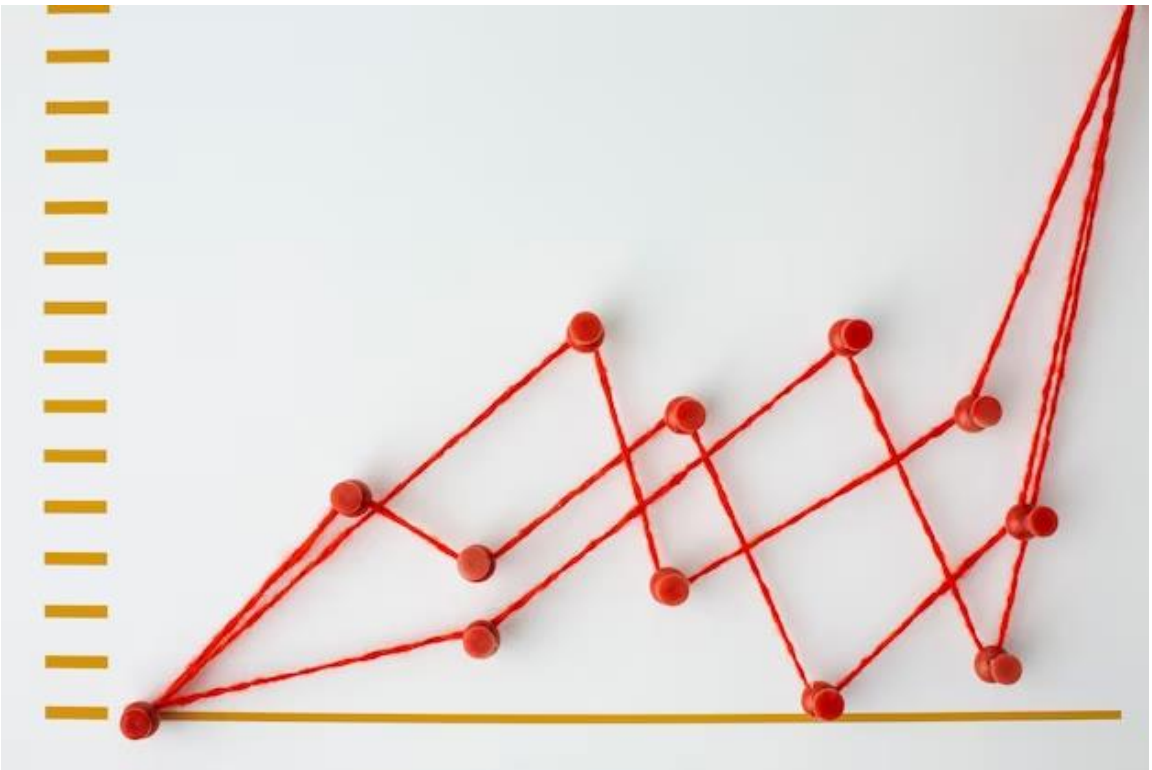
Γενικευμένα Γραμμικά Μοντέλα
Στατιστική μοντελοποίηση μέσω της γλώσσας προγραμαμτισμού R
Παραδείγματα κάποιων Πρακτικών Εφαρμογών

Χατζόπουλος Γεράσιμος



Contents

ΠΑΡΑΔΕΙΓΜΑ ΠΡΩΤΟ (BINARY ΔΕΔΟΜΕΝΑ)

Τα δεδομένα της σελίδας δείχνουν τον αριθμό των εντόμων που πέθαναν μετά από έκθεση 5 ωρών στις δόσεις κάποιας τοξικής ουσίας.
Σκοπός της εταιρίας είναι να μοντελοποιήσει τα δεδομένα εκφράζοντας την γραμμική τάση της τδόσης της τοξικής ουσίας και να προβλεφτεί βάση αυτό το μοντέλο η πιθανότητα θανάτου του εντόμους σε συγκεκριμένες δοσολογίες (1.7709, 1.8403 και 1.8865) που σκοπεύει να ρυθμίσει για δοσολογία στο προϊόν της

Δόση τοξικής ουσίας	Αρ. εντόμων	Αρ. θανάτων
1.6907	59	6
1.7242	60	13
1.7552	62	18
1.7842	56	28
1.8113	63	52
1.8369	59	53
1.8610	62	61
1.8839	60	60

Λύση:

Πρώτα από όλα θα φτιάξουμε τον πίνακα αυτόν στην *R*:

Στήλη 1:

```
dose <- c(1.6907, 1.7242, 1.7552, 1.7842, 1.8113, 1.8369, 1.8610, 1.8839)
dose
[1] 1.6907 1.7242 1.7552 1.7842 1.8113 1.8369 1.8610 1.8839
```

Στήλη 2:

```
n.bugs <- c(59, 60, 62, 56, 63, 59, 62, 60)
n.bugs
[1] 59 60 62 56 63 59 62 60
```

Στήλη 3:

```
n.deaths <- c(6, 13, 18, 28, 52, 53, 61, 60)
n.deaths
[1]  6 13 18 28 52 53 61 60
```

Πίνακας (Matrix):

```
m1 <- cbind(dose, n.bugs, n.deaths)
m1
      dose   n.bugs n.deaths
[1,] 1.6907      59        6
[2,] 1.7242      60       13
[3,] 1.7552      62       18
[4,] 1.7842      56       28
[5,] 1.8113      63       52
[6,] 1.8369      59       53
[7,] 1.8610      62       61
[8,] 1.8839      60       60
```

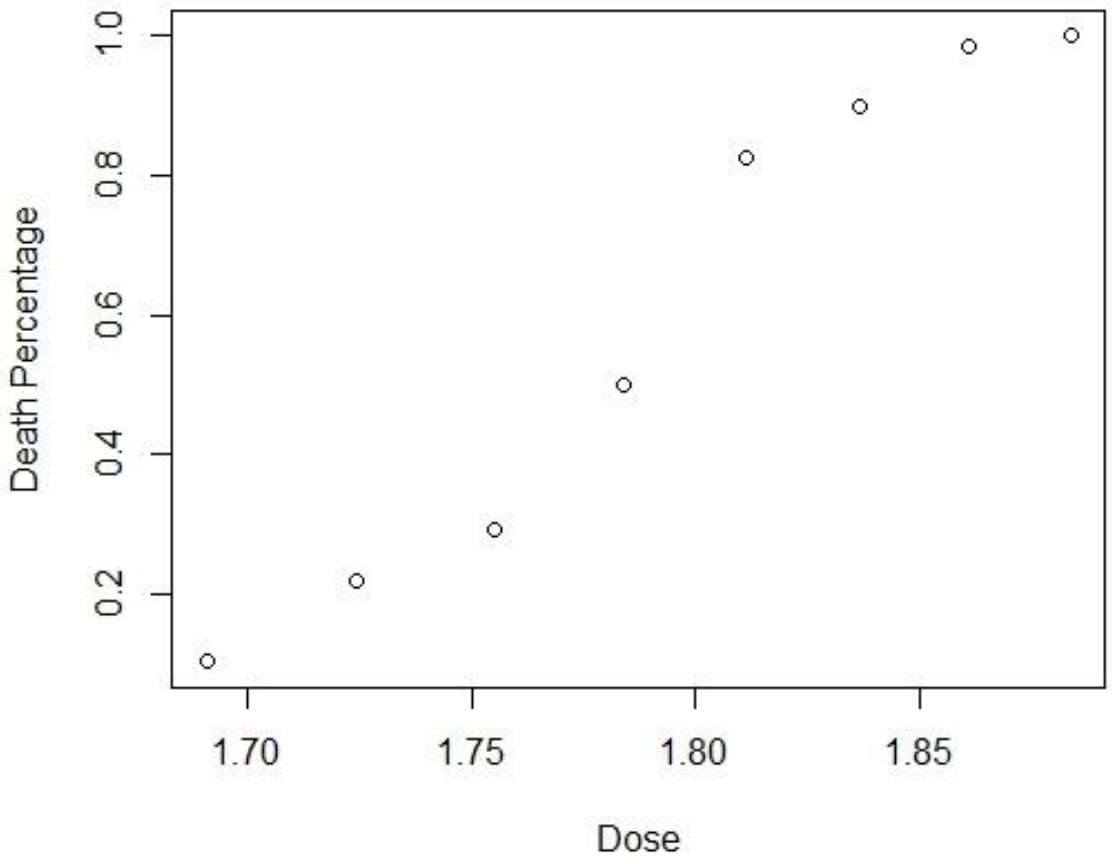
Για να πάρουμε αρχικά μία πρώτη οπτική εικόνα θα μπορούσαμε να δημιουργήσουμε ένα διάγραμμα των ποσοστών θανάτου έναντι των δόσεων της τοξικής ουσίας.
Θέλουμε το ποσοστό των θανάτων, που είναι ο αριθμός των θανάτων (η τρίτη στήλη της μήτρας) διαιρεμένο με το συνολικό αριθμό των εντόμων (η δεύτερη στήλη της μήτρας). Δηλαδή:

Ποσοστά Θανάτου:

```
y1 <- m1[,3]/m1[,2]
y1
[1] 0.1016949 0.2166667 0.2903226 0.5000000 0.8253968
                            [6] 0.8983051 0.9838710 1.0000000
```

Και τώρα θέλουμε το διάγραμμα των ποσοστών θανάτου έναντι των δόσεων της τοξικής ουσίας, που είναι:

```
plot(m1[,1],y1,xlab = "Dose",ylab = "Death Percentage")
```



Παρατήρηση

Βλέπουμε πως δεν φαίνεται να είναι γραμμική η σχέση των παρατηρήσεων, γιατί τα δεδομένα μας δεν μπαίνουν σε μια ευθεία γραμμή, αλλά μάλλον τετραγωνική (*quadratic*).

Αρχικά πριν περάσουμε στην μοντελοποίηση για να είναι περισσότερο ερμηνεύσημοι οι συντελεστές των μοντέλων θα μετατρέψουμε τις δόσεις πολλαπλασιάζοντάς τες με 100.

```
dose1 <- dose * 100
dose1
[1] 169.07 172.42 175.52 178.42 181.13 183.69 186.10 188.39
```

Θα ποροσαρμόσουμε ένα μοντέλο λογιστικής παλινδρόμησης που να εκφράζει τη γραμμική τάση της δόσης τοξικής ουσίας.

Το μοντέλο θα είναι της μορφής:

$$logit(\pi_i) = log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 \cdot x_i\,, \qquad i = 1, 2, \dots, 8$$

Γιατί υποθέτω ότι η δόση του φαρμάκου είναι μια συνεχόμενη μεταβλητή. Επιπλέον θεωρώ ότι:

$$n_i \cdot y_i \sim Binomial(n_i, \pi_i)$$

Όπου το να πεθάνει ένα έντομο θεωρείται **«επιτυχία»**, ενώ να επιβιώσει, **«αποτυχία»**.

Προσαρμόζουμε το μοντέλο:

```
mod.1 <- glm(yy~dose1, binomial)
mod.1

Call: glm(formula = yy ~ dose1, family = binomial)

Coefficients:
(Intercept)       dose1
   -60.7175       0.3427

Degrees of Freedom: 7 Total (i.e. Null);  6 Residual
Null Deviance:    284.2
Residual Deviance: 11.23        AIC: 41.43
```

Η σύνοψη του μοντέλου είναι:

```
summary(mod. 1)

Call:
glm(formula = yy ~ dose1, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5941  -0.3944   0.8329   1.2592   1.5940

Coefficients:
            Estimate                 Std. Error   z valu
(Intercept)  -60.71745     5.18070   -11.72
dose1          0.34270
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 284.202  on 7  degrees of freedom
Residual deviance:   11.232  on 6  degrees of freedom
AIC: 41.43

Number of Fisher Scoring iterations: 4
```

Έχουμε το μοντλεο μας αλλά πρώτα θα πρέπει να ελένξουμε την καλή προσαρμογη του (goodness of fit), δηλαδή αμά το συγκεκριμένο μοντέλο εγαρμόζει και εξηγεί σψστά στατιστικά τις μεταβλητές μας

Τα δεδομένα μας εδώ είναι ομαδοποιημένα (grouped). Άρα για να ελέγξουμε την καλή προσαρμογή του μοντέλου, μπορούμε να χρησιμοποιήσουμε:

1. Είτε τον έλεγχο καταλοίπων Pearson (πράγμα που γίνεται πιο μετά στο ερώτημα
2. Είτε τον έλεγχο Likelihood Ratio Test για την απόκλιση (deviance), για να δούμε αν το μοντέλο που προσαρμόσαμε είναι το ίδιο με το κορεσμένο (Saturated). Δηλαδή να ελέγξουμε αν:

$$H_0: M_1 = Saturated$$
$$H_1: M_1 \neq Saturated$$

Πράγμα που θα κάνουμε τώρα:

Ξέρουμε ότι εάν το μοντέλο προσαρμόζεται καλά στα δεδομένα, τότε η απόκλιση θα πρέπει να ακολουθεί περίπου την κατανομή $X^2_{N-p} = X^2_6$ επειδή υπάρχουν $N = 8$ covariate patterns (δηλαδή, διαφορετικές τιμές του x_i) και $p = 2$ παράμετροι.

Υπολογίζουμε την $p - value$ του:

```
pchisq(deviance(mod.1), df = mod.1$df.residual, lower.tail = FALSE)
[1] 0.08145881
```

Άρα για $\alpha = 0.05$ δεν έχουμε αρκετά στοιχεία για να απορρίψουμε την μηδενική υπόθεση 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.

Πρόσθεση τετραγωνικού ώρου στο μοντέλο μας (quadratic)

Το νέο μοντέλο θα είναι της μορφής:

$$logit(\pi_i) = log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 \cdot x_\iota + \beta_2 \cdot (x_\iota)^2, \quad i = 1, 2, \dots, 8$$

Στην R, αυτό γίνεται ως εξης:

```
mod.2 <- glm(yy~dose1 + I(dose1^2), binomial)
mod.2

Call:  glm(formula = yy ~ dose1 + I(dose1^2), family = binomial)

Coefficients:
(Intercept)        dose1    I(dose1^2)
 431.10580     -5.20615      0.01564

Degrees of Freedom: 7 Total (i.e. Null);  5 Residual
Null Deviance:    284.2
Residual Deviance: 3.195        AIC: 35.39
```

Ελέγγουμε τη στατιστική του σημαντικότητα και την καλή του προσαρμογή.

