

מודלים סטטיסטיים ויישומיים 52518 5 תשע"ח – תרגיל 10

להגשה עד 8.1.18 בשעה 23:55

1. מצורף קובץ בשם credit.csv עם נתונים בנוגע ל-1000 הלוואות של לקוחות בנק גרמני. המשתנים המסבירים בקובץ (רשימה חלקית מתוך הנתונים המקוריים) הם:
- good_credit: משתנה בינארי המציין האם רמת הסיכון של הלוקה טובה (1) או לא (0).
 - term:משך ההלוואה בחודשים.
 - amount: סכום ההלוואה המבוקש (בمارك גרמני).
 - age: גיל הלוקה.
 - land: משתנה בינארי המציין האם הלוקה בעל אדמה (1) או לא (0).
- המשתנה המסביר יהיה good_credit. בצעו ראשית ניתוח תיאורי של הנתונים והציבו על מאפייני הנתונים השונים (התפלגות, קשרים אפשריים בין משתנים, צפיפות חריגות ובן הלאה). הריצו מודל רגסיה לוגיסטיבית לבדיקת הקשר בין המשתנים המסבירים למשתנה המסביר. יישמו את הכלים בקובץ logcheck1.pdf לבדיקה טיב התאמת המודל ובמידת הצורך הריצו מודל חלופי. דנו בתוצאות.
- הערה: אין צורך למצוא מודל אופטימלי, מעבר למודל הבסיסי התאימו רק מודל חלופי אחד.

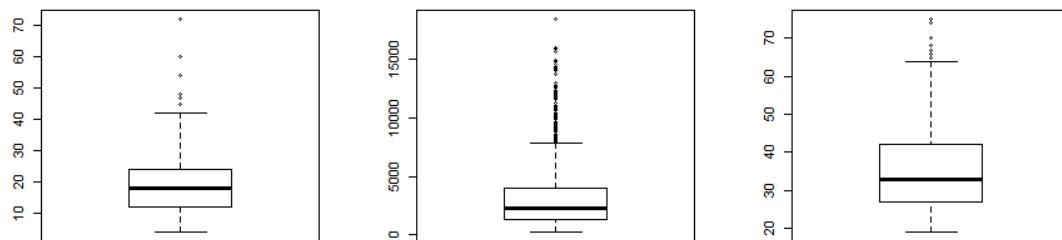
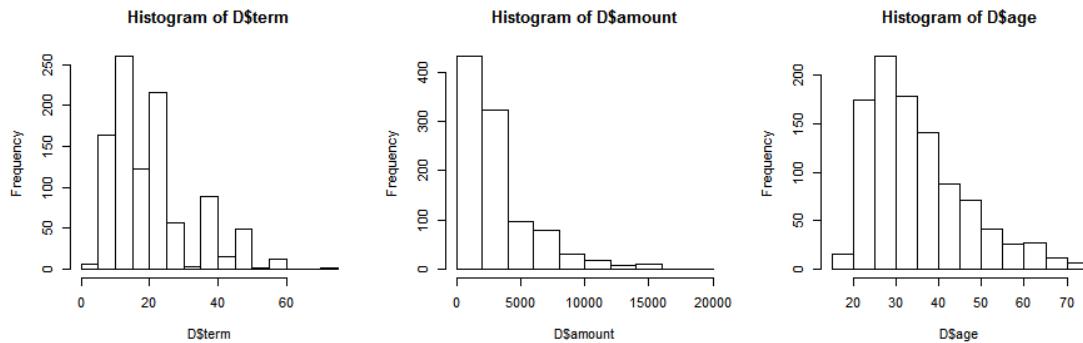
ניתוח כללי של הנתונים:

good_credit	term	amount	age	land
Min. :0.0	Min. : 4.0	Min. : 250	Min. :19.00	Min. :0.000
1st Qu.:0.0	1st Qu.:12.0	1st Qu.: 1366	1st Qu.:27.00	1st Qu.:0.000
Median :1.0	Median :18.0	Median : 2320	Median :33.00	Median :0.000
Mean : 0.7	Mean :20.9	Mean : 3271	Mean :35.54	Mean :0.154
3rd Qu.:1.0	3rd Qu.:24.0	3rd Qu.: 3972	3rd Qu.:42.00	3rd Qu.:0.000
Max. :1.0	Max. :72.0	Max. :18424	Max. :75.00	Max. :1.000

