M7222 – 5. cvičení : **GLM05b** (*Trojrozměrné kontingenční tabulky*)

Příklad: Průzkum na školách

V roce 1992 byl uskutečněn průzkum na školách Wright State University School of Medicine a United Health Service in Dayton Ohio. V průzkumu odpovídalo celkem 2276 studentů posledních ročníků na dotaz, zda zkusili alkohol, cigarety a marihuanu.

Výsledky jsou dány následující tabulkou.

Alcohol Use	Cigarette Use	Marijuana Use	
		Yes	No
Yes	Yes	911	538
	No	44	456
No	Yes	3	43
	No	2	279

Vytvoříme datový rámec.

```
> Data <- data.frame(expand.grid(marijuana = factor(c("Yes", "No"), levels = c("No",
    "Yes")), cigarette = factor(c("Yes", "No"), levels = c("No", "Yes")),
    alcohol = factor(c("Yes", "No"), levels = c("No", "Yes"))), count = c(911,
    538, 44, 456, 3, 43, 2, 279))
> str(Data)
```

```
'data.frame':   8 obs. of  4 variables:
 $ marijuana: Factor w/ 2 levels "No","Yes": 2 1 2 1 2 1 2 1
 $ cigarette: Factor w/ 2 levels "No","Yes": 2 2 1 1 2 2 1 1
 $ alcohol  : Factor w/ 2 levels "No","Yes": 2 2 2 2 1 1 1 1
 $ count    : num  911 538 44 456 3 43 2 279
```

```
> Data
```

```
  marijuana cigarette alcohol count
1       Yes       Yes     Yes   911
2        No       Yes     Yes   538
3       Yes        No     Yes    44
4        No        No     Yes   456
5       Yes       Yes      No     3
6        No       Yes      No    43
7       Yes        No      No     2
8        No        No      No   279
```

Nejprve vytvoříme maximální (také označovaný jako *saturovaný*) a minimální model (*grand mean model*, *equaprobability model*):

$$GLM_{max}: \quad H_0 : EY_{jkl} = N\pi_{jkl} \Rightarrow \eta_{jkl} = \log EY_{jkl} = \mu + \alpha_j + \beta_k + \gamma_l + (\alpha\beta)_{jk} + (\alpha\gamma)_{jl} + (\beta\gamma)_{kl} + (\alpha\beta\gamma)_{jkl}$$

$$GLM_{min}: \quad H_0 : EY_{jkl} = N\pi_{jkl} \Rightarrow \eta_{jkl} = \log EY_{jkl} = \mu$$

```
> m.max <- glm(count ~ alcohol * cigarette * marijuana, data = Data, family = "poisson",
     y = TRUE)
> summary(m.max)


Call:
glm(formula = count ~ alcohol * cigarette * marijuana, family = "poisson",
    data = Data, y = TRUE)

Deviance Residuals:
[1]  0  0  0  0  0  0  0  0

Coefficients:
                                  Estimate Std. Error z value Pr(>|z|)
(Intercept)                        5.63121    0.05987  94.060  < 2e-16 ***
alcoholYes                         0.49128    0.07601   6.464 1.02e-10 ***
cigaretteYes                      -1.87001    0.16383 -11.414  < 2e-16 ***
marijuanaYes                      -4.93806    0.70964  -6.959 3.44e-12 ***
alcoholYes:cigaretteYes            2.03538    0.17576  11.580  < 2e-16 ***
alcoholYes:marijuanaYes            2.59976    0.72698   3.576 0.000349 ***
cigaretteYes:marijuanaYes          2.27548    0.92746   2.453 0.014149 *
alcoholYes:cigaretteYes:marijuanaYes  0.58951    0.94236   0.626 0.531600
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance:  2.8515e+03  on 7  degrees of freedom
Residual deviance: -2.9199e-13  on 0  degrees of freedom
AIC: 65.043

Number of Fisher Scoring iterations: 3


> m.min <- glm(count ~ 1, data = Data, family = "poisson", y = TRUE)
> summary(m.min)


Call:
glm(formula = count ~ 1, family = "poisson", data = Data, y = TRUE)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-23.349  -19.213   -9.062   10.346   29.453

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  5.65073    0.02096   269.6   <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: 2851.5  on 7  degrees of freedom
Residual deviance: 2851.5  on 7  degrees of freedom
AIC: 2902.5

Number of Fisher Scoring iterations: 5
```