Δίνουμε την εντολή summary:

```
summary(mod. 2)
Call:
glm(formula = yy ~ dose1 + I(dose1^2), family = binomial)

Deviance Residuals:
      1       2       3       4       5       6
-0.4215  0.8189  -0.2769  -0.5232  0.8746  -0.8249
      7       8
0.1698  0.7224

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

(Intercept) 431.105804  180.653557    2.386   0.01702 *
dose1        -5.206153    2.045225   -2.546   0.01091 *
I(dose1^2)    0.015641    0.005786    2.703   0.00687 **
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 284.2024  on 7  degrees of freedom
Residual deviance:   3.1949  on 5  degrees of freedom
AIC: 35.393

Number of Fisher Scoring iterations: 4
```

Στατιστική σημαντικότητα των παραμέτρων:
α' τρόπος, Wald Test:
Το Wald Test ελέγχει την υπόθεση ότι:

$$H_0: \beta_i = 0$$
$$H_1: \beta_i \neq 0$$

Για κάθε μία παράμετρο.

Θα το ελέγξουμε από τον πίνακα Coefficients παραπάνω, όπου το $z - score$ είναι το

$$\sqrt{Wald - Statistic} \sim N(0,1)$$

Βλέπουμε από τα p-values στα δεξιά ότι:
α) Για την σταθερά (β_0). Σε επίπεδο στατιστικής σημαντικότητας $\alpha = 0.05$, απορρίπτουμε την μηδενική υπόθεση, διότι $(\boldsymbol{p - value < \alpha \Longrightarrow 0.01702 < 0.05})$, επομένως ο όρος είναι στατιστικά σημαντικός στο μοντέλο αυτό.
β) Για τον όρο (β_1). Σε επίπεδο στατιστικής σημαντικότητας $\alpha = 0.05$, απορρίπτουμε την μηδενική υπόθεση, διότι $(\boldsymbol{p - value < \alpha \Longrightarrow 0.01091 < 0.05})$, επομένως και αυτός ο όρος είναι στατιστικά σημαντικός στο μοντέλο μας.
γ) Για τον όρο (β_2). Σε επίπεδο στατιστικής σημαντικότητας $\alpha = 0.05$, απορρίπτουμε την μηδενική υπόθεση, διότι $(\boldsymbol{p - value < \alpha \Longrightarrow 0.00687 < 0.05})$, επομένως και αυτός ο όρος είναι στατιστικά σημαντικός στο μοντέλο μας.

Παρατηρούμε ότι όλοι οι όροι του μοντέλου M_2 είναι στατιστικά σημαντικοί

Ποιο μοντέλο θα πρέπει να επιλέξω; Το M_1 (απλό) η το M_2 (τετραγωνικό όρο); Ένας τρόπος είναι με το πληροφοριακό κριτήριο AIC και BIC (θα μπορούσε να γίνει και με άλλους τρόπους π.χ με Likelihood Ratio Test).

Οι τιμές των AIC και BIC, θα είναι:

```
AIC(mod.1)
[1] 41.43027
AIC(mod.2)
[1] 35.39294
BIC(mod.1)
[1] 41.58915
BIC(mod.2)
[1] 35.63127
```

Θέλουμε το μοντέλο με τις χαμηλότερες τιμές στα AIC και BIC και έτσι θα προτιμήσουμε το μοντέλο 2 (M_2).

Το ερώτημα της εταιρίας είναι ποια δοσολογία να επιλέξει ώστε να βρει την χρυσή τομή μεταξύ πιθανότητας θανάτου του εντόμου και δοσολογίας Έχει τρεις προτάσεις ως προς την δοσολογία, το έντομο να δεχτεί δόση 1.7709, 1.8403 και 1.8865.

Αποφασίσαμε ότι το μοντέλο M_2 είναι το καλύτερο από τα δύο, άρα θα χρησιμοποιήσω αυτό για την πρόβλεψη. Πριν το κάνω όμως, θα πρέπει να θυμόμαστε ότι κάναμε μια μετατροπή στις δόσεις. Τις πολλαπλασιάσαμε με το 100, για να είναι πιο ερμηνεύσιμοι οι συντελεστές. Οπότε και εδώ θα χρειαστεί να πολλαπλασιάσουμε αυτές τις δόσεις για πρόβλεψη επί 100, για να δουλέψει ο αλγόριθμος.

Πρώτα φτιάχνουμε ένα data frame με ολες τις τιμές που θέλουμε να προβλέψουμε:

```
newdose <- c(177.09, 184.03, 188.65)
new <- data.frame(dose1 = newdose)
new
     dose1
1   177.09
2   184.03
3   188.65
```

Και ύστερα κάνουμε πρόβλεψη της πιθανότητας:

```
predict(mod.2, newdata = new, type = "response")
        1         2         3
0.4178742 0.9391884 0.9963721
```

Άρα συμπεραίνουμε ότι:

- Αν χορηγηθεί δόση φαρμάκου ίση με 1.7709, τότε προβλέπουμε ότι η πιθανότητα του να πεθάνει το έντομο, θα είναι: 41.78742%.
- Αν χορηγηθεί δόση φαρμάκου ίση με 1.8403, τότε προβλέπουμε ότι η πιθανότητα του να πεθάνει το έντομο, θα είναι: 93.91884%.
- Αν χορηγηθεί δόση φαρμάκου ίση με 1.8865, τότε προβλέπουμε ότι η πιθανότητα του να πεθάνει το έντομο, θα είναι: 99.63721%.

BONUS (ΜΗ ΟΜΑΔΟΠΟΙΗΜΕΝΟ ΜΟΝΤΕΛΟ): Θα μπορούσα ακόμα για σφάλεια να κάνω και έναν έλεγχο για την προβλεπτική ικανότητα του παραπάνω μοντέλου μας με έναν έλεγχο Hosmer-Lemeshow.

Για να κάνω κάτι τέτοιο, θα χρειαστεί πρώτα να κάνω τα δεδομένα μου μη ομαδοποιημένα (ungrouped).
Για την δόση:

```
dose <- c(rep(1.6907,59),rep(1.7242,60),rep(1.7552,62),rep(1.7842,56),
          rep(1.8113,63),rep(1.8369,59),rep(1.8610,62),rep(1.8839,60))
table(dose)
dose
1.6907 1.7242 1.7552 1.7842 1.8113 1.8369  1.861 1.8839
    59     60     62     56     63     59     62     60
```

Για τον αριθμό επιτυχιών-αποτυχιών για κάθε δόση:

```
y <- c(rep(1,6),rep(0,53),rep(1,13),rep(0,47),rep(1,18),rep(0,44),
       rep(1,28),rep(0,28),rep(1,52),rep(0,11),rep(1,53),rep(0,6),
       rep(1,61),rep(0,1),rep(1,60))
table(y)
y
  0   1
190 291
```

Πολλαπλασιάζουμε τις δόσεις επί 100, για να είναι πιο ερμηνεύσιμοι οι συντελεστές:

```
dose1 <-dose * 100
table(dose1)
dose1
169.07 172.42 175.52 178.42 181.13 183.69  186.1 188.39
    59     60     62     56     63     59     62     60
```

Προσαρμόζουμε το μοντέλο M_2:

```
fit2.logit <- glm(y ~ dose1 + I(dose1^2),family = binomial(link = logit))
fit2.logit

Call: glm(formula = y ~ dose1 + I(dose1^2), family = binomial(link = logit))

Coefficients:
(Intercept)        dose1    I(dose1^2)
 431.10580    -5.20615       0.01564

Degrees of Freedom: 480 Total (i.e. Null);  478 Residual
Null Deviance:  645.4
Residual Deviance: 364.4  AIC: 370.4
```