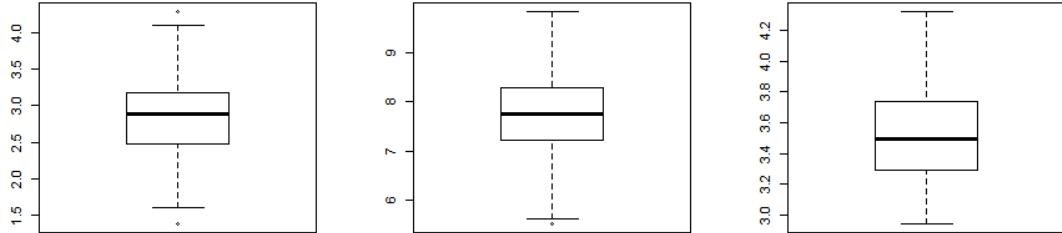
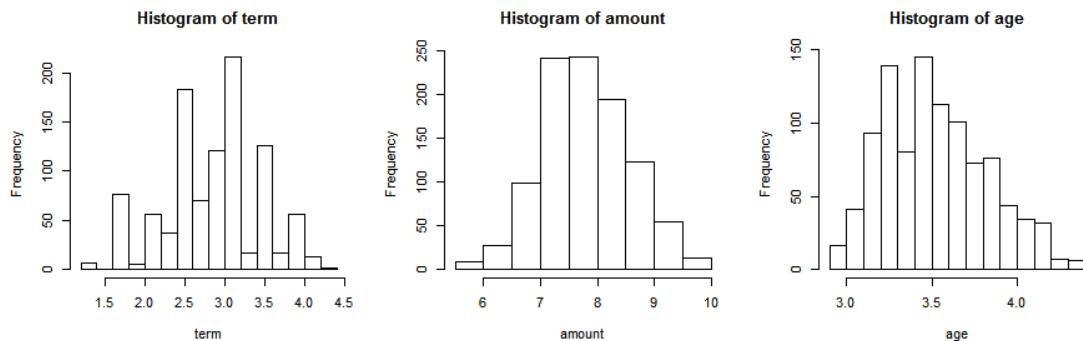
המשתנה good_credit מתפלג ברנולי, על פניו המדגם בולו נאמד $\hat{\theta} = 0.7$.

המשתנה land גם הוא ברנולי, על פניו המדגם בולו נאמד $\hat{\theta} = 0.154$.

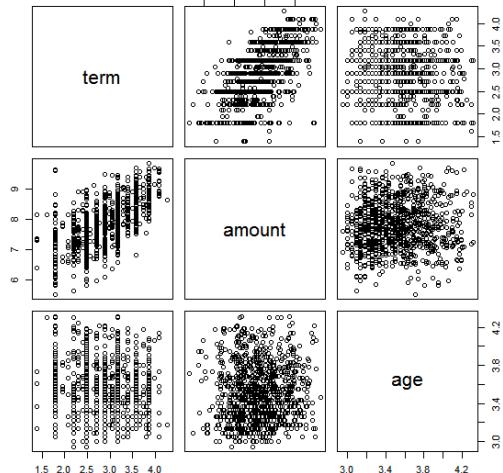
המשתנים age, term, amount, מתחולגים עם זנב ימני ארוך:



לבן נרצה לחת \log שלהם:



לא נראה שיש קשרים בין המשתנים, למעט אולי בין `term`-`amount`.