Pokud z maximálního modelu odebereme interakci $(\alpha\beta\gamma)$, dostáváme tzv. model párové závislosti:

$$GLM : H_0 : EY_{jkl} = N\pi_{jkl} \Rightarrow \eta_{jkl} = \log EY_{jkl} = \mu + \alpha_j + \beta_k + \gamma_l + (\alpha\beta)_{jk} + (\alpha\gamma)_{jl} + (\beta\gamma)_{kl}$$

```
> m.AC.AM.CM <- update(m.max, . ~ . - alcohol:cigarette:marijuana)
> summary(m.AC.AM.CM)


Call:
glm(formula = count ~ alcohol + cigarette + marijuana + alcohol:cigarette +
    alcohol:marijuana + cigarette:marijuana, family = "poisson",
    data = Data, y = TRUE)

Deviance Residuals:
       1         2         3         4         5         6         7         8
 0.02044  -0.02658  -0.09256   0.02890  -0.33428   0.09452   0.49134  -0.03690

Coefficients:
                         Estimate Std. Error z value Pr(>|z|)
(Intercept)               5.63342    0.05970  94.361  < 2e-16 ***
alcoholYes                0.48772    0.07577   6.437 1.22e-10 ***
cigaretteYes             -1.88667    0.16270 -11.596  < 2e-16 ***
marijuanaYes             -5.30904    0.47520 -11.172  < 2e-16 ***
alcoholYes:cigaretteYes   2.05453    0.17406  11.803  < 2e-16 ***
alcoholYes:marijuanaYes   2.98601    0.46468   6.426 1.31e-10 ***
cigaretteYes:marijuanaYes 2.84789    0.16384  17.382  < 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: 2851.46098  on 7  degrees of freedom
Residual deviance:    0.37399  on 1  degrees of freedom
AIC: 63.417

Number of Fisher Scoring iterations: 4
```

Pokud mezi dvojitými interakcemi jednu odstraníme, máme modely tzv. podmíněné nezávislosti (*conditional independence*)

```
> m.AC.AM <- glm(count ~ (cigarette + marijuana) * alcohol, data = Data,
     family = "poisson", y = TRUE)
> summary(m.AC.AM)


Call:
glm(formula = count ~ (cigarette + marijuana) * alcohol, family = "poisson",
    data = Data, y = TRUE)

Deviance Residuals:
       1         2         3         4         5         6         7         8
  7.2238   -7.7739  -15.8396   11.3171    2.0272   -0.3442   -1.2388    0.1379
```

```
Coefficients:
                        Estimate Std. Error z value Pr(>|z|)
(Intercept)              5.62295    0.06005  93.635   <2e-16 ***
cigaretteYes            -1.80971    0.15905 -11.378   <2e-16 ***
marijuanaYes            -4.16511    0.45067  -9.242   <2e-16 ***
alcoholYes              -0.08167    0.07810  -1.046    0.296
cigaretteYes:alcoholYes  2.87373    0.16730  17.178   <2e-16 ***
marijuanaYes:alcoholYes  4.12509    0.45294   9.107   <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: 2851.46  on 7  degrees of freedom
Residual deviance:  497.37  on 2  degrees of freedom
AIC: 558.41

Number of Fisher Scoring iterations: 5


> m.AC.CM <- glm(count ~ (alcohol + marijuana) * cigarette, data = Data,
     family = "poisson", y = TRUE)
> summary(m.AC.CM)


Call:
glm(formula = count ~ (alcohol + marijuana) * cigarette, family = "poisson",
    data = Data, y = TRUE)

Deviance Residuals:
      1       2       3       4       5       6       7       8
 0.8401  -1.0667   2.4964  -0.6743  -6.0678   5.0235  -4.5440   0.8867

Coefficients:
                         Estimate Std. Error z value Pr(>|z|)
(Intercept)               5.57765    0.06032  92.463  < 2e-16 ***
alcoholYes                0.57625    0.07456   7.729 1.08e-14 ***
marijuanaYes             -2.77123    0.15199 -18.233  < 2e-16 ***
cigaretteYes             -2.69414    0.16257 -16.572  < 2e-16 ***
alcoholYes:cigaretteYes   2.87373    0.16730  17.178  < 2e-16 ***
marijuanaYes:cigaretteYes 3.22431    0.16098  20.029  < 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: 2851.461  on 7  degrees of freedom
Residual deviance:   92.018  on 2  degrees of freedom
AIC: 153.06

Number of Fisher Scoring iterations: 6


> m.AM.CM <- glm(count ~ (alcohol + cigarette) * marijuana, data = Data,
     family = "poisson", y = TRUE)
> summary(m.AM.CM)
```

```
Call:
glm(formula = count ~ (alcohol + cigarette) * marijuana, family = "poisson",
    data = Data, y = TRUE)

Deviance Residuals:
       1         2         3         4         5         6         7         8
 0.05836   4.57017  -0.26193  -4.34413  -0.86631  -9.77162   2.22872   6.83535

Coefficients:
                         Estimate Std. Error z value Pr(>|z|)
(Intercept)               5.19207    0.06088  85.285  < 2e-16 ***
alcoholYes                1.12719    0.06412  17.579  < 2e-16 ***
cigaretteYes             -0.23512    0.05551  -4.235 2.28e-05 ***
marijuanaYes             -6.62092    0.47370 -13.977  < 2e-16 ***
alcoholYes:marijuanaYes   4.12509    0.45294   9.107  < 2e-16 ***
cigaretteYes:marijuanaYes 3.22431    0.16098  20.029  < 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: 2851.46  on 7  degrees of freedom
Residual deviance:  187.75  on 2  degrees of freedom
AIC: 248.8

Number of Fisher Scoring iterations: 5
```