Το οποίο έχει:

```
summary(fit2.logit)

Call:
glm(formula = y ~ dose1 + I(dose1^2), family = binomial(link = logit))

Deviance Residuals:
     Min       1Q   Median       3Q      Max
-2.81548  -0.62088   0.09326   0.38744   2.06300

Coefficients:
              Estimate  Std.Error  z value  Pr(>|z|)
(Intercept)431.105804 180.650653    2.386   0.01701 *
dose1        -5.206153   2.045191   -2.546   0.01091 *
I(dose1^2)    0.015641   0.005786    2.703   0.00687 **
---
Signif.codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 645.44  on 480  degrees of freedom
Residual deviance: 364.43  on 478  degrees of freedom
AIC: 370.43

Number of Fisher Scoring iterations: 6
```

Receiver Operating Characteristic (*ROC*) curve and Area Under Curve (*AUC*) score:

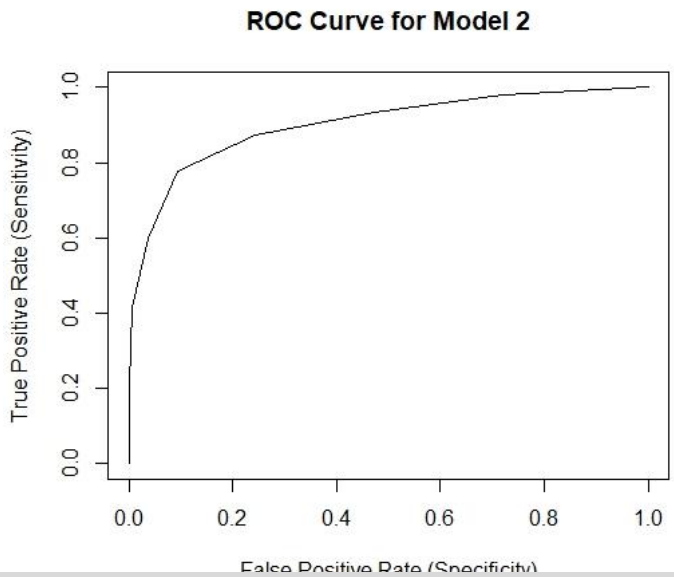
Τώρα αν θέλουμε να ελέγξουμε το πόσο καλά προβλέπουν τα δεδομένα, τα μοντέλα που προσαρμόσαμε, τότε θα χρειαστεί να εφαρμόσουμε τα διαγράμματα *ROC* και να δούμε τι *AUC* έχει το κάθε μοντέλο, για να κρίνουμε τις προβλεπτικές τους ικανότητες. Έτσι λοιπόν:

Εγκαθιστούμε και φορτώνουμε το πακέτο:

```
install.packages("ROCR")
library(ROCR)
```

Για το μοντέλο 2 (*ROC*):

```
pred2 <- -prediction(fitted(fit2.logit), fit2.logit$y)
perf2 <- -performance(pred2,"tpr","fpr")
plot(perf2,xlab = "False Positive Rate (Specificity)",
     ylab = "True Positive Rate (Sensitivity)",main = "ROC Curve for Model 2")
```



Και έτσι, το *AUC* είναι:

```
auc.m2 <- performance(pred2,"auc")
auc.m2@y.values
[[1]]
[1] 0.9010852
```

Οπότε ως προς την προβλεπτική ικανότητα, το μοντέλο 2, $AUC\ M2 = 0.9010852$ είναι αρκετά κοντά στο 1 που σημαίνει ότι έχει και πολύ καλή προβλεπτική ικανότητα.

Όρια ερμηνείας Area Under Curve (AUC):

- $\mathbf{80\% - 100\%}$ ⟶ **Εξαιρετική προβλεπτική ικανότητα.**
- $\mathbf{60\% - 80\%}$ ⟶ **Μέτρια προβλεπτική ικανότητα.**
- $\mathbf{< 60\%}$ ⟶ **Ότι χειρότερο μπορεί να συμβεί.**

ΠΑΡΑΔΕΙΓΜΑ 2 (BINARY DATA 2)

Τα παρακάτω δεδομένα αποτελούν δεδομένα αποτυχίας ενός εξαρτήματος (O-ring) στις εκτοξεύσεις των διαστημικών λεωφορείων πριν το Challenger. Το Challenger ήταν το διαστημικό λεωφορείο που εξερράγη στον αέρα κατά την εκτόξευσή του. Η θερμοκρασία είναι μια επεξηγηματική μεταβλητή. Το Challenger εξερράγη σε θερμοκρασία 31 βαθμών Fahrenheit. Κάθε πτήση θεωρείται σαν μια ανεξάρτητη δοκιμή. Το αποτέλεσμα του πειράματος είναι 1 αν ένα οποιοδήποτε εξάρτημα απέτυχε στην πτήση και 0 αν όλα τα εξαρτήματα δούλεψαν κανονικά.

Πτήση	Αποτέλεσμα	Θερμοκρασία
14	1	53
9	1	57
23	1	58
10	1	63
1	0	66
5	0	67
13	0	67
15	0	67
4	0	68
3	0	69
8	0	70
17	0	70
2	1	70
11	1	70
6	0	72
7	0	73
16	0	75
21	1	75
19	0	76
22	0	76
12	0	78
20	0	79
18	0	81

Ο σκοπός μας είναι με την χρήση του κατάλληλου GLM να συνδέσουμε την πιθανότητα αποτυχίας με τη θερμοκρασία. Εδώ η αναπαράσταση των δεδομένων θα γίνει ως 0 ή 1 (όχι ως μια συλλογή από διωνυμικές παρατηρήσεις αλλά ως συλλογή από δίτιμες Bernoulli).

Ερωτήσεις ερευνητών:

Προβλέψτε την πιθανότητα αποτυχίας κάποιου από τα εξαρτήματα στην θερμοκρασία που εκτοξεύτηκε το Challenger.

Ομαδοποιήστε τη θερμοκρασία σε μια νέα μεταβλητή με τρία επίπεδα που εκφράζουν τις τιμές θερμοκρασίας <=60, >60 και <=70, >70. Προσαρμόστε ένα μοντέλο για να ελέγξετε την επίδραση του παράγοντα ‘θερμοκρασία’.

Απάντηση

Πρώτα θα εισάγω τον πίνακα στην R:

```
x <- read.table(file = "C:\\Users\\30697\\Desktop\\challenger.txt", header = TRUE)
```

Για να δούμε αν όλα πήγαν καλά με την εισαγωγή:

```
head(x)
  res temp
1   1   53
2   1   57
3   1   58
4   1   63
5   0   66
6   0   67
```

Όλα καλά!

Μοντελοποίηση δεδομένων

Το μοντέλο θα είναι της μορφής:

$$logit(\pi_i) = log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 \cdot x_i\,, \qquad i = 1, 2, \dots, 23 \qquad \boxed{M_1}$$

Εδώ θα υποθέσω ότι η θερμοκρασία είναι μια επεξηγηματική, συνεχόμενη μεταβλητή. Επιπλέον θεωρώ ότι:

$$y_i \sim Binomial(1, \pi_i)$$

Δηλαδή ότι ακολουθούν μια Bernoulli.
Όπου θα θεωρήσω το να αποτύχει το εξάρτημα O-Ring ως «επιτυχία», ενώ το να δουλέψει κανονικά, ως «αποτυχία».

Προσοχή! Το μοντέλο αυτό δεν είναι ομαδοποιημένο (είναι δηλαδή ungrouped). Παρατηρήστε ότι ακόμα και αν γίνεται μιας μορφής ομαδοποίηση ως προς τις θερμοκρασίες (π.χ. στα 70 Fahrenheit), κάτι τέτοιο δεν μας συμφέρει, επειδή είναι ελάχιστες οι παρατηρήσεις που θα ομαδοποιηθούν και αυτές που θα γίνουν grouped, θα έχουν λίγες τιμές μέσα (π.χ. το covariate pattern στα 75 Fahrenheit με μία «επιτυχία» και μία «αποτυχία»). Απλά δεν μας συμφέρει να γίνει grouped το μοντέλο αυτό.

