 1801, 1756, 1760, 1666,
1733, 1831, 1609, 1973, 1944, 2003, 1774, 1811, 1873, 1862, 1897, 1866, 1921, 1867,
1789, 1944, 2094, 2097, 1981, 1887, 2024, 1928, 2032, 1978, 1859)
```

Ενώ το *offset* θα παίρνει τις εξής τιμές:

```
mod.offset <- rep(c(31, 30), times = 18)

mod.offset[2] <- 28

mod.offset[14] <- 28

mod.offset[26] <- 28

mod.offset

 [1] 31 28 31 30 31 30 31 30 31 30 31 30 31 28 31 30 31 30 31

[20] 30 31 30 31 30 31 28 31 30 31 30 31 30 31 30 31 30
```

Για να δούμε αν οι παρατηρήσεις μας προέρχονται από κανονικό πληθυσμό.

```
shapiro.test(n.suicides)


 Shapiro - Wilk normality test


data: n.suicides

W = 0.98675,      p - value = 0.9359
```

Έτσι σε επίπεδο στατιστικής σημαντικότητας α=0.05 δεν απορρίποτυμε την μηδενική υπόθεση, υποθέτωντας λοιπόν ότι τα δεδομένα μας προέρχονται από την κανονική κατανομή.

Όμως επειδή έχουμε τον αντισταθμιστή (*offset*) που θα θέλαμε να συμπεριλάβουμε στο μοντέλο μας για να κάνουμε πιο ακριβή τα συμπεράσματά μας, αντί για την Κανονική Κατανομή θα χρησιμοποιήσουμε την Gamma με *log link*.

Α. Υπάρχουν διαφορές μεταξύ των τριών ετών στον αναμενόμενο αριθμό αυτοκτονιών, αγνοώντας την επίδραση των μηνών;

Το υποερώτημα μας λέει να μην βάλουμε το o*ffset* μέσα στην συνάρτηση του μοντέλου. Τώρα ας καταλάβουμε διαισθητικά τι μας ζητάει. Αν ορίσω έναν παράγοντα $year$ με τα τρία έτη ως επίπεδα, τότε θα έχουμε ένα μοντέλο της μορφής:

$$log(suicides_i) = \beta_0 + \beta_1 \cdot year_{i2} + \beta_2 \cdot year_{i3} \qquad M_1$$

Όπου:

$$year_{ij} = \begin{cases} 0, & \alpha\nu\ i \neq j \\ 1, & \alpha\nu\ i = j \end{cases}, \qquad \mu\varepsilon\ i = 1, 2, 3.$$

Το παραπάνω μοντέλο θεωρεί ότι υπάρχουν διαφορές μεταξύ των τριών επιπέδων που έχουμε, επομένως κάθε φορά που αλλάζουμε επίπεδο αλλάζει και η τιμή της σταθεράς. Εάν δεν είχαμε διαφορές μεταξύ των τριών επιπέδων, θα ήταν σαν να λέγαμε ότι έχουμε μόνο ένα επίπεδο, οπότε το μοντέλο μας θα ήταν της μορφής:

$$log(suicides_i) = \beta_0 \qquad M_0$$

Εφόσον όλα είναι ένα επίπεδο, η αναμενόμενη τιμή των αυτοκτονιών δεν θα αλλάζει καθώς πάμε σε διαφορετικά έτη. Συνεπώς, το ερώτημα αυτό μας ζητάει να συγκρίνουμε τα δυο αυτά μοντέλα και να δούμε αν ο παράγοντας $year$ είναι στατιστικά σημαντικός.

Πάμε κατ'αρχάς να προσαρμόσουμε το μοντέλο M_1 στην R. Στην αρχή θα ορίσω τον παράγοντα $year$, ο οποίος έχει τρία επίπεδα:

```
year <- factor(rep(c(1,2,3), each = 12))
levels(year)
[1] "1" "2" "3"
```

Άρα το μοντέλο M_1, θα είναι:

```
mod1 <- glm(n.suicides ~ year, family = Gamma(link = "log"))
summary(mod1)


Call:
glm(formula = n.suicides ~ year, family = Gamma(link = "log"))


Deviance Residuals:
      Min         1Q     Median         3Q        Max
-0.143421  -0.034996   0.000178   0.039138   0.078238


Coefficients:
```

```
                Estimate  Std.Error  t value  Pr(>|t|)
(Intercept)     7.48493   0.01478    506.32   < 2e-16 ***
year2           0.04537   0.02091      2.17   0.0373 *
year3           0.09407   0.02091      4.50   7.99e-05 ***
---
Signif.codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Gamma family taken to be 0.002622402)

    Null deviance: 0.141116  on 35  degrees of freedom
Residual deviance: 0.087982  on 33  degrees of freedom
AIC: 435.84

Number of Fisher Scoring iterations: 3
```

Έχωντας κάνει αυτό πάμε να ελέγξουμε αν ο παράγοντας $year$ είναι στατιστικά σημαντικός στο μοντέλο αυτό.

Αυτό που θέλουμε να ελέγξουμε είναι το ποιά από τις παρακάτω δύο υποθέσεις ισχύει:

$$H_0: M_0 = M_1$$

$$H_1: M_0 \neq M_1$$

Έτσι, θα κάνουμε Likelihood Ratio Test, όπως ακριβώς ξέρουμε.

```
anova(mod1, test = F)
Analysis of Deviance Table

Model: Gamma, link: log

Response: n.suicides

Terms added sequentially (first to last)
```

```
      Df  Deviance  Resid.Df  Resid.Dev      F   Pr(>F)
NULL                     35   0.141116
year   2  0.053134        33   0.087982 10.131 0.0003712 ***
---
Signif.codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Συνεπώς σε επίπεδο στατιστικής σημαντικότητας $\alpha = 0.05$ απορρίπτουμε την μηδενική υπόθεση (διότι $\boldsymbol{p - value < a \Longrightarrow 0.0003712 < 0.05)}$ και έτσι καταλήγουμε στο ότι υπάρχουν ισχυρές ενδείξεις πως τα μοντέλα αυτά δεν είναι ίσα. Και επειδή πάντα προτιμούμε το μεγαλύτερο, θα διαλέξουμε το μοντέλο M_1.

Αυτό το συμπέρασμα όμως επίσης μας λέει ότι <u>υπάρχουν διαφορές</u> μεταξύ των τριών ετών στον αναμενόμενο αριθμό αυτοκτονιών.

Β. Υπάρχουν διαφορές μεταξύ των τεσσάρων εποχών θεωρώντας ότι οι διαφορές αυτές είναι ίδιες για όλα τα έτη;

Πάμε να δούμε λοιπόν το ίδιο πράγμα, όμως τώρα για τις εποχές. Εφόσον δεν μας λέει κάτι για το *offset*, θα το συμπεριλάβουμε στο μοντέλο, γιατί αυτό είναι το σωστό. Το μοντέλο μου με τον παράγοντα *season* με τέσσερα επίπεδα (όπου το πρώτο επίπεδο "χειμώνας" θα θεωρηθεί σαν reference group) θα είναι το εξής:

$$log(suicides_i) = log(days) + \beta_0 + \beta_1 \cdot season_{i2} + \beta_2 \cdot season_{i3} + \beta_3 \cdot season_{i4} \qquad M_2$$

Όπου:

$$season_{ij} = \begin{cases} 0, & \alpha\nu\ i \neq j \\ 1, & \alpha\nu\ \ i = j \end{cases}, \qquad i = 1, 2, 3, 4.$$

Όπως ακριβώς και στο προηγούμενο ερώτημα, το μοντέλο αυτό υποθέτει ότι υπάρχουν διαφορές μεταξύ των τεσσάρων εποχών. Φτιάχνουμε πρώτα τον παράγοντα εποχές:

```
season <-rep(c(1,rep(2,3),rep(3,3),rep(4,3),1,1),times = 3)
season <- factor(season)
```

Και ύστερα προσαρμόζουμε το νέο μας μοντέλο M_2:

```
mod2 <- glm(n.suicides ~ season, offset = log(mod.offset), family
          = Gamma(link = "log"))
summary(mod2)

Call:
glm(formula = n.suicides ~ season, family = Gamma(link = "log"),
  offset = log(mod.offset))

Deviance Residuals:
      Min         1Q     Median         3Q        Max
-0.10150  -0.03251  -0.00663   0.02941   0.10200

Coefficients:
            Estimate  Std.Error   t value    Pr(>|t|)
(Intercept) 4.08744   0.01947   209.934   < 2e-16 ***
season2     0.05790   0.02753     2.103   0.0434 *
season3     0.03123   0.02753     1.134   0.2651
season4     0.04054   0.02753     1.472   0.1507
---
Signif.codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Gamma family taken to be 0.003411757)

    Null deviance: 0.12464  on 35  degrees of freedom
Residual deviance: 0.10878  on 32  degrees of freedom
AIC: 445.49

Number of Fisher Scoring iterations: 3
```

Πάμε να ελέγξουμε αν ο παράγοντας $season$ είναι στατιστικά σημαντικός στο μοντέλο αυτό.

Αυτό που θέλουμε να ελέγξουμε είναι το ποιά από τις παρακάτω δύο υποθέσεις ισχύει:

$$H_0: M_0 = M_2$$

$$H_1: M_0 \neq M_2$$

Το μηδενικό μοντέλο τώρα περιλαμβάνει και το *offset*.