Ponecháme–li pouze jedinou dvojitou interakci, dostaneme modely tzv. sdružené nezávislosti:

```
> m.AC.M <- glm(count ~ alcohol * cigarette + marijuana, data = Data,
     family = "poisson", y = TRUE)
> summary(m.AC.M)


Call:
glm(formula = count ~ alcohol * cigarette + marijuana, family = "poisson",
    data = Data, y = TRUE)

Deviance Residuals:
      1       2       3       4       5       6       7       8
 11.297  -11.092  -13.996    9.045   -4.648    2.917  -14.721    8.286

Coefficients:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)             5.09053    0.06228  81.732  < 2e-16 ***
alcoholYes              0.57625    0.07456   7.729 1.08e-14 ***
cigaretteYes           -1.80971    0.15905 -11.378  < 2e-16 ***
marijuanaYes           -0.31542    0.04244  -7.431 1.08e-13 ***
alcoholYes:cigaretteYes 2.87373    0.16730  17.178  < 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: 2851.46  on 7  degrees of freedom
Residual deviance:  843.83  on 3  degrees of freedom
```

```
AIC: 902.87

Number of Fisher Scoring iterations: 6


> m.AM.C <- glm(count ~ alcohol * marijuana + cigarette, data = Data,
     family = "poisson", y = TRUE)
> summary(m.AM.C)


Call:
glm(formula = count ~ alcohol * marijuana + cigarette, family = "poisson",
    data = Data, y = TRUE)

Deviance Residuals:
       1         2         3         4         5         6         7         8
 10.6031   -4.6398  -19.7664    5.9142   -0.1592  -14.1425    0.2114   13.4104

Coefficients:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)             4.70495    0.06282  74.894   <2e-16 ***
alcoholYes              1.12719    0.06412  17.579   <2e-16 ***
marijuanaYes           -4.16511    0.45067  -9.242   <2e-16 ***
cigaretteYes            0.64931    0.04415  14.707   <2e-16 ***
alcoholYes:marijuanaYes 4.12509    0.45294   9.107   <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: 2851.46  on 7  degrees of freedom
Residual deviance:  939.56  on 3  degrees of freedom
AIC: 998.6

Number of Fisher Scoring iterations: 5


> m.CM.A <- glm(count ~ alcohol + cigarette * marijuana, data = Data,
     family = "poisson", y = TRUE)
> summary(m.CM.A)


Call:
glm(formula = count ~ alcohol + cigarette * marijuana, family = "poisson",
    data = Data, y = TRUE)

Deviance Residuals:
       1         2         3         4         5         6         7         8
  4.4691    1.7907    0.7207   -7.2723  -15.2958   -4.8889   -2.1064   13.9760

Coefficients:
                         Estimate Std. Error z value Pr(>|z|)
(Intercept)               4.65966    0.06308  73.867  < 2e-16 ***
alcoholYes                1.78511    0.05976  29.872  < 2e-16 ***
cigaretteYes             -0.23512    0.05551  -4.235 2.28e-05 ***
marijuanaYes             -2.77123    0.15199 -18.233  < 2e-16 ***
cigaretteYes:marijuanaYes  3.22431    0.16098  20.029  < 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: 2851.46  on 7  degrees of freedom
Residual deviance:  534.21  on 3  degrees of freedom
AIC: 593.26

Number of Fisher Scoring iterations: 6
```