הרגסיה הלוגיסטיבית עליהם מניבה את הנתונים הבאים:

```
Call:
glm(formula = good_credit ~ ., family = binomial(link = "logit"),
     data = Dlog)
```

```
Deviance Residuals:
    Min      1Q  Median      3Q      Max 
-2.2110 -1.2288  0.6796  0.8698  1.5314
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-1.1179	1.1539	-0.969	0.332664
term	-0.9297	0.1714	-5.425	5.78e-08 ***
amount	0.1870	0.1230	1.521	0.128353
age	0.9560	0.2565	3.726	0.000194 ***
land	-0.7341	0.1980	-3.708	0.000209 ***

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

(Dispersion parameter for binomial family taken to be 1)

```
Null deviance: 1221.7 on 999 degrees of freedom
Residual deviance: 1149.9 on 995 degrees of freedom
AIC: 1159.9
```

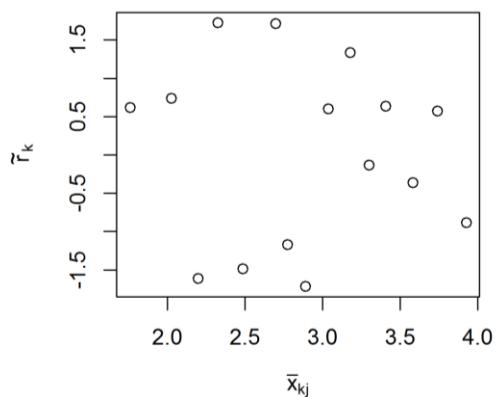
מספר הפלוקצייה Fisher Scoring iterations: 4

בעזרת הפונקציה `resid_anal` (שינוי כל מהקוד בקובץ `logistic.R` שבתיקו Logistic Regression Materials) ניתן ליצור קבצי `pdf` ובן ניהול ב-`Rich Text` ננתן שאריות עבור המשתנים החזויים:

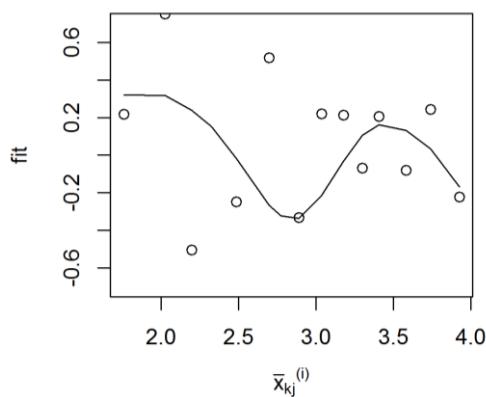
מנת ליצור קבצי `pdf` במקום `pdf` וכן טיפול ב-`Rich Text` ננתן שאריות עבור המשתנים החזויים:

Term

Norm. Res. for log(term)

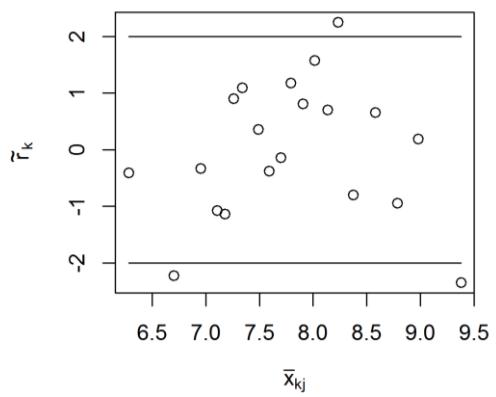


Fit vs. Ordered Res. for log(term)

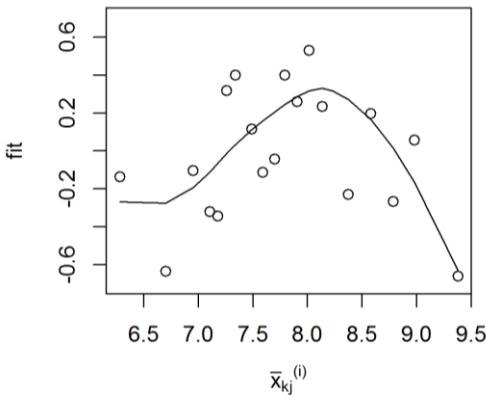


Amount

Norm. Res. for log(amount)

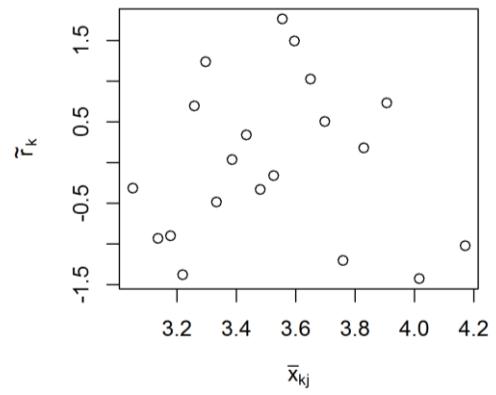


Fit vs. Ordered Res. for log(amount)

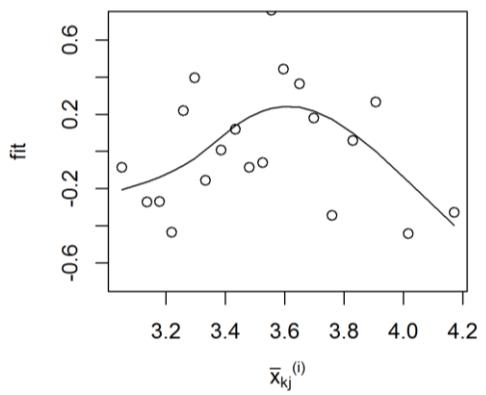


Age

Norm. Res. for log(age)



Fit vs. Ordered Res. for log(age)



ניתן לראות כי יש חריגות ב-amount ולבן נרצה להוסף למודל טרנספורמציות שלו – נתחיל עם חזקה שנייה
ושלישית:

```
Call:
glm(formula = good_credit ~ ., family = binomial(link = "logit"),
      data = D2)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.3628	-1.1979	0.6667	0.8472	2.0297

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	103.43497	43.04129	2.403	0.016254 *
term	-0.98021	0.17315	-5.661	1.5e-08 ***
amount	-43.21312	16.73423	-2.582	0.009814 **
age	0.99314	0.25875	3.838	0.000124 ***
land	-0.66944	0.20150	-3.322	0.000893 ***
amount.2	5.92123	2.15633	2.746	0.006033 **
amount.3	-0.26567	0.09201	-2.887	0.003885 **

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

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 1221.7 on 999 degrees of freedom
Residual deviance: 1130.8 on 993 degrees of freedom
AIC: 1144.8

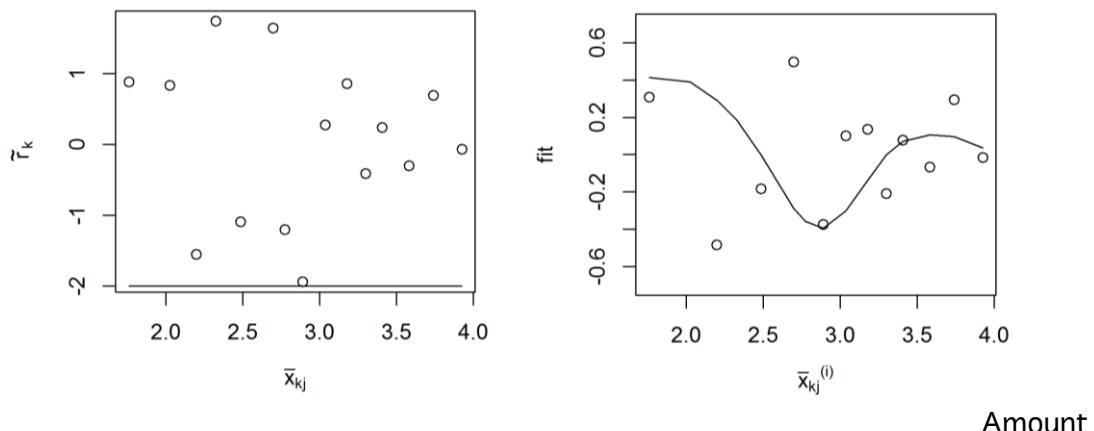
Number of Fisher Scoring iterations: 4

ניתן לראות כי שני המשתנים שהוספנו מובחקים.