Ας το προσαρμόσουμε στην R:

```
mod.1 <- glm(res~temp,family = binomial(link = logit),data = x)
mod.1

Call:  glm(formula = res ~ temp, family = binomial(link = logit), data = x)

Coefficients:
(Intercept)        temp
    15.0429     -0.2322

Degrees of Freedom: 22 Total (i.e. Null);  21 Residual
Null Deviance:    28.27
Residual Deviance: 20.32    AIC: 24.32
```

Δίνω και την εντολή ***summary***:

```
summary(mod.1)

Call:
glm(formula = res ~ temp, family = binomial(link = logit), data = x)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.0611  -0.7613  -0.3783   0.4524   2.2175

Coefficients:
            Estimate  Std. Error  z value  Pr(>|z|)
(Intercept)  15.0429     7.3786    2.039   0.0415 *
temp         -0.2322     0.1082   -2.145   0.0320 *
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 28.267  on 22  degrees of freedom
Residual deviance: 20.315  on 21  degrees of freedom
AIC: 24.315

Number of Fisher Scoring iterations: 5
```

Πρώτα θα πρέπει να ελένξω την στατισστική σημαντικότητα των συντελεστών του μοντέλου μας

Wald Test

α'τρόπος ελέγχου, Wald Test:
Ας μην ξεχνάμε ότι το Wald Test ελέγχει την υπόθεση ότι:

$$H_0: \beta_i = 0$$
$$H_1: \beta_i \neq 0$$

Για κάθε μία παράμετρο.

Θα το ελέγξουμε από τον πίνακα Coefficients παραπάνω, όπου το $z - score$ είναι το

$$\sqrt{Wald - Statistic} \sim N(0,1)$$

Βλέπουμε από τα p-values στα δεξιά ότι:
α) Για την σταθερά (β_0). Σε επίπεδο στατιστικής σημαντικότητας $\alpha = 0.05$, απορρίπτουμε την μηδενική υπόθεση, διότι $(\boldsymbol{p - value < \alpha \Rightarrow 0.0415 < 0.05})$, επομένως θεωρούμε ότι ο όρος είναι στατιστικά σημαντικός στο μοντέλο αυτό.
β) Για τον όρο (β_1). Σε επίπεδο στατιστικής σημαντικότητας $\alpha = 0.05$, απορρίπτουμε την μηδενική υπόθεση, διότι $(\boldsymbol{p - value < \alpha \Rightarrow 0.0320 < 0.05})$, επομένως και αυτός ο όρος είναι στατιστικά σημαντικός στο μοντέλο μας.

Θα περάσουμε και στην ερμηνε΄λια των συντελεστών μας.

Οι συντελεστές μας στο μοντέλο M_1 είναι αυτοί:

```
coef(mod.1)
 (Intercept)         temp
15.0429016  -0.2321627
```

Πράγμα που σημαίνει ότι:
α) Για την σταθερά (y axis intercept). Όταν η θερμοκρασία είναι ίση με το μηδέν $(x_i = 0)$, τότε η πιθανότητα (odds) αποτυχίας του εξαρτήματος O-Ring, είναι:

$$O_{\pi_i} = exp\{\hat{\beta}_0\} = exp\{15.0429016\}$$

Εξαιρετικά ψηλά odds όταν η θερμοκρασία είναι στο μηδέν.

β) Για την παράμετρο β_1 $(temp)$. Κάθε φορά που η θερμοκρασία αυξάνεται κατά μία μονάδα Fahrenheit, τότε η αναλογία των πιθανοτήτων (δηλαδή Odds Ratio), μειώνεται κατά $exp\{-0.2321627\}$. Με άλλα λόγια η πιθανότητα (odds) αποτυχίας του εξαρτήματος *O-Ring* μειώνεται κατά $exp\{-0.2321627\}$, κάθε φορά που η θερμοκρασία αυξάνεται κατά μία μονάδα. Δηλαδή:

$$Odds\ Ratio\ (OR) = \frac{O_{\pi_{i+1}}}{O_{\pi_i}} = exp\{\hat{\beta}_1\} = exp\{-0.2321627\} \Rightarrow O_{\pi_{i+1}} = exp\{-0.2321627\} \cdot O_{\pi_i}$$

Πράγμα που μπορούμε επίσης να επιβεβαιώσουμε δίνοντας την εντολή fitted, μιας και τα δεδομένα μας είναι ταξινομημένα από την χαμηλότερη προς την ψηλότερη θερμοκρασία.

```
fitted(mod.1)
          1           2           3           4           5
0.93924781 0.85931657 0.82884484 0.60268105 0.43049313
          6           7           8           9          10
0.37472428 0.37472428 0.37472428 0.32209405 0.27362105
         11          12          13          14          15
0.22996826 0.22996826 0.22996826 0.22996826 0.15804910
         16          17          18          19          20
0.12954602 0.08554356 0.08554356 0.06904407 0.06904407
         21          22          23
0.04454055 0.03564141 0.02270329
```

Και παρατηρούμε ότι όσο αυξάνεται η θερμοκρασία, τόσο μειώνεται η πιθανότητα αποτυχίας του εξαρτήματος *O-Ring*.

Ξέρουμε ότι το Challenger εκτοξεύτηκε σε θερμοκρασία 31 Fahrenheit όταν είχε αποτύχει. Μπορούμε να υπολογίσουμε ποια ήταν η πιθανότητα αποτυχίας των εξαστημάτων του τους την θερμοκρασία αυτή

Εισάγουμε την θερμοκρασία αυτή σε ένα data frame για να δουλέψει η εντολή ***predict***:

```
newtemp <- c(31)
new <- data.frame(temp = newtemp)
new
  temp
1   31
```

Και κάνουμε την πρόβλεψη με βάση το μοντέλο που προσαρμόσαμε:

```
predict(mod.1,newdata = new,type = "response")
          1
0.9996088
```

Επομένως, παρατηρούμε ότι αν εκτοξευτεί το Challenger σε θερμοκρασία 31 Fahrenheit, τότε η πιθανότητα αποτυχίας του εξαρτήματος O-Ring, είναι: 99.96088%. Παρατηρείστε ότι τα δεδομένα μας ξεκινούν από την τιμή 53 Fahrenheit, οπότε είνα λογικό να καταλήγουμε σε τέτοιο συμπέρασμα, μιας και η θερμοκρασία 31 Fahrenheit είναι μια ακραία τιμή που είναι έξω από τα όρια των δεδομένων μου. Ίσως όμως το μοντέλο να χάνει και σε προβλεπτική ικανότητα. Αν και βγάζει νόημα, μιας και όπως αποδείξαμε, όσο χαμηλότερη είναι η θερμοκρασία τόσο υψηλότερη είναι η πιθανότητα αποτυχίας *O-Ring*.

Για ασφάλεια θα μπορούσαμε να ελένξουμε την προβλεπτική ικανότητα του μοντέλου.