Έπομένως:

```
anova(mod2, test = "F")
Analysis of Deviance Table

Model: Gamma, link: log

Response: n.suicides

Terms added sequentially (first to last)
```

	Df	Deviance	Resid.Df	Resid.Dev	F	Pr(>F)
NULL			35	0.12464		
season	3	0.015859	32	0.10878	1.5495	0.2208

Σε επίπεδο στατιστικής σημαντικότητας $\alpha = 0.05$ δεν απορρίπτουμε την μηδενική υπόθεση (διότι $\boldsymbol{p - value > a \Longrightarrow 0.2208 > 0.05}$) και έτσι καταλήγουμε στο ότι υπάρχουν ισχυρές ενδείξεις πως τα μοντέλα αυτά είναι ίσα. Δηλαδή στο μοντέλο αυτό ο παράγοντας $season$ βγαίνει να είναι στατιστικά ασήμαντος.

Παρατηρήστε όμως ότι αν είχαμε το μοντέλο:

$$log(suicides_{ik}) = log(days) + \beta_0 + \beta_1 \cdot year_{i2} + \beta_2 \cdot year_{i3} + \beta_3 \cdot season_{k2} + \\ + \beta_4 \cdot season_{k3} + \beta_5 \cdot season_{k4} \qquad M_3$$

Όπου:

$$year_{ij} = \begin{cases} 0, & \alpha\nu\ i \neq j \\ 1, & \alpha\nu\ i = j \end{cases}, \; season_{kj} = \begin{cases} 0, & \alpha\nu\ k \neq j \\ 1, & \alpha\nu\ k = j \end{cases}, \qquad i = 1,2,3 \quad \kappa\alpha\iota \quad k = 1,2,3,4.$$

Τότε αν το συγκρίναμε με ένα ολόιδιο όπου δεν υπάρχει ο παράγοντας $season$, δηλαδή με ένα μοντέλο όπως το παρακάτω:

$$log(suicides_{ik}) = log(days) + \beta_0 + \beta_1 \cdot year_{i2} + \beta_2 \cdot year_{i3} \qquad M_4$$

Όπου:

$$year_{ij} = \begin{cases} 0, & \alpha\nu\ i \neq j \\ 1, & \alpha\nu\ i = j \end{cases}, \quad i = 1, 2, 3$$

Τότε θα παρατηρούσαμε το εξής:

```
mod3 <- glm(n.suicides ~ year + season, offset = log(mod.offset), family
          = Gamma(link = "log"))

mod4 <- glm(n.suicides ~ year, offset = log(mod.offset), family
          = Gamma(link = "log"))

anova(mod4, mod3, test = "F")

Analysis of Deviance Table


Model 1: n.suicides ~ year

Model 2: n.suicides ~ year + season

  Resid. Df  Resid. Dev  Df   Deviance        F    Pr(>F)
1        33    0.071532
2        30    0.055336   3   0.016196   2.9489   0.04862 *
---

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Τώρα έστω και οριακά απορρίπτουμε την μηδενική υπόθεση, έτσι λοιπόν ο παράγοντας $season$ γίνεται στατιστικά σημαντικός σε επίπεδο στατιστικής σημαντικότητας $\alpha = 0.05$, αλλά μόνο αν συμπεριλάβουμε τον παράγοντα $year$ μέσα στο μοντέλο μας.

Γ. Οι διαφορές των τεσσάρων εποχών φαίνεται να είναι διαφορετικές μεταξύ των τριών ετών;

Οι διαφορές αυτές αναφέρονται στην αλληλεπίδραση μεταξύ των δύο αυτών παραγόντων. Αν δεν υπάρχει αλληλεπίδραση, δεν υπάρχουν και διαφορές μεταξύ των εποχών ανα κάθε έτος. Το νέο μας μοντέλο θα είναι:

$$log(suicides_{ik}) = log(days) + \beta_0 + \beta_1 \cdot year_{i2} + \beta_2 \cdot year_{i3} + \beta_3 \cdot season_{k2} + \\ + \beta_4 \cdot season_{k3} + \beta_5 \cdot season_{k4} + \beta_6 \cdot year_{i2} \cdot season_{k2} + \beta_7 \cdot year_{i3} \cdot season_{k2} + \\ + \beta_8 \cdot year_{i2} \cdot season_{k3} + \beta_9 \cdot year_{i3} \cdot season_{k3} + \beta_{10} \cdot year_{i2} \cdot season_{k4} + \\ + \beta_{11} \cdot year_{i3} \cdot season_{k4} \qquad M_5$$

Όπου:

$$year_{ij} = \begin{cases} 0, & \alpha\nu\ i \neq j \\ 1, & \alpha\nu\ i = j \end{cases}, \quad season_{kj} = \begin{cases} 0, & \alpha\nu\ k \neq j \\ 1, & \alpha\nu\ k = j' \end{cases} \qquad i = 1,2,3 \quad \kappa\alpha\iota \quad k = 1,2,3,4.$$

Το μοντέλο M_5 προσαρμόζεται στην *R*, ως εξής:

```
mod5 <- glm(n.suicides ~ year * season, offset = log(mod.offset), family
            = Gamma(link = "log"))
summary(mod5)


Call:
glm(formula = n.suicides ~ year * season, family = Gamma(link = "log"),
  offset = log(mod.offset))

Deviance Residuals:
    Min      1Q  Median      3Q     Max
-0.074546 -0.030381  0.002545  0.026815  0.064772

Coefficients:
             Estimate  Std.Error  t value   Pr(>|t|)
(Intercept)  4.01934   0.02540  158.238   < 2e-16 ***
year2        0.09313   0.03592    2.593   0.01597 *
year3        0.10779   0.03592    3.001   0.00619 **
```

```
season2           0.10748   0.03592    2.992 0.00632 **
season3           0.05583   0.03592    1.554 0.13323
season4           0.04881   0.03592    1.359 0.18687
year2:season2    -0.09330   0.05080   -1.837 0.07870 .
year3:season2    -0.05305   0.05080   -1.044 0.30673
year2:season3    -0.06199   0.05080   -1.220 0.23423
year3:season3    -0.01088   0.05080   -0.214 0.83217
year2:season4    -0.03558   0.05080   -0.700 0.49037
year3:season4     0.01065   0.05080    0.210 0.83570
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Gamma family taken to be 0.001935579)

    Null deviance: 0.124642  on 35  degrees of freedom
Residual deviance: 0.046631  on 24  degrees of freedom
AIC: 430.98

Number of Fisher Scoring iterations: 3
```

Και αυτό που θέλουμε να ελέγξουμε είναι αν η αλληλεπίδραση είναι στατιστικά σημαντική στο μοντέλο μας. Οπότε θα συγκρίνουμε το μοντέλο M_3 με το μοντέλο M_5, ως εξής:

Θα ελέγξουμε ποιά από τις παρακάτω δύο υποθέσεις ισχύει:

$$H_0: M_3 = M_5$$

$$H_1: M_3 \neq M_5$$

Άρα:

```
anova(mod3, mod5, test = "F")
Analysis of Deviance Table
```

$$\boldsymbol{Model\ 1: n.suicides \sim year + season}$$

$$\boldsymbol{Model\ 2: n.suicides \sim year * season}$$

$\boldsymbol{Resid.Df}$	$\boldsymbol{Resid.Dev}$	$\boldsymbol{Df}$	$\boldsymbol{Deviance}$	$\boldsymbol{F}$	$\boldsymbol{Pr(>F)}$
1	**30**	**0.055336**			
2	**24**	**0.046631**	**6** **0.0087057**	**0.7496**	**0.6157**

Οπότε σε επίπεδο στατιστικής σημαντικότητας $\alpha = 0.05$ δεν απορρίπτουμε την μηδενική υπόθεση (διότι $\boldsymbol{p - value > a \Longrightarrow 0.6157 > 0.05)}$ και έτσι καταλήγουμε στο ότι υπάρχουν ισχυρές ενδείξεις πως τα μοντέλα αυτά είναι ίσα. Με άλλα λόγια, στο μοντέλο αυτό η αλληλεπίδραση μεταξύ του παράγοντα $season$ με τον παράγοντα $year$ δεν είναι στατιστικά σημαντική. Αυτό είναι ένα αποτέλεσμα που περιμέναμε πάνω κάτω, διότι δεν έχουμε διαφορετικές εποχές σε κάθε έτος.

Συνεπώς, οι διαφορές των τεσσάρων εποχών φαίνεται να μην είναι διαφορετικές μεταξύ των τριών ετών.

Δ. Για το μοντέλο του $(B.)$ εκτιμήστε το λόγο του ρυθμού αυτοκτονιών μεταξύ Άνοιξης και καλοκαιριού για το 1969.

Για να λύσουμε αυτό το ερώτημα, θα χρειαστεί να αλλάξουμε τα επίπεδα αναφοράς στους παράγοντές μας, που μέχρι στιγμής για τον $season$ είναι ο "Χειμώνας", ενώ για τον $year$ είναι το έτος 1968. Οπότε θα τα αλλάξουμε σε "Άνοιξη" καθώς και σε έτος 1969 αντίστοιχα. Και τα δύο αυτά επίπεδα είναι στο επίπεδο 2 σε κάθε έναν από τους παράγοντες. Οπότε στην *R*, αυτό γίνεται ως εξής:

```
season2 <- relevel(season, ref = 2)
levels(season2)
[1] "2" "1" "3" "4"
```

και