ניתוח השאריות המעודכן:

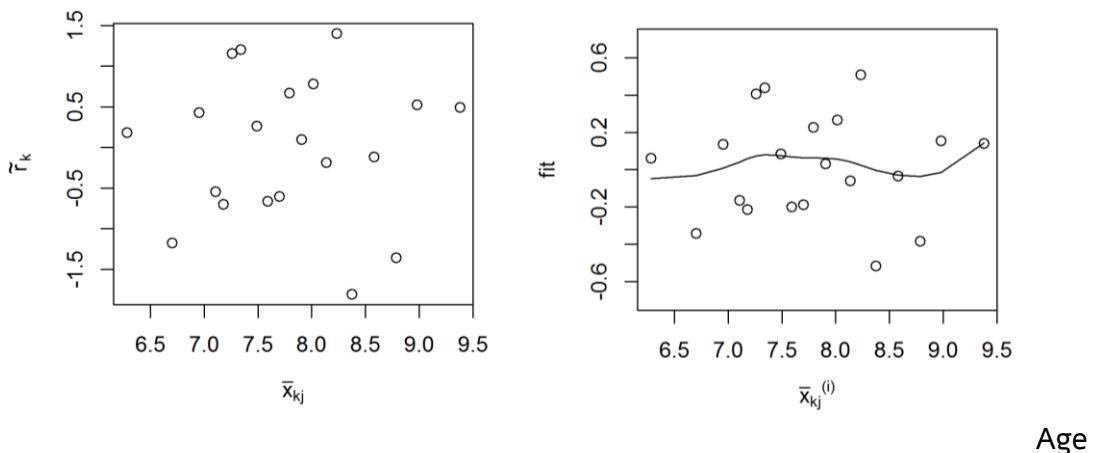
Term

Norm. Res. for log(term)_new_model Fit vs. Ordered Res. for log(term)_new_mc



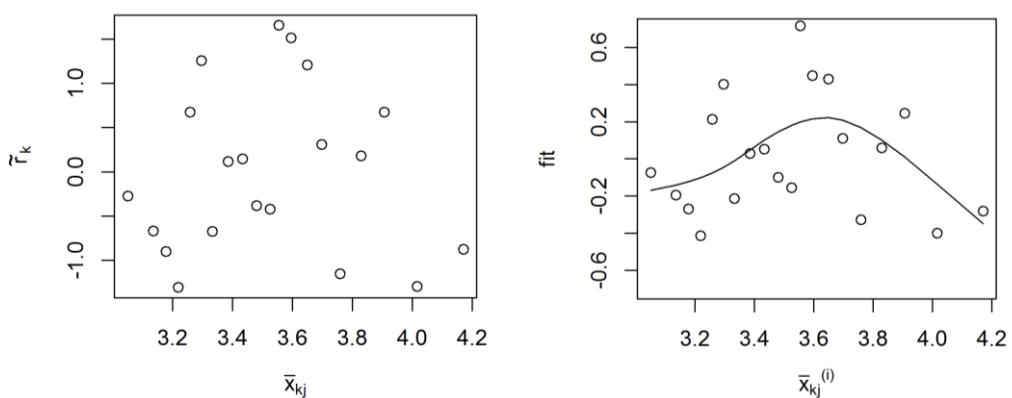
Amount

Norm. Res. for log(amount)_new_model Fit vs. Ordered Res. for log(amount)_new_mc



Age

Norm. Res. for log(age)_new_model Fit vs. Ordered Res. for log(age)_new_mc



```

D <- read.csv('credit.csv')
#summary
summary(D)
#hist + box
par(mfrow=c(2,3))
hist(D$term)
hist(D$amount)
hist(D$age)
boxplot(D$term)
boxplot(D$amount)
boxplot(D$age)
par(mfrow=c(1,1))
#log
term <- log(D$term)
amount <- log(D$amount)
age <- log(D$age)
#plot after log
par(mfrow=c(2,3))
hist(term)
hist(amount)
hist(age)
boxplot(term)
boxplot(amount)
boxplot(age)
par(mfrow=c(1,1))
#new dataframe
Dlog <- data.frame(cbind(good_credit=D$good_credit, term, amount, age, land=D$land))
pairs(Dlog[,2:4])
#log reg
LR <- glm(good_credit ~ ., data = Dlog, family = binomial(link = "logit"))
summary(LR)
#analyze residuals
y.hat <- LR$fitted.values
resid_anal(term, Dlog$good_credit, y.hat, 20, "log(term)")
resid_anal(amount, Dlog$good_credit, y.hat, 20, "log(amount)")
resid_anal(age, Dlog$good_credit, y.hat, 20, "log(age)")
#add terms of amount
D2 <- cbind(Dlog, amount.2 = amount^2, amount.3 = amount^3)
LR2 <- glm(good_credit ~ ., data = D2, family = binomial(link = "logit"))
summary(LR2)
#new residuals analysis
y.hat.new <- LR2$fitted.values
resid_anal(term, Dlog$good_credit, y.hat.new, 20, "log(term)_new_model")
resid_anal(amount, Dlog$good_credit, y.hat.new, 20, "log(amount)_new_model")
resid_anal(age, Dlog$good_credit, y.hat.new, 20, "log(age)_new_model")

