Άσκηση 1

Δημιουργώ ένα αρχείο txt με όνομα “seira3_ask1” το οποίο περιέχει τα εξής:

season	lanes	accidents
1	2	35
1	4	199
2	2	56
2	4	184
3	2	53
3	4	227
4	2	50
4	4	270

Για την εισαγωγή των δεδομένων στην R δίνω τις εντολές:

> data<-read.csv(file.choose(),header=TRUE,sep='')

> attach(data)

Στη συνέχεια κάνω κωδικοποίηση για τις λωρίδες (αν είναι 2 τότε l=1 και αν είναι 4 τότε l=0) με την εξής εντολή:

> l <- c(1, 0, 1, 0, 1, 0, 1, 0)

Για την προσαρμογή του μοντέλου Poisson δίνουμε την εντολή:

> poisson_model <- glm(accidents ~ season + l, family = poisson)

Και με την εντολή

> summary(poisson_model)

παίρνουμε συνοπτικά τα στοιχεία του μοντέλου

```
Call:
glm(formula = accidents ~ season + l, family = poisson)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  5.10745    0.07986  63.954  < 2e-16 ***
season       0.11138    0.02744   4.059 4.92e-05 ***
l           -1.51206    0.07932 -19.064  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 499.1102  on 7  degrees of freedom
Residual deviance:   8.2766  on 5  degrees of freedom
AIC: 65.992

Number of Fisher Scoring iterations: 4
```

Από τις πιο πάνω πληροφορίες μπορούμε να εφαρμόσουμε τον έλεγχο Wald στο οποίο παρατηρούμε ότι οι μεταβλητές season και l είναι σημαντικές αφού p-value<0.05.

Για την εξέταση της καταλληλότητας του μοντέλου με την ελεγχοσυνάρτηση Deviance μπορούμε από τις πιο πάνω πληροφορίες να παρατηρήσουμε ότι:

$D(\beta.null)-D(\beta)=499.1102-8.2766=490.8336\sim X^2_2$

Επίσης από τις πιο κάτω εντολές:

```
> fit_null<-glm(accidents~1, family=poisson)
```

```
> anova(fit_null, poisson_model, test="Chisq")
Analysis of Deviance Table

Model 1: accidents ~ 1
Model 2: accidents ~ season + l
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1         7     499.11
2         5       8.28  2   490.83 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Παρατηρούμε ότι οι πιο πάνω υπολογισμοί που κάναμε επαληθεύονται και πως p-value<0.05 επομένως απορρίπτουμε την $H_0:\beta_j=0$,null έναντι της $H_1:\beta_j\neq0$

Επιπρόσθετα με την πιο κάτω εντολή υπολογίζεται ο συντελεστής προσδιορισμού Deviance

```
> r.squared<-(fit_null$deviance-poisson_model$deviance)/fit_null$deviance
```

```
> r.squared
```

[1] 0.9834173

Παρατηρούμε ότι είναι αρκετά κοντά στη μονάδα.

Επομένως, με τη χρήση της ελεγχοσυνάρτηση Deviance παρατηρούμε ότι το μοντέλο που προσαρμόσαμε είναι κατάλληλο.

Για τον έλεγχο της καταλληλότητας του μοντέλου μέσω του ελέγχου των υπολοίπων δίνουμε τις εντολές:

```
> r.pears<-residuals(poisson_model,type='pearson')
```

```
> sum(r.pears^2)
```

[1] 8.309523

```
> r.pears.null<-residuals(fit_null,type='pearson')
```

```
> sum(r.pears.null^2)
```

[1] 472.041

Από τα πιο πάνω αποτελέσματα παρατηρούμε ότι sum(r.pears^2) είναι πολύ μικρότερο από το sum(r.pears.null^2) επομένως το μοντέλο Poisson έχει καλύτερη προσαρμογή από το null model, εξηγεί δηλαδή καλύτερα τη διακύμανση των δεδομένων. Επομένως, το μοντέλο που προσαρμόσαμε είναι κατάλληλο.

Άσκηση 2

Δημιουργώ ένα αρχείο txt με όνομα “seira3_ask2” το οποίο περιέχει τα εξής:

y	n	E	x
3	50	1	2.00
5	49	1	2.64
19	47	1	3.48
19	38	1	4.59
24	29	1	6.06
35	50	1	8.00
2	50	2	2.00
14	49	2	2.64
20	50	2	3.48
27	50	2	4.59
41	50	2	6.06
40	50	2	8.00
28	50	3	2.00
37	50	3	2.64
46	50	3	3.48
48	50	3	4.59
48	50	3	6.06
50	50	3	8.00

Για την εισαγωγή των δεδομένων στην R δίνω τις εντολές:

> data<-read.csv(file.choose(),header=TRUE,sep='')

> attach(data)

Για την προσαρμογή του μοντέλου λογιστικής παλινδρόμησης δίνουμε την εντολή

> model <- glm(cbind(y, n - y) ~ E + x, family = binomial)

και πιο κάτω βλέπουμε τα χαρακτηριστικά του μοντέλου

> summary(model)