Odstraníme–li všechny dvojité interakce, tj. máme pouze hlavní faktory , pak získáme tzv. model úplné nezávislosti–*mutually independent model*:

$$GLM : H_0 : EY_{jkl} = N\pi_{jkl} \Rightarrow \eta_{jkl} = \log EY_{jkl} = \mu + \alpha_j + \beta_k + \gamma_l$$

```
> m.A.C.M <- glm(count ~ alcohol + cigarette + marijuana, data = Data,
     family = "poisson", y = TRUE)
> summary(m.A.C.M)


Call:
glm(formula = count ~ alcohol + cigarette + marijuana, family = "poisson",
    data = Data, y = TRUE)

Deviance Residuals:
      1       2       3       4       5       6       7       8
 14.522  -7.817 -17.683   3.426 -12.440  -8.436  -8.832  19.639

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   4.17254    0.06496  64.234  < 2e-16 ***
alcoholYes    1.78511    0.05976  29.872  < 2e-16 ***
cigaretteYes  0.64931    0.04415  14.707  < 2e-16 ***
marijuanaYes -0.31542    0.04244  -7.431 1.08e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 2851.5  on 7  degrees of freedom
Residual deviance: 1286.0  on 4  degrees of freedom
AIC: 1343.1

Number of Fisher Scoring iterations: 6
```

Abychom mohli jednotlivé modely srovnat, nejprve vytvoříme pomocnou funkci

```
> lrt.test <- function(model) {
     G2 <- model$deviance
     df <- model$df.residual
     pval <- 1 - pchisq(G2, df)
```

```
    vysl <- c(dev = round(G2, 2), df, pval)
    names(vysl) <- c("dev", "df", "p-value")
    return(vysl)
}
```

Vytvoříme seznam modelů a provedeme srovnání

```
> seznam.modelu <- c("m.max", "m.AC.AM.CM", "m.AC.AM", "m.AC.CM", "m.AM.CM",
    "m.AC.M", "m.AM.C", "m.CM.A", "m.A.C.M", "m.min")
> n <- length(seznam.modelu)
> vysl <- matrix(NA, nrow = n, ncol = 3)
> for (i in 1:n) vysl[i, ] <- lrt.test(get(seznam.modelu[i]))
> rownames(vysl) <- seznam.modelu
> colnames(vysl) <- c("dev", "df", "p-value")
> vysl
```

```
              dev df   p-value
m.max        0.00  0 1.0000000
m.AC.AM.CM   0.37  1 0.5408396
m.AC.AM    497.37  2 0.0000000
m.AC.CM     92.02  2 0.0000000
m.AM.CM    187.75  2 0.0000000
m.AC.M     843.83  3 0.0000000
m.AM.C     939.56  3 0.0000000
m.CM.A     534.21  3 0.0000000
m.A.C.M   1286.02  4 0.0000000
m.min     2851.46  7 0.0000000
```

Z výsledků je zřejmé, že jediným modelem, který se významně nezhoršil oproti maximálnímu, je model párové závislosti.