resid_anal <- function(z,y,phat,nbin,varname) {
  #DIVIDE THE DATA INTO BINS BASED ON z
  n = length(y)
  zrnk = rank(z,ties.method="average")
  rnk = nbin*(zrnk/n)
  grp = trunc(rnk-0.001)+1
  #COMPUTE THE RESIDUAL STATISTICS FOR EACH BIN
  zmean = NULL
  pmean = NULL
  ymean = NULL
  ng = NULL
  bins_actual <- NULL
  for (b in 1:nbin) {
    idx = which(grp==b)

```

```

ngcur = length(idx)
if(ngcur != 0){
  zm = mean(z[idx])
  pm = mean(phat[idx])
  ym = mean(y[idx])
  ng = c(ng,ngcur)
  zmean = c(zmean,zm)
  pmean = c(pmean,pm)
  ymean = c(ymean,ym)
  bins_actual <- c(bins_actual, b)
}
}
resid = ymean-pmean
pvar = pmean*(1-pmean)/ng
nresid = resid/sqrt(pvar)
lym = log(ymean/(1-ymean))
lpm = log(pmean/(1-pmean))
lresid = lym - lpm
#PRINT OUT THE RESIDUAL STATISTICS
cbind(bins_actual,zmean,pmean,ymean,resid,nresid)
#PLOT NORMALIZED RESIDUALS VS. z TO CHECK FOR OUTLIERS
png(paste(varname,"%d.png",sep = " "), width = 1200, height = 1200, res = 300)
plot(zmean,nresid, xlab = expression(bar(x)[kj]), ylab = expression(tilde(r)[k]), main = paste("Norm. Res. for",varname))
zmin = min(zmean)
zmax = max(zmean)
lines(c(zmin,zmax),c(2,2))
lines(c(zmin,zmax),c(-2,-2))
#PLOT LOG-ODDS RESIDUALS vs. z TO CHECK FOR TREND
lc = loess.control(cell=0.6)
lyloess = loess(lresid ~ zmean, control=lc)
lfv = lyloess$fitted
zm_ord = order(zmean)
zmin = min(zmean)
zmax = max(zmean)
lmin = -0.7 #0.7 is approximately log(2)
lmax = 0.7
plot(zmean[zm_ord],lfv[zm_ord],type="l",xlim=c(zmin,zmax),ylim=c(lmin,lmax),
      xlab = expression({bar(x)[kj]}^(i)), ylab = "fit", main = paste("Fit vs. Ordered Res. for",varname))
points(zmean,lresid)
dev.off()
}