```
Call:
glm(formula = cbind(y, n - y) ~ E + x, family = binomial)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.83600    0.36732  -13.17   <2e-16 ***
E            1.30789    0.11744   11.14   <2e-16 ***
x            0.61710    0.05109   12.08   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 413.644  on 17  degrees of freedom
Residual deviance:  78.359  on 15  degrees of freedom
AIC: 146.56

Number of Fisher Scoring iterations: 5
```

Για τον έλεγχο της καταλληλότητας μπορούμε να συγκρίνουμε το μοντέλο μέσω Deviance με το μοντέλο που περιέχει μόνο τη β_0:

> fit_null<-glm(cbind(y, n - y) ~1, family=binomial)

> anova(fit_null, model, test="Chisq")

```
Analysis of Deviance Table

Model 1: cbind(y, n - y) ~ 1
Model 2: cbind(y, n - y) ~ E + x
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1        17     413.64
2        15      78.36  2   335.29 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Παρατηρούμε ότι p-value<0.05 επομένως απορρίπτουμε την H_0:μοντέλο με τη β_0 έναντι της H_1: μοντέλο με τύπο και ποσότητες εντομοκτόνων. Επομένως, το μοντέλο που προσαρμόσαμε είναι καλύτερο.

Για να εξετάσουμε αν υπάρχουν σημεία επιρροής στο μοντέλο μας δίνουμε τις εντολές:

> cooks<-cooks.distance(model)

> cooks

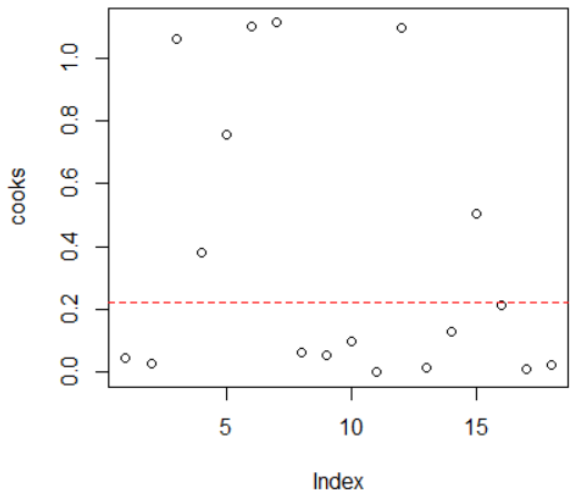
```
           1            2            3            4            5            6
4.601498e-02 2.849906e-02 1.060406e+00 3.790004e-01 7.547549e-01 1.098881e+00
           7            8            9           10           11           12
1.111419e+00 6.212636e-02 5.385975e-02 1.006340e-01 4.397794e-06 1.095059e+00
          13           14           15           16           17           18
1.452459e-02 1.292736e-01 5.065981e-01 2.143010e-01 1.091585e-02 2.385933e-02
```

Από τα πιο πάνω αποτελέσματα παρατηρούμε ότι υπάρχουν αρκετά σημεία που εμφανίζουν υψηλές τιμές (3, 4, 5, 6, 7, 12, 15), γεγονός που δείχνει ότι ενδέχεται να έχουν επιρροή στο μοντέλο.