```
year2 <- relevel(year, ref = 2)
levels(year2)
[1] "2" "1" "3"
```

Θα υποθέσουμε ότι το μοντέλο μας είναι το:

$$log(suicides_{ik}) = log(days) + \beta_0 + \beta_1 \cdot year_{i2} + \beta_2 \cdot year_{i3} + \beta_3 \cdot season_{k2} + \\ + \beta_4 \cdot season_{k3} + \beta_5 \cdot season_{k4} \qquad M_3$$

Όπου:

$$year_{ij} = \begin{cases} 0, & \alpha\nu\ i \neq j \\ 1, & \alpha\nu\ i = j \end{cases}, \; season_{kj} = \begin{cases} 0, & \alpha\nu\ k \neq j \\ 1, & \alpha\nu\ k = j' \end{cases} \qquad i = 1,2,3 \quad \kappa\alpha\iota \quad k = 1,2,3,4.$$

Τώρα λόγω του *offset*, δεν εκτιμάμε απλά τον αναμενόμενο αριθμό αυτοκτονιών, αλλά τον ρυθμό αυτοκτονιών, διότι:

$$log(suicides_{ik}) = log(days) + \beta_0 + \beta_1 \cdot year_{i2} + \beta_2 \cdot year_{i3} + \beta_3 \cdot season_{k2} + \\ + \beta_4 \cdot season_{k3} + \beta_5 \cdot season_{k4} \Longrightarrow$$

$$\Longrightarrow log\,(suicides_{ik}) - log(days) = \beta_0 + \beta_1 \cdot year_{i2} + \beta_2 \cdot year_{i3} + \beta_3 \cdot season_{k2} + \\ + \beta_4 \cdot season_{k3} + \beta_5 \cdot season_{k4} \Longrightarrow$$

$$\Longrightarrow log\left(\frac{suicides_{ik}}{days}\right) = \beta_0 + \beta_1 \cdot year_{i2} + \beta_2 \cdot year_{i3} + \beta_3 \cdot season_{k2} + \\ + \beta_4 \cdot season_{k3} + \beta_5 \cdot season_{k4}$$

Τώρα λοιπόν το νέο μας μοντέλο είναι:

```
mod3.new <- glm(n.suicides ~ year2 + season2, offset
          = log(mod.offset), family = Gamma(link = "log"))

summary(mod3.new)


Call:

glm(formula = n.suicides ~ year2 + season2, family = Gamma(link
          = "log"),
  offset = log(mod.offset))


Deviance Residuals:

      Min         1Q     Median         3Q        Max
-0.09144  -0.03614   0.01080   0.02773   0.07501
```

Coefficients:

	Estimate	Std.Error	t value	Pr(>|t|)	
(Intercept)	4.14402	0.01747	237.239	< 2e − 16	***
year21	− 0.04538	0.01747	− 2.598	0.01439	*
year23	0.04896	0.01747	2.803	0.00879	**
season21	− 0.05870	0.02017	− 2.910	0.00675	**
season23	− 0.02734	0.02017	− 1.355	0.18545	
season24	− 0.01836	0.02017	− 0.910	0.37004	

− − −

Signif.codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(**Dispersion parameter for Gamma family taken to be** 0.001830727)

Null deviance: 0.124642 **on** 35 **degrees of freedom**

Residual deviance: 0.055336 **on** 30 **degrees of freedom**

AIC: 425.14

Number of Fisher Scoring iterations: 3

Οπότε σύμφωνα με τα επίπεδα αναφοράς που έχουμε ορίσει, για $i = 1$, έχουμε το έτος 1969, για $k = 1$ έχουμε την εποχή "Άνοιξη" και για $k = 3$, έχουμε την εποχή "Καλοκαίρι".

Αυτό που μας ζητάει το ερώτημα στην ουσία είναι:

$$log\left(\frac{suicides_{13}}{days}\right) - log\left(\frac{suicides_{11}}{days}\right) = \beta_4 + \beta_0 - \beta_0 \Longrightarrow$$

$$\Longrightarrow log\left(\frac{\frac{suicides_{13}}{days}}{\frac{suicides_{11}}{days}}\right) = \beta_4 \Longrightarrow \boxed{\frac{\textbf{Summer Suicide Rate}}{\textbf{Spring Suicide Rate}} = \boldsymbol{exp}\{-\mathbf{0.02734}\}}$$

Επομένως, ο ρυθμος αυτοκτονιών καλοκαιριού προς τον ρυθμό αυτοκτονιών την άνοιξη είναι $exp\{-0.02734\} = 0.9730304$. Δηλαδή ο ρυθμός αυτοκτονιών το Καλοκαίρι είναι περίπου κατα 2.7% μικρότερος από τον αντίστοιχο ρυθμό αυτοκτονιών την Άνοιξη το έτος 1969.

Ε. Για το μοντέλο του ($B.$) ελέγξτε αν δικαιολογείται η υπόθεση ο αναμενόμενος αριθμός αυτοκτονιών είναι ανάλογος του αριθμού των ημερών του μήνα παρατήρησης. Περιγράψτε τι σημαίνει η μη ισχύς της υπόθεσης αυτής για το πρόβλημα που εξετάζουμε.

Αυτό που θέλουμε να ελέγξουμε είναι αν ισχύει η υπόθεση της αναλογικότητας. Ακόμα και αν ξέρουμε ήδη την απάντηση, ας ελέγξουμε αν η αναλογικότητα είναι στατιστικά σημαντική στο μοντέλο μας.

```
mod6 <- glm(n.suicides ~ log(mod.offset), family = Gamma(link = "log"))
summary(mod6)


Call:
glm(formula = n.suicides ~ log(mod.offset), family = Gamma(link
    = "log"))


Deviance Residuals:
      Min         1Q     Median         3Q        Max
-0.128367  -0.037940  -0.001158   0.037541   0.125909


Coefficients:
                 Estimate  Std.Error  t value  Pr(>|t|)
(Intercept)        4.8428     1.1955    4.051  0.00028 ***
log(mod.offset)    0.7882     0.3504    2.249  0.03107 *
---
Signif.codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


(Dispersion parameter for Gamma family taken to be 0.003628157)


    Null deviance: 0.14112  on 35  degrees of freedom
Residual deviance: 0.12334  on 34  degrees of freedom
AIC: 446.01
```

```
Number of Fisher Scoring iterations: 3
```

Και έτσι:

```
anova(mod6, test = "F")

Analysis of Deviance Table


Model: Gamma, link: log


Response: n.suicides


Terms added sequentially (first to last)



                    Df  Deviance Resid. Df Resid. Dev      F
NULL                                    35     0.14112
log(mod.offset)  1          0.017777  34     0.12334 4.8998
                    Pr(>F)
NULL
log(mod.offset) 0.03367 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Επομένως σε επίπεδο στατιστικής σημαντικότητας α=0.05 απορρίπτουμε την μηδενική υπόθεση και έτσι η τιμή αυτή είναι στατιστικά σημαντική για το μοντέλο μας. Ας το ελέγξουμε και για το μονέλο M_2.

```
mod3.2 <- glm(n.suicides ~ year + season + log(mod.offset), family
          = Gamma(link = "log"))

summary(mod3.2)


Call:
```

$$glm(formula = n.suicides \sim year + season + log(mod.offset), family = Gamma(link = "log"))$$

Deviance Residuals:

Min	1Q	Median	3Q	Max
−0.08559	−0.03506	0.01035	0.02647	0.07734

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	3.68367	0.98834	3.727	0.000835	***
year2	0.04535	0.01774	2.557	0.016056	*
year3	0.09436	0.01774	5.321	1.04e−05	***
season2	0.06212	0.02270	2.737	0.010488	*
season3	0.03131	0.02048	1.529	0.137155	
season4	0.04149	0.02072	2.002	0.054674	.
log(mod.offset)	1.10409	0.28868	3.825	0.000643	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(**Dispersion parameter for Gamma family taken to be** 0.001887265)

Null deviance: 0.141116 **on** 35 **degrees of freedom**

Residual deviance: 0.055095 **on** 29 **degrees of freedom**

AIC: 426.99

Number of Fisher Scoring iterations: 4

Οπότε:

```
anova(mod3.2, test = "F")
Analysis of Deviance Table

Model: Gamma, link: log

Response: n.suicides

Terms added sequentially (first to last)


                Df  Deviance  Resid. Df  Resid. Dev       F
NULL                                 35    0.141116
year             2   0.053134        33    0.087982  14.0769
season           3   0.005880        30    0.082102   1.0385
log(mod.offset)  1   0.027007        29    0.055095  14.3103
                Pr(>F)
NULL
year            5.341e-05 ***
season          0.3902299
log(mod.offset) 0.0007189 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Οπότε για άλλη μια φορά συμπεραίνουμε ότι το offset είναι στατιστικά σημαντικό για το μοντέλο μας.

Τώρα όμως ποιό θα μπορούσε να είναι το θέμα; Θα μπορούσε να μην ισχύει η υπόθεση αναλογικότητας. Και έτσι ο αναμενόμενος αριθμός αυτοκτονιών δεν θα ήταν ανάλογος των ημερών του μήνα παρατήρησης, δηλαδή για το μοντέλο M_3 δεν θα είχαμε αυτό:

$$log\left(\frac{suicides_{ik}}{days}\right) = \beta_0 + \beta_1 \cdot year_{i2} + \beta_2 \cdot year_{i3} + \beta_3 \cdot season_{k2} +$$

$$+ \beta_4 \cdot season_{k3} + \beta_5 \cdot season_{k4}$$

Αλλά θα ήταν ανάλογος του μήνα παρατήρησης υψωμένου σε κάποια δύναμη $\boldsymbol{\gamma}$. Δηλαδή θα είχαμε αυτό:

$$log(suicides_{ik}) = \gamma \cdot log(days) + \beta_0 + \beta_1 \cdot year_{i2} + \beta_2 \cdot year_{i3} + \beta_3 \cdot season_{k2} +$$

$$+ \beta_4 \cdot season_{k3} + \beta_5 \cdot season_{k4} \Longrightarrow$$

$$\Longrightarrow log\,(suicides_{ik}) - log(days^{\gamma}) = \beta_0 + \beta_1 \cdot year_{i2} + \beta_2 \cdot year_{i3} + \beta_3 \cdot season_{k2} +$$

$$+ \beta_4 \cdot season_{k3} + \beta_5 \cdot season_{k4} \Longrightarrow$$

$$\Longrightarrow log\left(\frac{suicides_{ik}}{days^{\gamma}}\right) = \beta_0 + \beta_1 \cdot year_{i2} + \beta_2 \cdot year_{i3} + \beta_3 \cdot season_{k2} +$$

$$+ \beta_4 \cdot season_{k3} + \beta_5 \cdot season_{k4}$$

Οπότε λοιπόν για να βγάλουμε κάποιο συμπέρασμα, θα πρέπει να ελέγξουμε ποιά από τις δύο υποθέσεις ισχύει:

$$H_0: \gamma = 1$$

$$H_1: \gamma \neq 1$$

Α'τρόπος:

Από τις τιμές του πίνακα ***summary***(), φτιάχνουμε την $z - value$:

$$z = \frac{\hat{\gamma} - 1}{Standard\ deviation_{\gamma}} = \frac{1.10409 - 1}{0.28868} = \frac{0.10409}{0.28868} = 0.3605723$$

Και ξέρουμε ότι:

$$z \sim N(0,1)$$

Οπότε πάμε να βρούμε την $p - value$ του:

```
2 * (1 - pnorm(0.3605723, 0, 1))

[1] 0.7184192
```

Παρατηρούμε λοιπόν ότι $\boldsymbol{p - value > a \Longrightarrow 0.7184192 > 0.05}$ οπότε δεν απορρίπτουμε την μηδενική υπόθεση, καταλήγοντας στο συμπέρασμα ότι υπάρχουν ισχυρές ενδείξεις ότι ισχύει η αναλογικότητα. Ή με άλλα λόγια, ότι δικαιολογείται η υπόθεση ο αναμενόμενος αριθμός αυτοκτονιών να είναι ανάλογος του αριθμού των ημερών του μήνα παρατήρησης.

Β'τρόπος:

Μπορούμε επίσης να δούμε αν υπάρχει η μονάδα στο 95% διάστημα εμπιστοσύνης του μοντέλου μας.

```
confint(mod3.2)
Waiting for profiling to be done...
                    2.5 %       97.5 %
(Intercept)     1.7332978157  5.63935515
year2           0.0105825586  0.08011364
year3           0.0596004085  0.12912742
season2         0.0178286084  0.10649758
season3        -0.0088346167  0.07145048
season4         0.0008593156  0.08212159
log(mod.offset) 0.5328018350  1.67384931
```

Στην τελευταία γραμμή μπορούμε να δούμε ότι ο αριθμός 1 είναι μέσα στο διάστημα εμπιστοσύνης και έτσι δεν απορρίπτουμε την μηδενική υπόθεση, καταλήγοντας στο ότι υπάρχει αναλογικότητα στο μοντέλο μου.

Τι θα σήμαινε η μη ισχύς της υπόθεσης αυτής για το πρόβλημα που εξετάζουμε;

Αν δεν ισχύει η αναλογικότητα, δηλαδή η $\gamma \neq 1$, τότε:

$$suicides' = \frac{suicides}{days^{\gamma}} \Longrightarrow suicides = days^{\gamma} \cdot suicides'$$

Αν δηλαδή διπλασιαστεί ο αριθμός των ημερών (θεωρητικά, γιατί έχουμε καθορισμένο αριθμό ημερών σε κάθε μήνα), τότε περιμένουμε ο αναμενόμενος αριθμός αυτοκτονιών να αυξηθεί κατά $days^{\gamma}$.

Πρακτικά βέβαια αυτό δεν μπορεί να γίνει και άλλωστε για αυτό καταλήξαμε στο ότι $\gamma = 1$.

Άσκηση 2:

Τα παρακάτω δεδομένα αποτελούν μέρος ενός πειράματος για να προσδιοριστεί η επίδραση της θερμοκρασίας και του χρόνου αποθήκευσης στην έλλειψη ασκορβικού οξέος σε φασόλια. Τα φασόλια μαζεύτηκαν κάτω από ομοιόμορφες συνθήκες πριν από τις 8 το πρωί. Ετοιμάστηκαν και πήραν τη θερμοκρασία συντήρησης πριν το μεσημέρι της ίδιας ημέρας. Τρία πακέτα από το προϊόν δόθηκαν στην τύχη σε κάθε συνδυασμό θερμοκρασίας και χρόνου αποθήκευσης. Το άθροισμα των τριών συκεντρώσεων ασκορβικού οξέος δίνεται στον πίνακα παρακάτω. Θεωρείστε τα δεδομένα ανεξάρτητα ως ακολουθούντα την κανονικη κατανομή.

	Εβδομάδες Αποθήκευσης			
Θερμοκρασία	2	4	6	8
0	45	47	46	46
10	45	43	41	37
20	34	28	21	16

Υποθέστε ότι η αρχική συγκέντρωση του ασκορβικού οξέος είναι ανεξάρτητη του χρόνου αποθήκευσης (ή αλλιώς ότι οι συγκεντρώσεις ασκορβικού οξέος είναι ίδιες στον χρόνο 0). Θεωρείστε την θερμοκρασία αποθήκευσης ως παράγοντα, ενώ το χρόνο αποθήκευσης ως συνεχή επεξηγηματική μεταβλητή.

Α. Προσαρμόστε ένα μοντέλο για τη μέση συγκέντρωση ασκορβικού οξέος μιας τριάδας πακέτων με τη βοήθεια της θερμοκρασίας και του χρόνου αποθήκευσης. Το μοντέλο θα υποθέτει μια σταθερά και μια γραμμική επίδραση των εβδομάδων αποθήκευσης για κάθε θερμοκρασία.

Β. Δώστε ερμηνεία στις παραμέτρους του μοντέλου.

Γ. Εκτιμήστε την αρχική συγκέντρωση (στον χρόνο 0) ασκορβικού οξέος μιας τριάδας πακέτων όταν η θερμοκρασία είναι 20 βαθμοί και δώσατε 95% Διάστημα Εμπιστοσύνης.

Δ. Δώστε το τυπικό σφάλμα της πρόβλεψης του $(\Gamma.)$ καθώς και το προσεγγιστικό 95% Διάστημα Εμπιστοσύνης.

Ε. Ποιά είναι η εκτίμηση της διακύμανσης των δεδομένων που λαμβάνετε από το μοτνέλο που χρησιμοποιήσατε στο $(A.)$;

ΣΤ. Προσθέστε την αλληλεπίδραση εβδομάδων αποθήκευσης και θερμοκρασίας στο μοντέλο ελέγξτε τη στατιστική του σημαντικότητα και σχολιάστε την επάρκεια του νέου μοντέλου κάνοντας οπτικό έλεγχο καταλοίπων.

ΛΥΣΗ:

Από την εκφώνηση μας δίνεται ότι τα δεδομένα μου ακολουθούν την κανονική κατανομή. Πάμε να τα εισάγουμε στην *R*:

```
acid <- c(45,45,34,47,43,28,46,41,21,46,37,16)
```

```
Temp <- rep(c(0,10,20), times = 4)
Temp <- factor(Temp)
weeks <- rep(c(2,4,6,8), each = 3)
```

Α. Προσαρμόστε ένα μοντέλο για τη μέση συγκέντρωση ασκορβικού οξέος μιας τριάδας πακέτων με τη βοήθεια της θερμοκρασίας και του χρόνου αποθήκευσης. Το μοντέλο θα υποθέτει μια σταθερά και μια γραμμική επίδραση των εβδομάδων αποθήκευσης για κάθε θερμοκρασία.

Το μοντέλο μας θα είναι της μορφής:

$$log(acid_{i,weeks}) = \beta_0 + \beta_1 \cdot Temp_{i1} \cdot weeks + \beta_2 \cdot Temp_{i2} \cdot weeks + \beta_3 \cdot Temp_{i3} \cdot weeks \quad \boxed{M_1}$$

Όπου:

$$Temp_{ij} = \begin{cases} 0, & \alpha v\ i \neq j \\ 1, & \alpha v\ i = j \end{cases}$$

Οπότε πάμε να το προσαρμόσουμε στην *R*:

```
mod1 <- glm(acid ~ Temp:weeks, family = gaussian(link = "log"))
summary(mod1)

Call:
glm(formula = acid ~ Temp:weeks, family = gaussian(link = "log"))

Deviance Residuals:
      Min         1Q     Median         3Q        Max
-1.49820  -0.38546   0.08059   0.80866   0.86028