```

2. נניח כי $N = \sum_{m=1}^M Q_m$ ב"ת עבור $Q_m \sim Poi(e^{\alpha+\theta_m})$ ונתנו $\sum_{m=1}^M e^{\theta_m} = 1$, $m = 1, \dots, M$ באשר $m = 1, \dots, M$ בוגר $n = N$ הינה מולטינומית:

$$(Q|N=n) \sim Multinomial(n; e^{\theta_1}, \dots, e^{\theta_M})$$

נסמן $\sum_m e^{\theta_m} = 1, \sum_m Q_m = N$ ובן $Q_m \sim Pois(e^{\alpha+\theta_m})$

$$P(N=n) = \frac{(\sum_m \zeta_m)^n e^{-\sum_m \zeta_m}}{n!}$$

ζ_m מתייחסו לעצמו פואסון עם פרמטר $\sum_m \zeta_m$ לבן $x_1 \sim Pois(\zeta_1), \dots, x_M \sim Pois(\zeta_M)$

$$P(N=n) = \frac{(\sum_m e^{\alpha+\theta_m})^n e^{-\sum_m e^{\alpha+\theta_m}}}{n!} = \frac{e^{n\alpha} e^{-e^\alpha \sum_m e^{\theta_m}}}{n!} = \frac{e^{n\alpha} e^{-e^\alpha}}{n!}$$

ובמקרה זה באשר $\zeta_m = e^{\alpha+\theta_m}$ קיבל $\zeta_m = e^{\alpha+\theta_m}$ מכון $P(N=n)$.

בי הפילוג המותנה יהיה:

$$\begin{aligned}
P(Q_1, \dots, Q_M | N = n) &= \frac{\prod_{m=1}^M \frac{(e^{\alpha+\theta_m})^{Q_m} e^{-e^{\alpha+\theta_m}}}{Q_m!}}{\frac{e^{n\alpha} e^{-e^\alpha}}{n!}} \\
&= \frac{1}{\prod_{m=1}^M Q_m!} (e^{\alpha+\theta_1})^{Q_1} \cdot \dots \cdot (e^{\alpha+\theta_M})^{Q_M} \cdot e^{-e^{\alpha+\theta_1}} \cdot \dots \cdot e^{-e^{\alpha+\theta_M}} \cdot \frac{n!}{e^{n\alpha} e^{-e^\alpha}} \\
&= \frac{n!}{\prod_{m=1}^M Q_m!} \cdot e^{\sum_m \theta_m Q_m} \cdot \frac{e^{\alpha \sum_m Q_m} e^{-e^{\alpha+\theta_1}-\dots-e^{\alpha+\theta_M}}}{e^{n\alpha} e^{-e^\alpha}} \\
&= \frac{n!}{\prod_{m=1}^M Q_m!} \prod_{m=1}^M (e^{\theta_m})^{Q_m} \frac{e^{\alpha n} e^{-e^{\alpha}(\sum_m e^{\theta_m})}}{e^{n\alpha} e^{-e^\alpha}} = \frac{n!}{\prod_{m=1}^M Q_m!} \prod_{m=1}^M (e^{\theta_m})^{Q_m} \frac{e^{\alpha n} e^{-e^{\alpha}}}{e^{n\alpha} e^{-e^\alpha}} \\
&= \frac{n!}{\prod_{m=1}^M Q_m!} \prod_{m=1}^M (e^{\theta_m})^{Q_m} \sim Mult(n; e^{\theta_1}, \dots, e^{\theta_M})
\end{aligned}$$

3. הריצו את מודל הרגסיה הפואסונית שפורסם באתר בקובץ Poisson Regression Example. חשבו אומד עבור x^* : ($Sex = 1, Age = 0.6, Income = 0.9, HScore = 11$) $E[Y|X = x]$ ורוח ברמת סמך של 95% עבור הוקטור

$$\hat{y} = 0.6689, \quad y \in [0.5279, 0.8479]$$

```

source("pois1.r")
g <- function(u){
  return(exp(u))
}
x.star <- c(1, 1, 0.6, 0.9, 11)
theta.hat <- x.star %*% pois$coefficients
y.hat <- g(theta.hat)
alpha <- 0.05
z.crit <- qnorm(1 - alpha / 2)
c.crit <- z.crit * sqrt(t(x.star) %*% covb %*% x.star)
CI_theta <- theta.hat + c(-c.crit, c.crit)
CI <- g(CI_theta)