Για περαιτέρω εξέταση δίνουμε τις εντολές:

> plot(cooks)

> abline(h=4/length(cooks),col="red",lty=2)



Παρατηρούμε ότι οι παρατηρήσεις 3, 4, 5, 6, 7, 12, 15 είναι πάνω από το όριο επομένως θεωρούνται σημεία επιρροής

Επίσης, με την εντολή:

> 1-pchisq(model$deviance, model$df.residual)

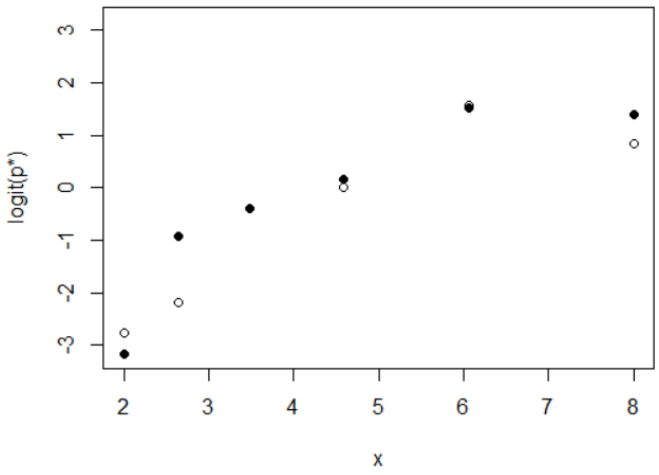
[1] 1.39226e-10

Ελέγχω την σημαντικότητα του μοντέλου. Ουσιαστικά παίρνω το p-value στον έλεγχο H_0: μοντέλο με τύπο και ποσότητες εντομοκτόνων έναντι της H_1:saturated model. Επειδή το p-value<0.05 απορρίπτω την H_0 και άρα δέχομαι το saturated model ως καλύτερο.

Βασικό λόγος αδυναμίας του μοντέλου μας μπορεί να είναι το γεγονός ότι μια μεταβλητή χρειάζεται μετασχηματισμό. Για να το βρω αυτό εργάζομαι γραφικά δίνοντας τις εντολές:

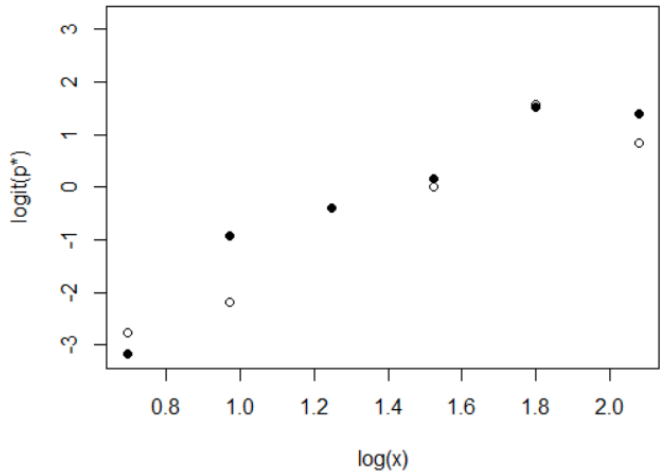
> pstar<-y/n

> plot(log(pstar/(1-pstar))~x, ylab="logit(p*)",pch=c(1,16)[E])



Προτείνουμε το μετασχηματισμό ln(x). Θέλουμε ουσιαστικά το logit να σχετίζεται γραμμικά με το x.

> plot(log(pstar/(1-pstar))~log(x), ylab="logit(p*)",pch=c(1,16)[E])



Για την προσαρμογή του νέου μοντέλου δίνω τις εντολές:

> model2 <- glm(cbind(y, n - y) ~ E + log(x), family = binomial)

> summary(model2)

```
Call:
glm(formula = cbind(y, n - y) ~ E + log(x), family = binomial)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -5.9975     0.4412  -13.59   <2e-16 ***
E             1.3480     0.1210   11.14   <2e-16 ***
log(x)        2.7527     0.2149   12.81   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 413.644  on 17  degrees of freedom
Residual deviance:  58.206  on 15  degrees of freedom
AIC: 126.4

Number of Fisher Scoring iterations: 4
```

Από το παραπάνω παρατηρούμε πως ότι το Residual deviance αυτού του μοντέλου με 15 βαθμούς ελευθερίας είναι καλύτερο από πριν αφού έχει χαμηλότερη τιμή (πριν ήταν 78.359). Επίσης, η σημαντικότητα των μεταβλητών δεν επηρεάστηκε από το μετασχηματισμό.

Επιπλέον, παρατηρώντας την τιμή του κριτηρίου AIC=126.4 βλέπουμε ότι είναι μικρότερη από του προηγούμενου μοντέλου που προσαρμόσαμε που ήταν 146.56, επομένως αυτό το μοντέλο είναι καλύτερο.