Coefficients:
                Estimate  Std. Error   t value  Pr(>|t|)
```

$(Intercept)$ 3.8354389 0.0186464 205.694 $3.49e-16$ $***$

$Temp0:weeks$ -0.0007716 0.0037540 -0.206 0.842286

$Temp10:weeks$ -0.0236121 0.0040351 -5.852 0.000382 $***$

$Temp20:weeks$ -0.1329785 0.0061597 -21.588 $2.23e-08$ $***$

$---$

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

($Dispersion\ parameter\ for\ gaussian\ family\ taken\ to\ be\ 1.124727$)

$Null\ deviance$: 1226.9167 on 11 $degrees\ of\ freedom$

$Residual\ deviance$: 8.9978 on 8 $degrees\ of\ freedom$

AIC: 40.599

$Number\ of\ Fisher\ Scoring\ iterations$: 3

Εδώ βάζουμε $link = log$, διότι μας δίνει καλύτερο $residual\ deviance$ απο την $identity$.

Β. Δώστε ερμηνεία στις παραμέτρους του μοντέλου.

Για το β_0: Όταν ο χρόνος αποθήκευσης των τριών πακέτων από τα φασόλια είναι ίσος με το μηδέν, τότε η αναμενόμενη τιμή της συγκέντρωσης ασκορβικού οξέος μέσα στα τρία πακέτα αυτά θα είναι:

$$acid_{i,0} = \exp\{3.8354389\} = 46.31375$$

Για το β_1: Εαν τα φασόλια από τα τρία πακέτα φυλαχτούν σε θερμοκρασία 0, τότε η αύξηση του χρόνου φύλαξης φασολιών κατά μιά μονάδα, θα δημιουργήσει λόγο αναμενόμενων τιμών:

$$\frac{acid_{1,weeks+1}}{acid_{1,weeks}} = \exp\{-0.0007716\} = 0.9992287$$

Ή με άλλα λόγια, αν αυξηθεί ο χρόνος φύλαξης κατά μία μονάδα σε τριάδα πακέτων φασολιών που φυλάχτηκαν σε θερμοκρασία 0 βαθμών, τότε θα υπάρξει μείωση της συγκέντρωσης του ασκορβικού οξέος κατά 0.007713%.

Για το $\boldsymbol{\beta_2}$: Αν τα φασόλια από τα τρία πακέτα φυλαχτούν σε θερμοκρασία 10 βαθμών, τότε η αύξηση του χρόνου φύλαξης φασολιών κατά μιά μονάδα, θα δημιουργήσει λόγο αναμενόμενων τιμών:

$$\frac{acid_{2,weeks+1}}{acid_{2,weeks}} = \exp\{-0.0236121\} = 0.9766645$$

Ή με άλλα λόγια, αν αυξηθεί ο χρόνος φύλαξης κατά μία μονάδα σε τριάδα πακέτων φασολιών που φυλάχτηκαν σε θερμοκρασία 10 βαθμών, τότε θα υπάρξει μείωση της συγκέντρωσης του ασκορβικού οξέος κατά 2.33355%.

Για το $\boldsymbol{\beta_3}$: Αν τα φασόλια από τα τρία πακέτα φυλαχτούν σε θερμοκρασία 20 βαθμών, τότε η αύξηση του χρόνου φύλαξης φασολιών κατά μιά μονάδα, θα δημιουργήσει λόγο αναμενόμενων τιμών:

$$\frac{acid_{3,weeks+1}}{acid_{3,weeks}} = \exp\{-0.1329785\ \} = 0.8754839$$

Ή με άλλα λόγια, αν αυξηθεί ο χρόνος φύλαξης κατά μία μονάδα σε τριάδα πακέτων φασολιών που φυλάχτηκαν σε θερμοκρασία 20 βαθμών, τότε θα υπάρξει μείωση της συγκέντρωσης του ασκορβικού οξέος κατά 12.45161%.

Γ. Εκτιμήστε την αρχική συγκέντρωση (στον χρόνο 0) ασκορβικού οξέος μιας τριάδας πακέτων όταν η θερμοκρασία είναι 20 βαθμοί και δώσατε 95% Διάστημα Εμπιστοσύνης.

Στο μοντέλο που προσαρμόσαμε, ότι θερμοκρασία και να βάλουμε στον παράγοντα θα πάρουμε την ίδια τιμή, εφόσον ο χρόνος φύλαξης είναι ίσος με το 0. Αυτό που θα πάρουμε είναι απλά το

β_0 στο μοντέλο μας. Αυτό βγάζει νόημα γιατί βρισκόμαστε σε περίοδο πριν τοποθετήσουμε τα τρία πακέτα στη συντήρηση και επομένως είναι λογικό η αναμενόμενη τιμή της συγκέντρωσης του ασκορβικού οξέος να είναι ίδια. Μπορούμε φυσικά να κάνουμε και πρόβλεψη στην *R*.

```
new <- data.frame(Temp = factor(20), weeks = 0)
pred <- predict(mod1, newdata = new, type = 'response', se.fit = TRUE)
$fit
       1
46.31375

$se.fit
         1
0.8635833

$residual.scale
[1]1.060532

exp(confint(mod1))
Waiting for profiling to be done...
                          2.5 %      97.5 %
(Intercept)          44.6459356  48.0175765
Temp0:weeks           0.9918958   1.0065679
Temp10:weeks          0.9689969   0.9843153
Temp20:weeks          0.8647605   0.8859540
```

Όπως ακριβώς το περιμέναμε, μας πρόβλεψε την e^{β_0}. Το διάστημα εμπιστοσύνης 95% είναι αυτό της τιμής της σταθεράς (*intercept*). Δηλαδή:

$$[44.6459356,\ \ 48.0175765]$$

Δ. Δώστε το τυπικό σφάλμα της πρόβλεψης του (*Γ.*) καθώς και το προσεγγιστικό 95% Διάστημα Εμπιστοσύνης.

Το τυπικό σφάλμα της πρόβλεψης ήταν:

```
pred$se.fit
       1
0.8635833
```

Τώωωωωωωωωρα... Προσεγγιστικό, μα το θεό δεν έχω ιδέα τι εννοεί.

Έστω ότι είναι αυτό:

$$\left[\hat{\beta}_0 - t\ value \cdot standard\ error, \qquad \hat{\beta}_0 + t\ value \cdot standard\ error\right]$$

Αυτό στην R θα είναι:

```
exp(3.8354389) - 0.0186464 * 205.694
[1] 42.4783
exp(3.8354389) + 0.0186464 * 205.694
[1] 50.1492
```

Οπότε το προσεγγιστικό 95% διάστημα εμπιστοσύνης θα είναι:

$$[\mathbf{42.4783}, \qquad \mathbf{50.1492}\ ]$$

Ε. Ποιά είναι η εκτίμηση της διακύμανσης των δεδομένων που λαμβάνετε από το μοντέλο που χρησιμοποιήσατε στο (*Α.*);

$$Variance = \frac{Residual\ Sum\ of\ Squares}{n - p} = \frac{8.9978}{8} = 1.124725$$

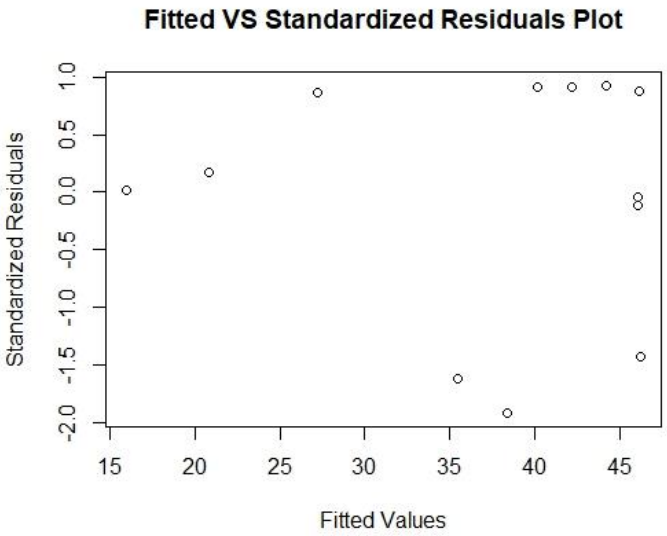
ΣΤ. Προσθέστε την αλληλεπίδραση εβδομάδων αποθήκευσης και θερμοκρασίας στο μοντέλο ελέγξτε τη στατιστική του σημαντικότητα και σχολιάστε την επάρκεια του νέου μοντέλου κάνοντας οπτικό έλεγχο καταλοίπων.

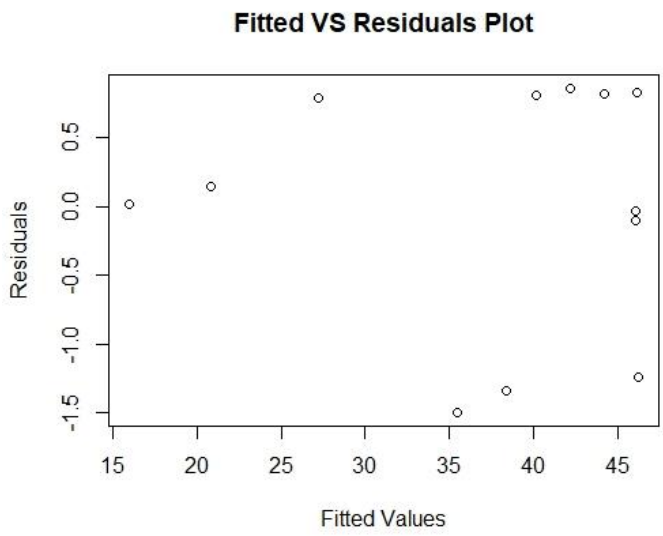
Το έχουμε ήδη κάνει στο ερώτημα Α. Στατιστική σημαντικότητα λόγω του ότι το φ είναι άγνωστο δυστυχώς μπορεί να γίνει μόνο με διαγράμματα.

Για να δούμε λοιπόν τι θα μας δείξουν:

```
plot(fitted(mod1), resid(mod1, type = "pearson"), xlab = "Fitted Values", ylab
        = "Residuals", main = "Fitted VS Residuals Plot")

plot(fitted(mod1), rstandard(mod1, type = "pearson"), xlab
        = "Fitted Values", ylab = "Standardized Residuals", main
        = "Fitted VS Standardized Residuals Plot")
```





Φαίνονται εντάξει τα διαγράμματα, δεν φαίνεται να έχουμε και *outliers*. Δηλαδή φαίνεται σαν μια αρκετά καλή προσαρμογή.

Άσκηση 2 (Take Two, Vasdekis Revenge):

Τα παρακάτω δεδομένα αποτελούν μέρος ενός πειράματος για να προσδιοριστεί η επίδραση της θερμοκρασίας και του χρόνου αποθήκευσης στην έλλειψη ασκορβικού οξέος σε φασόλια. Τα φασόλια μαζεύτηκαν κάτω από ομοιόμορφες συνθήκες πριν από τις 8 το πρωί. Ετοιμάστηκαν και πήραν τη θερμοκρασία συντήρησης πριν το μεσημέρι της ίδιας ημέρας. Τρία πακέτα από το προϊόν δόθηκαν στην τύχη σε κάθε συνδυασμό θερμοκρασίας και χρόνου αποθήκευσης. Το άθροισμα των τριών συκεντρώσεων ασκορβικού οξέος δίνεται στον πίνακα παρακάτω. Θεωρείστε τα δεδομένα ανεξάρτητα ως ακολουθούντα την κανονικη κατανομή.

	Εβδομάδες Αποθήκευσης			
Θερμοκρασία	2	4	6	8
0	45	47	46	46
10	45	43	41	37
20	34	28	21	16

Υποθέστε ότι η αρχική συγκέντρωση του ασκορβικού οξέος είναι ανεξάρτητη του χρόνου αποθήκευσης (ή αλλιώς ότι οι συγκεντρώσεις ασκορβικού οξέος είναι ίδιες στον χρόνο 0). Θεωρείστε την θερμοκρασία αποθήκευσης ως παράγοντα, ενώ το χρόνο αποθήκευσης ως συνεχή επεξηγηματική μεταβλητή.

Α. Προσαρμόστε ένα μοντέλο για τη μέση συγκέντρωση ασκορβικού οξέος μιας τριάδας πακέτων με τη βοήθεια της θερμοκρασίας και του χρόνου αποθήκευσης. Το μοντέλο θα υποθέτει μια σταθερά και μια γραμμική επίδραση των εβδομάδων αποθήκευσης για κάθε θερμοκρασία.

Β. Δώστε ερμηνεία στις παραμέτρους του μοντέλου.

Γ. Εκτιμήστε την αρχική συγκέντρωση (στον χρόνο 0) ασκορβικού οξέος μιας τριάδας πακέτων όταν η θερμοκρασία είναι 20 βαθμοί και δώσατε 95% Διάστημα Εμπιστοσύνης.

Δ. Δώστε το τυπικό σφάλμα της πρόβλεψης του $(\Gamma.)$ καθώς και το προσεγγιστικό 95% Διάστημα Εμπιστοσύνης.

Ε. Ποιά είναι η εκτίμηση της διακύμανσης των δεδομένων που λαμβάνετε από το μοτνέλο που χρησιμοποιήσατε στο $(A.)$;

ΣΤ. Προσθέστε την αλληλεπίδραση εβδομάδων αποθήκευσης και θερμοκρασίας στο μοντέλο ελέγξτε τη στατιστική του σημαντικότητα και σχολιάστε την επάρκεια του νέου μοντέλου κάνοντας οπτικό έλεγχο καταλοίπων.

ΛΥΣΗ:

Από την εκφώνηση μας δίνεται ότι τα δεδομένα μου ακολουθούν την κανονική κατανομή. Πάμε να τα εισάγουμε στην *R*:

```
acid <- c(45,45,34,47,43,28,46,41,21,46,37,16)
```

```
Temp <- rep(c(0,10,20), times = 4)
Temp <- factor(Temp)
weeks <- rep(c(2,4,6,8), each = 3)
```

Α. Προσαρμόστε ένα μοντέλο για τη μέση συγκέντρωση ασκορβικού οξέος μιας τριάδας πακέτων με τη βοήθεια της θερμοκρασίας και του χρόνου αποθήκευσης. Το μοντέλο θα υποθέτει μια σταθερά και μια γραμμική επίδραση των εβδομάδων αποθήκευσης για κάθε θερμοκρασία.

Το μοντέλο μας θα είναι απλά το αθροιστικό, δηλαδή θα είναι της μορφής:

$$acid_{i,weeks} = \beta_0 + \beta_1 \cdot Temp_{i2} + \beta_2 \cdot Temp_{i3} + \beta_3 \cdot weeks \qquad M_1$$

Όπου:

$$Temp_{ij} = \begin{cases} 0, & \alpha v\ i \neq j \\ 1, & \alpha v\ i = j \end{cases}$$

Οπότε πάμε να το προσαρμόσουμε στην *R*:

```
mod1 <- glm(acid ~ Temp + weeks, family = gaussian)
summary(mod1)
```

```
Call:
glm(formula = acid ~ Temp + weeks, family = gaussian)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-5.2500  -1.1458  -0.0833   1.5208   5.0000

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
```

$(Intercept)$ 53.083 2.934 18.090 $8.95e-08$ $***$

$Temp10$ -4.500 2.541 -1.771 0.1146

$Temp20$ -21.250 2.541 -8.362 $3.17e-05$ $***$

$weeks$ -1.417 0.464 -3.053 0.0157 $*$

$---$

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

($Dispersion\ parameter\ for\ gaussian\ family\ taken\ to\ be\ 12.91667$)

$Null\ deviance$: 1226.92 on 11 $degrees\ of\ freedom$

$Residual\ deviance$: 103.33 on 8 $degrees\ of\ freedom$

AIC: 69.891

$Number\ of\ Fisher\ Scoring\ iterations$: 2

Β. Δώστε ερμηνεία στις παραμέτρους του μοντέλου.

Για το β_0: Αν είμαστε στον χρόνο 0, τότε η αναμενόμενη συγκέντρωση του ασκορβικού οξέος σε θερμοκρασία συντήρησης 0 βαθμών, εκτιμάται να είναι:

$$acid_{1,0} = \beta_0 = 53.083$$

Για το β_1: Αν είμαστε στον χρόνο 0, τότε η αναμενόμενη συγκέντρωση του ασκορβικού οξέος καθώς η θερμοκρασία συντήρησης αυξάνει από 0 βαθμούς σε 10, εκτιμάται να μειωθεί κατά 4.5 μονάδες.

Για το β_2: Αν είμαστε στον χρόνο 0, τότε η αναμενόμενη συγκέντρωση του ασκορβικού οξέος καθώς η θερμοκρασία συντήρησης αυξάνει από 0 βαθμούς σε 20, εκτιμάται να μειωθεί κατά 21.25 μονάδες.

Για το β_3: Αν ο χρόνος συντήρησης αυξηθεί κατά μία μονάδα, με όλες τις άλλες παραμέτρους να παραμένουν σταθερές, τότε αναμένουμε μείωση της συγκέντρωσης του ασκορβικού οξέος κατά 1.417 μονάδες.

Γ. Εκτιμήστε την αρχική συγκέντρωση (στον χρόνο 0) ασκορβικού οξέος μιας τριάδας πακέτων όταν η θερμοκρασία είναι 20 βαθμοί και δώσατε 95% Διάστημα Εμπιστοσύνης.

Αυτό γίνεται στην R ως εξής:

```
new <- data.frame(Temp = factor(20),weeks = 0)
pred <- predict(mod1,newdata = new,type = 'response',se.fit = TRUE)
pred
$fit
       1
31.83333

$se.fit
[1] 2.934469

$residual.scale
[1] 3.593976
```

Ενώ το διάστημα εμπιστοσύνης 95% θα βρεθεί από:

```
confint(mod1)
Waiting for profiling to be done...
              2.5 %     97.5 %
(Intercept) 47.331879 58.8347878
Temp10      -9.480906  0.4809057
Temp20     -26.230906 -16.2690943
```

```
weeks       −2.326051   −0.5072819
```

Όπου για την τιμή της θερμοκρασία να είναι 20, αναμένουμε μείωση της συγκέντρωσης του ασκορβικού οξέος κάπου ανάμεσα στο διάστημα του:

$$[-26.230906, -16.2690943]$$

Δηλαδή η αναμενόμενη τιμή της συγκέντρωσης του ασκορβικού οξέος θα είναι κάπου στο διάστημα:

$$[26.85209, \quad 36.81391]$$

Δ. Δώστε το τυπικό σφάλμα της πρόβλεψης του $(\Gamma.)$ καθώς και το προσεγγιστικό 95% Διάστημα Εμπιστοσύνης.

Το τυπικό σφάλμα πρόβλεψης είναι το:

```
pred$se.fit

[1] 2.934469
```

Το προσεγγιστικό 95% Διάστημα Εμπιστοσύνης είναι:

$$\left[\hat{\beta}_0 - t\ value \cdot standard\ error, \quad \hat{\beta}_0 + t\ value \cdot standard\ error\right]$$

Αυτό στην R θα είναι:

```
53.083 − 21.250 − 2.541 ∗ (−8.362)

[1] 53.08084

53.083 − 21.250 + 2.541 ∗ (−8.362)

[1] 10.58516
```

Άρα ίσως να είναι το:

$$[10.58516, \quad 53.08084]$$

Ε. Ποιά είναι η εκτίμηση της διακύμανσης των δεδομένων που λαμβάνετε από το μοντέλο που χρησιμοποιήσατε στο $(\boldsymbol{A}.)$;

$$Variance = \frac{Residual\ Sum\ of\ Squares}{n-p} = \frac{103.33}{8} = 12.91625$$

ΣΤ. Προσθέστε την αλληλεπίδραση εβδομάδων αποθήκευσης και θερμοκρασίας στο μοντέλο ελέγξτε τη στατιστική του σημαντικότητα και σχολιάστε την επάρκεια του νέου μοντέλου κάνοντας οπτικό έλεγχο καταλοίπων.

Το νέο μας μοντέλο θα είναι της μορφής:

$$acid_{i,weeks} = \beta_0 + \beta_1 \cdot Temp_{i2} + \beta_2 \cdot Temp_{i3} + \beta_3 \cdot weeks + \beta_4 \cdot Temp_{i2} \cdot weeks + \quad M_2$$
$$+ \beta_5 \cdot Temp_{i3} \cdot weeks$$

Όπου:

$$Temp_{ij} = \begin{cases} 0, & \alpha v\ i \neq j \\ 1, & \alpha v\ i = j \end{cases}$$

Πάμε να το προσαρμόσουμε στην *R*:

```
mod2 <- glm(acid ~ Temp * weeks, family = gaussian)
summary(mod2)


Call:
glm(formula = acid ~ Temp * weeks, family = gaussian)


Deviance Residuals:
   Min      1Q  Median      3Q     Max
-0.700  -0.450   0.000   0.250   1.100


Coefficients:
                Estimate  Std.Error  t value  Pr(>|t|)
```

```
(Intercept)       45.5000    0.9618    47.309  5.98e-09 ***
Temp10             2.5000    1.3601     1.838  0.11569
Temp20            -5.5000    1.3601    -4.044  0.00677 **
weeks              0.1000    0.1756     0.569  0.58969
Temp10:weeks      -1.4000    0.2483    -5.638  0.00133 **
Temp20:weeks      -3.1500    0.2483   -12.685  1.47e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 0.6166667)

    Null deviance: 1226.9  on 11  degrees of freedom
Residual deviance:    3.7  on  6  degrees of freedom
AIC: 33.936

Number of Fisher Scoring iterations: 2
```

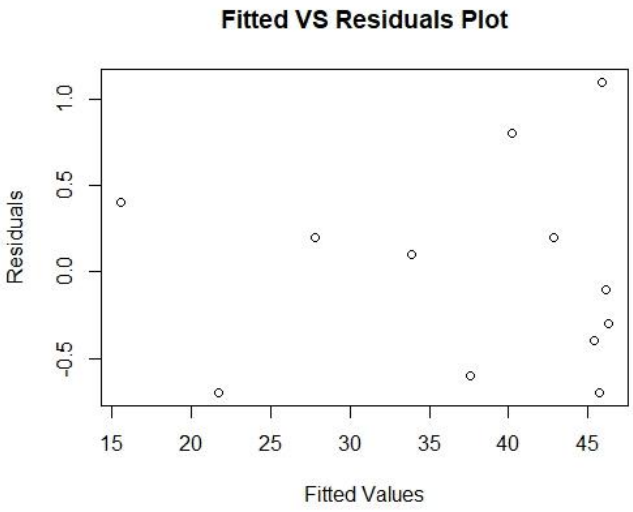
Ας δούμε ποιό είναι το καλύτερο από τα δύο μοντέλα:

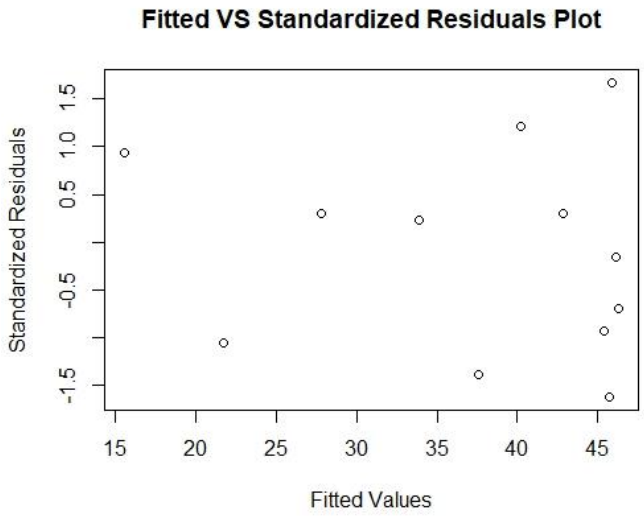
```
AIC(mod1)
[1] 69.89117
AIC(mod2)
[1] 33.93564
```

Το μοντέλο 2 είναι προφανώς καλύτερο από το πρώτο. Ας ελέγξουμε και την καλή του προσαρμογή μέσω διαγραμμάτων:

```
plot(fitted(mod2), resid(mod2, type = "pearson"), xlab = "Fitted Values", ylab
     = "Residuals", main = "Fitted VS Residuals Plot")
```

```
plot(fitted(mod2), rstandard(mod2, type = "pearson"), xlab
        = "Fitted Values", ylab = "Standardized Residuals", main
        = "Fitted VS Standardized Residuals Plot")
```





Αρκετά καλό fit, με διάσπαρτες τιμές και χωρίς *outliers*.

Άσκηση 3:

Σε μια έρευνα για τη συσχέτιση των επιπέδων πίεσης αίματος και χοληστερόλης μετείχαν 1267 υποκείμενα. Λάβαμε τα παρακάτω αποτελέσματα:

	Επίπεδα Πίεσης Αίματος		
Επίπεδα Χοληστερόλης	<130	130-155	155+
<200	117	168	22
200-219	85	141	20
220-259	119	277	43
>=260	67	145	33

Το ενδιαφέρον μας βρίσκεται στην πιθανότητα $P(επίπεδα\ πίεσης \mid επίπεδα\ χοληστερόλης)$. Προτείνετε μια λογική κωδικοποίηση για τα επίπεδα χοληστερόλης. Θεωρήστε τα επίπεδα χοληστερόλης ως μια συνεχή μεταβλητή.

Α. Προσαρμόστε ένα μοντέλο αθροιστικών logits με proportional odds με την κύρια επίδραση των επιπέδων χοληστερόλης. Είναι η επίδραση του επιπέδου χοληστερόλης στατιστικά σημαντική;

Β. Ερμηνεύστε τον συντελεστή της επίδρασης της χοληστερόλης.

Γ. Ελέγξτε την καλή προσαρμογή του μοντέλου.

Δ. Στο μοντέλο που προσαρμόσατε στο (Α.) κάντε έλεγχο παραλληλίας γράφοντας καθαρά ποιά είναι τα μοντέλα που συγκρίνονται μεταξύ τους και τη μηδενική υπόθεση.

Ε. Προβλέψτε την πιθανότητα κάθε ενός από τα επίπεδα αίματος για ένα υποκείμενο επίπεδο χοληστερόλης 270.

ΛΥΣΗ:

Επειδή τα επίπεδα πίεσης αίματος καθώς και τα επίπεδα χοληστερόλης δίνονται σε διαστήματα, μπορώ να προχωρήσω με δύο τρόπους:

- Πρώτον, να πάρω τις μέσες τιμές των διαστημάτων και να βάλω αυτά για τα επίπεδα πίεσης αίματος και τα επίπεδα χοληστερόλης. Παρατηρήστε όμως εδώ, ότι έχουμε τα διαστήματα με μία τιμή μόνο όπως το πρώτο και το τελευταίο διάστημα.
- Δεύτερον, να ορίσω το κατώτερο όριο του διαστήματος ως την τιμή του κάθε επιπέδου.

Αυτό που θα κάνω λοιπόν εδώ είναι το δεύτερο, δηλαδή θα υποθέσω ότι το κατώτερο νούμερο ορίζει τα επίπεδα πίεσης αίματος καθώς και τα επίπεδα χοληστερόλης. Δεν μπορούμε να έχουμε αρνητική χοληστερόλη ή επίπεδα πίεσης, οπότε υποθέτω ότι η κατώτερη τιμή είναι το μηδέν. Άρα ο παραπάνω πίνακας γίνεται:

	Επίπεδα Πίεσης Αίματος		
Επίπεδα Χοληστερόλης	0	130	155
0	117	168	22
200	85	141	20
220	119	277	43
260	67	145	33

Τα δεδομένα μας είναι κατηγορικά, γιατί χωρίζονται σε κατηγορίες, καθώς και ομαδοποιημένα.

Πάμε να τα εισάγουμε στην *R*:

```
cholesterol <- c(0,200,220,260)
resp <- matrix(c(117,85,119,67,168,141,277,145,22,20,43,33),ncol = 3)
resp
     [,1] [,2] [,3]
[1,]  117  168   22
[2,]   85  141   20
[3,]  119  277   43
[4,]   67  145   33
```

Α. Προσαρμόστε ένα μοντέλο αθροιστικών logits με proportional odds με την κύρια επίδραση των επιπέδων χοληστερόλης. Είναι η επίδραση του επιπέδου χοληστερόλης στατιστικά σημαντική;

```
library(car)
library(VGAM)
```

Το μοντέλο που θα προσαρμόσουμε θα είναι της μορφής:

$$logit(\pi_{j,cholesterol}) = \beta_{0,j} + \beta_1 \cdot cholesterol, \qquad με\ j = 2, 3. \qquad M_1$$

Έχουμε δηλαδή την ίδια επεξηγηματική μεταβλητή για κάθε ένα από τα δύο μοντέλα. Όπου:

$$\pi_{j,cholesterol} = P(επίπεδα\ πίεσης \mid επίπεδα\ χοληστερόλης)$$

Ή με άλλα λόγια:

$$log\left(\frac{\pi_{1,cholesterol}}{\pi_{2,cholesterol}+\pi_{3,cholesterol}}\right)=\beta_{0,2}+\beta_1\cdot cholesterol$$

$$log\left(\frac{\pi_{1,cholesterol}+\pi_{2,cholesterol}}{\pi_{3,cholesterol}}\right)=\beta_{0,3}+\beta_1\cdot cholesterol$$

Αυτό στην *R* γίνεται ως εξής:

```
mod1 <- vglm(resp ~ cholesterol, cumulative(parallel = T))
summary(mod1)

Call:
vglm(formula = resp ~ cholesterol, family = cumulative(parallel = T))

Coefficients:
                Estimate  Std.Error  z value  Pr(>|z|)
(Intercept):1 -0.4599708  0.1113983  -4.129  3.64e-05 ***
(Intercept):2  2.5936677  0.1409108   18.406  < 2e-16 ***
cholesterol   -0.0019483  0.0005673  -3.434  0.000595 ***
---
Signif.codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Names of linear predictors: logitlink(P[Y <= 1]),
logitlink(P[Y <= 2])

Residual deviance: 5.5349 on 5 degrees of freedom
```

$$Log-likelihood: \ -24.8302 \ on \ 5 \ degrees \ of \ freedom$$

$$Number \ of \ Fisher \ scoring \ iterations: 3$$

$$No \ Hauck-Donner \ effect \ found \ in \ any \ of \ the \ estimates$$

$$Exponentiated \ coefficients:$$

$$cholesterol$$

$$0.9980536$$

Β. Ερμηνεύστε τον συντελεστή της επίδρασης της χοληστερόλης.

Όταν η χοληστερόλη αυξάνεται κατά μία μονάδα, τότε ο λόγος των πιθανοτήτων (Odds Ratio) μεταξύ αυτών που έχουν χοληστερόλη και αυτών που δεν έχουν, είναι: $exp\{-0.20448\} = 0.815071$.

Γ. Ελέγξτε την καλή προσαρμογή του μοντέλου.

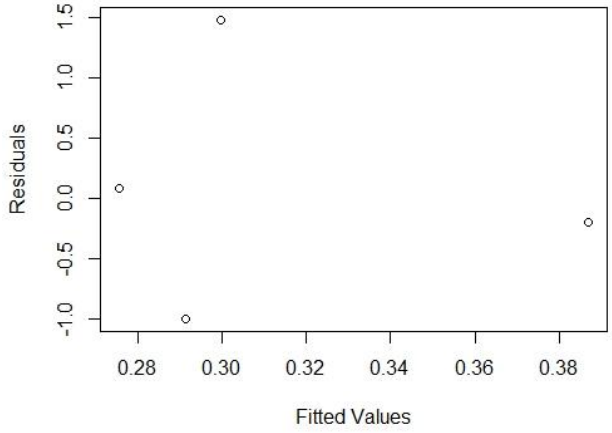
Εφόσον τα δεομένα μας είναι grouped, θα κάνουμε τους κλασσικούς τρεις ελέγχους:

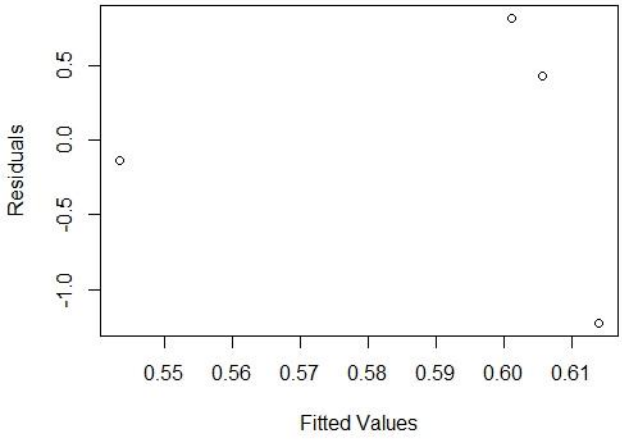
1. Έλεγχος μέσω διαγραμμάτων:

```
plot(fitted(mod1)[,1], resid(mod1, type = "pearson")[,1], xlab
        = "Fitted Values", ylab = "Residuals")

plot(fitted(mod1)[,2], resid(mod1, type = "pearson")[,2], xlab
        = "Fitted Values", ylab = "Residuals")
```

Οπότε τα διαγάμματε θα είναι:





‘Οπως φαίνεται, έχουμε καλή προσαρμογή των δεδομένων.

2. Έλεγχος με deviance:

```
pchisq(deviance(mod1), df = mod1@df.residual, lower.tail = FALSE)
```

```
[1] 0.3541358
```

Άρα για $\alpha = 0.05$, δεν έχουμε αρκετά στοιχεία για να απορρίψουμε την μηδενική υπόθεση (διότι $p - value > a \Longrightarrow 0.3541358 > 0.05$) και έτσι μπορούμε να πούμε ότι το μοντέλο προσαρμόζει καλά τα δεδομένα, ή με άλλα λόγια, ότι το μοντέλο αυτό δεν διαφέρει πολύ από το κορεσμένο (Saturated) μοντέλο.

3. Έλεγχος με κατάλοιπα Pearson:

```
pchisq(sum(resid(mod1, type = "pearson")^2), df
                = mod1@df.residual, lower.tail = FALSE)
```

```
[1] 0.3436032
```

Άρα για $\alpha = 0.05$, δεν έχουμε αρκετά στοιχεία για να απορρίψουμε την μηδενική υπόθεση (διότι $p - value > a \Longrightarrow 0.3436032 > 0.05$) και έτσι μπορούμε να πούμε ότι το μοντέλο προσαρμόζει καλά τα δεδομένα, ή με άλλα λόγια, ότι το μοντέλο αυτό δεν διαφέρει πολύ από το κορεσμένο (Saturated) μοντέλο.

Δ. Στο μοντέλο που προσαρμόσατε στο (Α.) κάντε έλεγχο παραλληλίας γράφοντας καθαρά ποιά είναι τα μοντέλα που συγκρίνονται μεταξύ τους και τη μηδενική υπόθεση.

Εδώ υποθέσαμε αναλογικότητα, αλλά αν δεν ελιχαμε υποθέσει, τότε το μοντέλο μας θα ήταν:

$$logit\left(\pi_{j,cholesterol}\right) = \beta_{0,j} + \beta_{1,j} \cdot cholesterol, \qquad με\ j = 2, 3. \qquad M_2$$

Αυτό που θέλουμε να ελέγξουμε είναι το ποιά από τις παρακάτω υποθέσεις ισχύει:

$$H_0: \beta_{1,2} = \beta_{1,3} = \beta_1$$

$$H_1: Όχι\ H_0$$

Δηλαδή το να έχουμε ή όχι αναλογικότητα (proportionality). Θα συγκρίνουμε τα δυο μοντέλα στην R.

```
mod2 <- vglm(resp ~ cholesterol, cumulative)
summary(mod2)

Call:
vglm(formula = resp ~ cholesterol, family = cumulative)

Coefficients:
                  Estimate  Std.Error  z value  Pr(>|z|)
(Intercept):1 -0.467149   0.115802   -4.034  5.48e-05 ***
(Intercept):2  2.632340   0.223163    11.796  < 2e-16 ***
cholesterol:1 -0.001902   0.000602   -3.160   0.00158 **
cholesterol:2 -0.002156   0.001085   -1.988   0.04686 *
---
Signif.codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Names of linear predictors: logitlink(P[Y <= 1]),
logitlink(P[Y <= 2])
```

$$\boldsymbol{Residual\ deviance}: \mathbf{5.4833\ on\ 4\ degrees\ of\ freedom}$$

$$\boldsymbol{Log-likelihood}: \mathbf{-24.8043\ on\ 4\ degrees\ of\ freedom}$$

$$\boldsymbol{Number\ of\ Fisher\ scoring\ iterations}: \mathbf{3}$$

$$\boldsymbol{Warning}: \boldsymbol{Hauck-Donner\ effect\ detected\ in\ the\ following\ estimate(s)}:$$

$$'(\boldsymbol{Intercept}): \mathbf{2}'$$

$$\boldsymbol{Exponentiated\ coefficients}:$$

$$\boldsymbol{cholesterol}: \mathbf{1} \quad \boldsymbol{cholesterol}: \mathbf{2}$$

$$\mathbf{0.9980995} \qquad \mathbf{0.9978461}$$

Πάμε να κάνουμε τον έλεγχο Likelihood Ratio για σύγκριση μεταξύ δύο μοντέλων μέσω των αποκλίσεών τους (*deviances*).

$$\boldsymbol{pchisq}(\boldsymbol{deviance}(\boldsymbol{mod1}) - \boldsymbol{deviance}(\boldsymbol{mod2}), \boldsymbol{df} = \boldsymbol{mod1@df.residual} - \boldsymbol{mod2@df.residual}, \boldsymbol{lower.tail} = \boldsymbol{FALSE})$$

$$[\mathbf{1}]\ \mathbf{0.8202424}$$

Επομένως σε επίπεδο στατιστικής σημαντικότητας $\alpha = 0.05$, δεν έχουμε αρκετά στοιχεία για να απορρίψουμε την μηδενική υπόθεση (διότι $p - value > a \Longrightarrow 0.8202424 > 0.05$) και έτσι μπορούμε να πούμε ότι το μοντέλο μας είναι αναλογικό (*proportional*). Οπότε ισχύει η υπόθεση παραλληλίας.

Ε. Προβλέψτε την πιθανότητα κάθε ενός από τα επίπεδα αίματος για ένα υποκείμενο επίπεδο χοληστερόλης 270.

Αυτό θα γίνει ως εξής:

```
new <- data.frame(cholesterol = 270)
predict(mod1, newdata = new, type = "response")
        mu1       mu2       mu3
1 0.2717008 0.6160159 0.1122833
```

Αυτές είναι λοιπόν οι πιθανότητες κάθε ενός από τα επίπεδα αίματος για ένα υποκείμενο επίπεδο χοληστερόλης 270.