Για να δούμε αν κάτι πάει στραβά με την προβλεπτική ικανότητα του μοντέλου. Τα δεδομένα μας είναι ήδη μη ομαδοποιημένα (ungrouped), οπότε δεν χρειάζεται κάποια μετρατροπή για αυτά που θα κάνω τώρα. Επομένως:

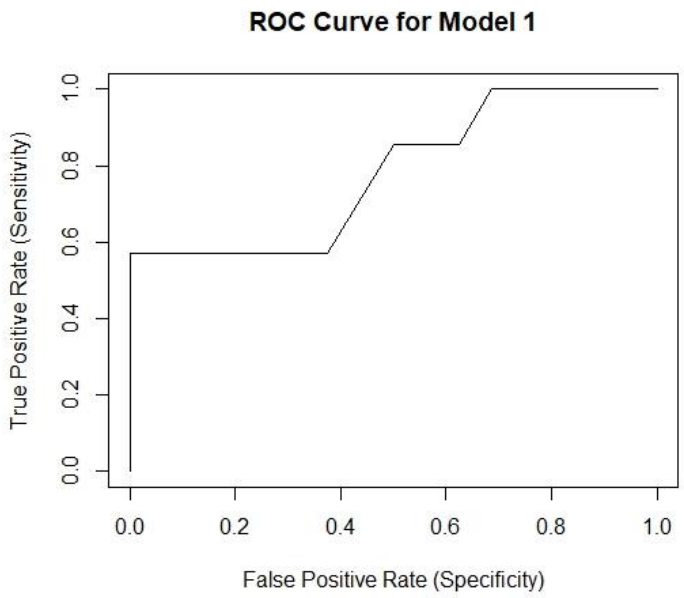
Φορτώνουμε το πακέτο:

```
library(ROCR)
```

Κάνω το διάγραμμα Receiver Operating Characteristic (*ROC*):

```
pred1 <- prediction(fitted(mod.1), mod.1$y)
perf1 <- performance(pred1,"tpr","fpr")
plot(perf1,xlab = "False Positive Rate (Specificity)",
     ylab = "True Positive Rate (Sensitivity)",main = "ROC Curve for Model 1")
```

Και έτσι, το διάγραμμα είναι:



Φαίνεται σαν να χάνει κάπου, αλλά ας δούμε και την τιμή του Area Under Curve (AUC):

```
auc.m1 <- performance(pred1,"auc")
auc.m1@y.values
[[1]]
[1] 0.78125
```

Επομένως συμπεραίνουμε ότι το μοντέλο μας έχει μέτρια προβλεπτική ικανότητα, μιας και η τιμή της $AUC = 78.125\%$, που είναι ανάμεσα στο 60% και το 80%.

Η ερευνητική ομάδα θα ήθελε να δει την πιθανότητα αποτυχίας σε ομαδοποιημένες κατηγορίες σε τρία επίπεδα θερμοκρασιών <=60, >60 και <=70, >70.

Έχουμε έναν παράγοντα, την θερμοκρασία, με τρία επίπεδα $i = 1, 2, 3$. Η link function δεν αλλάζει και παραμένει *logit*. Οπότε, σύμφωνα με το ερώτημα αυτό, το νέο μας μοντέλο θα είναι:

$$logit(\pi_i) = log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 \cdot z_{i1} + \beta_2 \cdot z_{i2}\,, \qquad i = 1, 2, 3 \qquad \boxed{M_2}$$

Όπου:

$$z_{i1} = \begin{cases} 0\ if\ i = 1 \\ 1\ if\ i = 2 \\ 0\ if\ i = 3 \end{cases}, \quad z_{i2} = \begin{cases} 0\ if\ i = 1 \\ 0\ if\ i = 2 \\ 1\ if\ i = 3 \end{cases}$$

Και ο νέος πίνακας των δεδομένων μας εκφρασμένων σε συχνότητες, θα είναι:

		Θερμοκρασία		
		≤ 60	**$(>60)\&(\leq 70)$**	**>70**
Απέτυχε το O-Ring?	**NAI**	**3**	**3**	**1**
	OXI	**0**	**8**	**8**

Πάμε λοιπόν να το προσαρμόσουμε στην R:

Πρώτα θα φτιάξω τον πίνακα αυτόν:

```
my.data <- matrix(c(3,0,3,8,1,8),ncol=3)
my.data
     [,1] [,2] [,3]
[1,]    3    3    1
[2,]    0    8    8
```

Και θα τον αναστρέψω με τις «επιτυχίες» να είναι στην αριστερή στήλη:

```
t <- t(my.data)
t
     [,1] [,2]
```

```
[1,]  3    0
[2,]  3    8
[3,]  1    8
```

Θα φτιάξω τον παράγοντα θερμοκρασία, που έχει τρία επίπεδα:

```
temp <- factor(c(1,2,3))
temp
[1] 1 2 3
Levels: 1 2 3
```

Και τώρα προσαρμόζω το μοντέλο, με reference group την θερμοκρασία μικρότερη των 60 Fahrenheit:

```
mod.2 <- glm(t ~ temp, binomial)
mod.2

Call:  glm(formula = t ~ temp, family = binomial)

Coefficients:
(Intercept)       temp2       temp3
     23.16       -24.14      -25.24

Degrees of Freedom: 2 Total (i.e. Null);  0 Residual
Null Deviance:    9.097
Residual Deviance: 5.249e-10  AIC: 10.56
```

Δίνω και την εντολή summary:

```
summary(mod.2)

Call:
glm(formula = t ~ temp, family = binomial)

Deviance Residuals:
[1]  0  0  0

Coefficients:
             Estimate  Std. Error  z value  Pr(>|z|)
(Intercept)     23.16    37437.85    0.001     1.000
temp2          -24.14    37437.85   -0.001     0.999
temp3          -25.24    37437.85   -0.001     0.999

(Dispersion parameter for binomial family taken to be 1)

      Null deviance: 9.0972e+00  on 2  degrees of freedom
Residual deviance: 5.2494e-10  on 0  degrees of freedom
AIC: 10.564

                    Number of Fisher Scoring iterations: 21
```

Παρ' όλο που το νέο μας μοντέλο είναι ομαδοποιημένο (grouped), όταν φτιάξαμε τις συχνότητες «επιτυχίας» και «αποτυχίας» για ομάδες θερμοκρασιών, αντί για κάθε θερμοκρασία ξεχωριστά, χάσαμε τις πληροφορίες που μας έδιναν τα ungrouped δεδομένα. Περιμένουμε τα συμπεράσματά μας να αλλάξουν σε σύγκριση με το πρώτο μοντέλο.

Σημαντικότητα των συντελεστών.
Για να το ελέγξω, θα κάνω Wald Test. Κοιτώντας στον πίνακα Coefficients, τα $p-values$, μας βγήκε ότι όλοι οι συντελεστές είναι μη στατιστικά σημαντικοί σε επίπεδο στατιστικής σημαντικότητας, έστω $\alpha = 0.05$.

Ερμηνεία των συντελεστών.

Για $i = 0$, έχουμε:

$$logit(\pi_0) = log\left(\frac{\pi_0}{1-\pi_0}\right) = \beta_0 + \beta_1 \cdot z_{01} + \beta_2 \cdot z_{02} \Longrightarrow$$
$$\Longrightarrow log(O_{\pi_0}) = \beta_0 \Longrightarrow O_{\pi_0} = \exp\{23.16\}$$

Επομένως, για το β_0 βλέπουμε ότι όταν η θερμοκρασία είναι μικρότερη ή ίση των 60 Fahrenheit, τότε η πιθανότητα (odds) του να αποτύχει το εξάρτημα O-ring, είναι ίση με $\exp\{23.16\}$.