```

4. שאלת זו עוסקת במודלים לוג-לוג לינאריים

- א. כתבו פונקציה ב-R המתקבלת בקלט מערך תלת-ממדי של ערבי π ומחשבת את $\bar{\theta}$ ואת ערבי λ השוניים.
 ב. מערך תלת-ממדי מוגדר ב-R באופן הבא:

```
g <- array(data, c(A,B,C))
```

באשר **data** הוא וקטור המוביל את הנתונים A, B, C הם מספרי הרמות התואמים. עבור נתוני הקובץ ex10q4.csv צרו מערך תלת-ממדי תואם והריצו את הפונקציה אותה כתבתם על **g**.
 ג. ודאו כי מתקיים $P(A = i, B = j, C = k) = P(A = i)P(B = j)P(C = k)$, כלומר כי זהה המודל (AB, C) .

```
# FUNCTION TO TAKE ARRAY AND COMPUTE LOGLIN PARAMETERS
loglinpars = function(tbl) {
```

```
theta = log(tbl)
```

```
dimth = dim(theta)
```

```
I = dimth[1]
```

```
J = dimth[2]
```

```
K = dimth[3]
```

```
thb.ddd = mean(theta)
```

```
thb.idd = apply(theta, 1, mean)
thb.djd = apply(theta, 2, mean)
```

```

thb.ddk = apply(theta,3,mean)
thb.iid = apply(theta,c(1,2),mean)
thb.idk = apply(theta,c(1,3),mean)
thb.djk = apply(theta,c(2,3),mean)

lam.A = thb.idd - thbddd
lam.B = thb.djd - thbddd
lam.C = thb.ddk - thbddd

lam.AB = matrix(0,I,J)
for (i in 1:I) {
  for (j in 1:J) {
    lam.AB[i,j] = thb.iid[i,j] - thb.idd[i] - thb.djd[j] + thbddd
  }
}

lam.AC = matrix(0,I,K)
for (i in 1:I) {
  for (k in 1:K) {
    lam.AC[i,k] = thb.idk[i,k] - thb.idd[i] - thb.ddk[k] + thbddd
  }
}

lam.BC = matrix(0,J,K)
for (j in 1:J) {
  for (k in 1:K) {
    lam.BC[j,k] = thb.djk[j,k] - thb.djd[j] - thb.ddk[k] + thbddd
  }
}

lam.ABC = array(rep(0,I*J*K),dim=c(I,J,K))
for (i in 1:I) {
  for (j in 1:J) {
    for (k in 1:K) {
      lam.ABC[i,j,k] = theta[i,j,k] - (thbddd + lam.A[i] + lam.B[j] + lam.C[k]
        + lam.AB[i,j] + lam.AC[i,k] + lam.BC[j,k])
    }
  }
}

ans = list(thbddd=thbddd, lam.A=lam.A, lam.B=lam.B, lam.C=lam.C,
  lam.AB=lam.AB, lam.AC=lam.AC, lam.BC=lam.BC, lam.ABC=lam.ABC)
return(ans)

}

# SET UP DATA
indat = read.csv('ex10q4.csv',header=T)
A = indat$A
B = indat$B
C = indat$C
pr = indat$n
pr = pr/sum(pr)
pi = array(rep(0,12),dim=c(3,2,2))
for (z in 1:12) {
  pi[A[z],B[z],C[z]] = pr[z]
}

#RUN LOGLINEAR MODEL FUNCTION AND PRINT OUT RESULTS
llp = loglinpars(pi)
llp$thbddd

```

```
llp$lam.A
llp$lam.B
llp$lam.C
llp$lam.AB
llp$lam.AC
llp$lam.BC
llp$lam.ABC

# VERIFY THAT A AND B ARE JOINTLY INDEPENDENT OF C
pi.ijd = apply(pi,c(1,2),sum)
pi.ddk = apply(pi,3,sum)
pr1 = rep(1,12)
for (z in 1:12) {
  pr1[z] = pi.ijd[A[z],B[z]]*pi.ddk[C[z]]
}
pr
pr1
```