Για $i = 1$, έχουμε:

$$logit(\pi_1) = log\left(\frac{\pi_1}{1-\pi_1}\right) = \beta_0 + \beta_1 \cdot z_{11} + \beta_2 \cdot z_{12} \Longrightarrow$$
$$\Longrightarrow log(O_{\pi_1}) = \beta_0 + \beta_1 \Longrightarrow O_{\pi_1} = \exp\{23.16\}$$

Άρα:

$$\beta_1 = log(O_{\pi_1}) - log(O_{\pi_0}) \Longrightarrow \beta_1 = log\left(\frac{O_{\pi_1}}{O_{\pi_0}}\right) \Longrightarrow \frac{O_{\pi_1}}{O_{\pi_0}} = \exp\{-24.14\}$$

Επομένως, για το β_1 βλέπουμε ότι όταν η θερμοκρασία αλλάξει το επίπεδο στο οποίο βρίσκεται και πάει από εκεί που ήταν μικρότερη ή ίση των 60 Fahrenheit σε μία θερμοκρασία μεταξύ των 60 και 70 Fahrenheit, τότε ο λόγος πιθανοτήτων (odds ratio) του να αποτύχει το εξάρτημα *O-ring*, θα μειωθεί κατά $\exp\{24.14\}$.

Για $i = 2$, έχουμε:

$$logit(\pi_2) = log\left(\frac{\pi_2}{1-\pi_2}\right) = \beta_0 + \beta_1 \cdot z_{21} + \beta_2 \cdot z_{22} \Longrightarrow$$
$$\Longrightarrow log(O_{\pi_2}) = \beta_0 + \beta_2 \Longrightarrow O_{\pi_2} = \exp\{-25.24\}$$

Άρα:

$$\beta_2 = log(O_{\pi_1}) - log(O_{\pi_0}) \Longrightarrow \beta_2 = log\left(\frac{O_{\pi_2}}{O_{\pi_0}}\right) \Longrightarrow \frac{O_{\pi_1}}{O_{\pi_0}} = \exp\{-25.24\}$$

Επομένως, για το β_2 βλέπουμε ότι όταν η θερμοκρασία αλλάξει το επίπεδο στο οποίο βρίσκεται και πάει από εκεί που ήταν μικρότερη ή ίση των 60 Fahrenheit σε μία θερμοκρασία μεγλύτερη των 70 Fahrenheit, τότε ο λόγος πιθανοτήτων (odds ratio) του να αποτύχει το εξάρτημα *O-ring*, θα μειωθεί κατά $\exp\{25.24\}$

ΠΑΡΑΔΕΙΓΜΑ 3 (BINARY)

Σε ένα πείραμα κάποιοι ιδιαίτερα εκκεντρικοί ερευνητές θέλουν να συσχετίσουν την πιθανότητα καρδιακής ανεπάρκειας με τη συχνότητα ροχαλητού κατά τη διάρκεια του βραδυνού ύπνου. Πήραν υποκείμενα με διαφορετική συχνότητα βραδυνού ροχαλητού και κατέγραψαν μετά από αρκετά χρόνια την εμφάνιση καρδιακής ανεπάρκειας. Τα αποτελέσματα φαίνονται στον παρακάτω πίνακα

	Συχνότητα ροχαλητού			
	Ποτέ	2 ώρες	4 ώρες	5 ώρες
Καρδιακή ανεπάρκεια	24	35	21	30
Οχι καρδιακή ανεπάρκεια	1355	603	192	224

Ο στόχος είναι να συσχετιστεί η πιθανότητα καρδιακής ανεπάρκειας με τη συχνότητα του ροχαλητού. Και στγκεκριμένα να προβλέωουμε την πιθανότητα καρδιακής ανεπάρκειας αν κάποιος ροχαλίζει 1, 2, 3 ώρες το βράδυ

Πρώτα από όλα θα φτιάξουμε τον πίνακα αυτόν στην R:

```
snoring <- matrix(c(24,1355,35,603,21,192,30,224),ncol=2,byrow=TRUE)
snoring
     [,1]  [,2]
[1,]   24  1355
[2,]   35   603
[3,]   21   192
[4,]   30   224
```

Φτιάχνω την στήλη scores:

```
scores <- c(0,2,4,5)    #scores correspond to never,occasionally,never,every night
scores
[1] 0 2 4 5
```

Τα ενώνω σε μια μήτρα:

```
m1 <- cbind(scores,snoring)
colnames(m1) <- c("Scores","Heart Disease","No Heart Disease" )
m1
     Scores  Heart Disease  No Heart Disease
[1,]      0             24              1355
[2,]      2             35               603
[3,]      4             21               192
[4,]      5             30               224
```

ΜΟΝΤΕΛΟΠΟΙΗΣΗ ΔΕΔΟΜΕΝΩΝ

Το μοντέλο θα είναι της μορφής:

$$logit(\pi_i) = log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 \cdot x_i\,, \qquad i = 1,2$$

Γιατί υποθέτω ότι η συχνότητα βραδυνού ροχαλητού είναι μια συνεχόμενη μεταβλητή. Επιπλέον θεωρώ ότι:

$$n_i \cdot y_i \sim Binomial(n_i, \pi_i)$$

Όπου το να έχει ένα άτομο καρδιακή ανεπάρκεια θεωρείται «επιτυχία», ενώ το να μην έχει, «αποτυχία».

Προσαρμόζω λοιπόν το μοντέλο στην R:

```
snoring.fit1 <- glm(snoring ~ scores, family = binomial(link = logit))
snoring.fit1

Call: glm(formula = snoring ~ scores, family = binomial(link = logit))

Coefficients:
(Intercept)     scores
    -3.8662     0.3973

Degrees of Freedom: 3 Total (i.e. Null);  2 Residual
Null Deviance:    65.9
Residual Deviance: 2.809   AIC: 27.06
```

ΕΡΜΗΝΕΙΑ ΣΥΝΤΕΛΕΣΤΩΝ

```
summary(snoring.fit1)

Call:
glm(formula = snoring ~ scores, family = binomial(link = logit))

Deviance Residuals:
      1       2       3       4
-0.8346  1.2521  0.2758  -0.6845

Coefficients:
             Estimate  Std. Error   z value    Pr(>|z|)
(Intercept)  -3.86625     0.16621  -23.261   < 2e-16 ***
scores        0.39734     0.05001    7.945   1.94e-15 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 65.9045 on 3 degrees of freedom
Residual deviance: 2.8089 on 2 degrees of freedom
AIC: 27.061

Number of Fisher Scoring iterations: 4

Πράγμα που σημαίνει ότι:
α) Για την σταθερά (intercept). Αν η συχνότητα βραδυνού ροχαλητού είναι ίση με το μηδέν ($x_i = 0$), τότε η πιθανότητα (odds) του να έχει κάποιος καρδιακή ανεπάρκεια, είναι:

$$O_{\pi_i} = exp\{\hat{\beta}_0\} = exp\{-3.86625\}$$

β) Για την παράμετρο β_1 $(Scores)$. Κάθε φορά που η συχνότητα βραδινού ροχαλητού αυξάνεται κατά μία μονάδα, τότε η αναλογία των πιθανοτήτων (δηλαδή Odds Ratio), αυξάνεται κατά $exp\{0.39734\}$. Δηλαδή:

$$\frac{O_{\pi_{i+1}}}{O_{\pi_i}} = exp\{\hat{\beta}_1\} = exp\{0.39734\} \Longrightarrow O_{\pi_{i+1}} = exp\{0.39734\} \cdot O_{\pi_i}$$

Πιθανότητα καρδιακής ανεπάρκειας αν κάποιος ροχαλίζει 1, 2, 3 ώρες το βράδυ.

Πρώτα φτιάχνουμε ένα data frame με ολες τις τιμές που θέλουμε να προβλέψουμε:

```
newscore <- c(1,2,3)
new <- data.frame(scores = newscore)
new
  scores
1      1
2      2
3      3
```

Και ύστερα κάνουμε πρόβλεψη της πιθανότητας:

```
predict(snoring.fit1, newdata = new, type = "response")
        1         2         3
0.03020986 0.04429511 0.06451072
```

Άρα λοιπόν συμπεραίνουμε ότι:

- Υπάρχει πιθανότητα 3.020986% του να αποκτήσει κάποιος καρδιακή ανεπάρκεια, αν ροχαλίζει μία ώρα το βράδυ.
- Υπάρχει πιθανότητα 4.429511% του να αποκτήσει κάποιος καρδιακή ανεπάρκεια, αν ροχαλίζει δύο ώρες το βράδυ.
- Υπάρχει πιθανότητα 6.451072% του να αποκτήσει κάποιος καρδιακή ανεπάρκεια, αν ροχαλίζει τρεις ώρες το βράδυ.

ΠΑΡΑΔΕΙΓΜΑ 4 (GAUSSIAN)

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.

Αυτά είναι τα ερωτήματα των ερευνητών που ζητήσανε

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.

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

Προσθέστε την αλληλεπίδραση εβδομάδων αποθήκευσης και θερμοκρασίας στο μοντέλο ελέγξτε τη στατιστική του σημαντικότητα

ΛΥΣΗ:

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

```
□□□□ <− □(□□, □□, □□, □□, □□, □□, □□, □□, □□, □□, □□, □□)
```

```
□□□□ <− □□□(□(□, □□, □□), □□□□□ = □)

□□□□ <− □□□□□□□(□□□□)

□□□□□ <− □□□(□(□, □, □, □), □□□□ = □)
```

ΜΟΝΤΕΛΟΠΟΙΗΣΗ

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

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

Όπου:

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

Οπότε πάμε να το προσαρμόσουμε στην *R*:

```
□□□□ <− □□□(□□□□ ~ □□□□ + □□□□□, □□□□□□ =
□□□□□□□□)



□□□□:

□□□(□□□□□□□ = □□□□ ~ □□□□ + □□□□□, □□□□□□ =


□□□□□□□□
    □□        □□              □□
−□. □□□□  − □. □□□□  − □. □□□□  □. □□□□


□□□□□□□□□□□□:
     □□□□□□□□ □□□. □□□□□ □
```

$(Intercept)$ 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

ΕΡΜΗΝΕΙΑ ΠΑΡΑΜΕΤΡΩΝ

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.

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:

```
□□□ <- □□□□□.□□□□□□(□□□□□ = □□□□□□□(□□), □□□□□□ = □)

□□□□□ <- □□□□□□□□(□□□□□, □□□□□□□□ = □□□, □□□□□ = '□□□□□□□□□□', □□.
□□□ = □□□□□)

□□□□□

$□□□

       □

     □□.
  □□□□□□



$□□.□□□

[□] □.□□□□□□□



$□□□□□□□□□.
```

Ενώ το διάστημα εμπιστοσύνης 95% θα βρεθεί από:

```
□□□□□□□□(□□□□□)

□□□□□□□□ □□□ □□□□□□□□□□ □□
□□ □□□□□...

        □.□ %   □□.□ %


□□□□□□     −□.□□□□□□□  □.

□□□□□□□  −□□.□□□□□□□ −□□.
```

□□□□□ − □. □□□□□□ − □. □□□□□□□

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

$$[-\mathbf{26.230906}, -\mathbf{16.2690943}]$$

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

$$[\mathbf{26.85209}, \quad \mathbf{36.81391}]$$

Προσθέστε την αλληλεπίδραση εβδομάδων αποθήκευσης και θερμοκρασίας στο μοντέλο ελέγξτε τη στατιστική του σημαντικότητα, ποιο μοντέλο προτείνεται να επιλέξουμε;

Το νέο μας μοντέλο θα είναι της μορφής:

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

Όπου:

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

Πάμε να το προσαρμόσουμε στην *R*:

```
□□□□ <- □□□(□□□□ ~ □□□□ * □□□□□, □□□□□□ =
□□□□□□□□□)



□□□□:

□□□(□□□□□□□□ = □□□□ ~ □□□□ * □□□□□, □□□□□□ =



□□□□□□□□□

  □□      □□          □□

-□. □□ - □.      □. □□ □. □□  □.



□□□□□□□□□

            □□□□□□□□□      □□□. □□□□□ □
```

```
(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
```

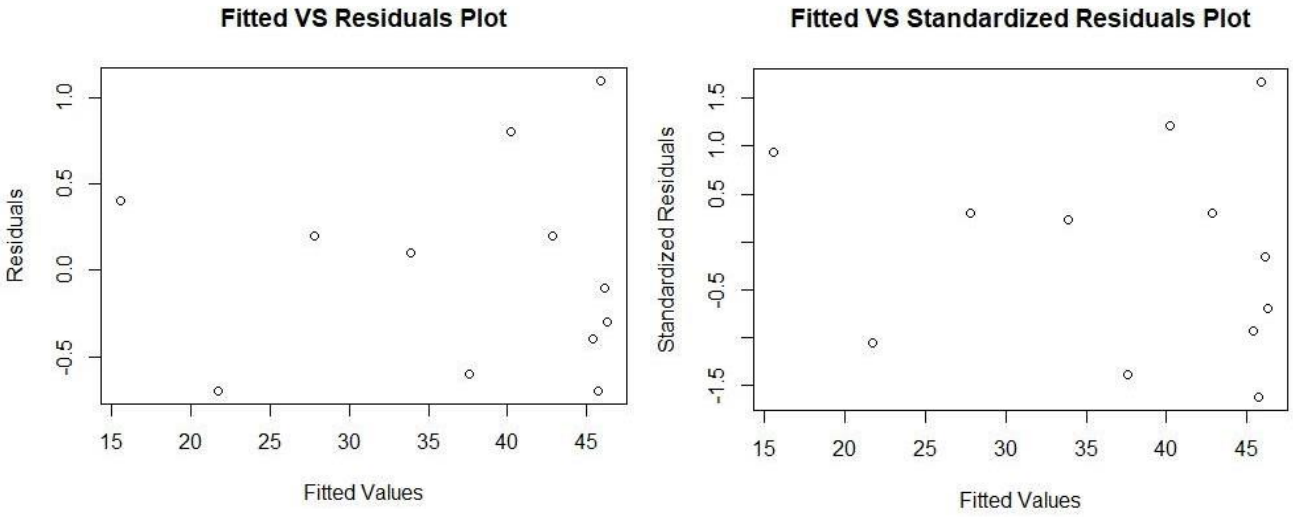
Ας δούμε ποιό είναι το καλύτερο από τα δύο μοντέλα:

```
□□□(□□□□)
[□] □□.
□□□□□
□□□(□□□□)
```

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

```
□□□□(□□□□□□□(□□□□), □□□□□(□□□□, □□□□ = "□□□□□□□□"), □□□□ =
"□□□□□□□ □□□□□□□", □□□□
          = "□□□□□□□□□□□", □□□□ = "□□□□□□□ □□ □□□□□□□□□□ □□□□□")
```

```
□□□□(□□□□□□□(□□□□□), □□□□□□□□□□(□□□□□, □□□□□ = "□□□□□□□□"), □□□□□
        = "□□□□□□□ □□□□□□□", □□□□□ = "□□□□□□□□□□□□□□□
        □□□□□□□□□□", □□□□□
```



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