Computer Assignment 3

Gabriel Lindqvist, Jiahui Li

2024-12-05

Wermuth (1976,Biometrics)reports data collected from a birth clinic,which includes information on the mother's age,her smoking habits (number of cigarettes per day),gestational age (in days)and the survival status of the child.Data were collected during a given period and neither the total sample size nor any margins were held fixed.The aim of the assignment is to find a log-linear model that can be used to gain an understanding of the association between the variables.The model should fit the data well but not be too complex to interpret.

Exercise3:1

Task1

In order to find a 'good'model, several models have to be fitted.Start with the saturated model (which has a perfect fit)and remove interaction terms in a systematic way,where higher order interactions are removed before lower order interactions.No main effect should be removed,since the interest here is the association between the variables.The goodness-of-fit of the different models should be evaluated with deviance (compared to the saturated model)and AIC,and the table below should be completed with the calculated values. The variable names to be used are:X=Mother's age,Y=Smoking habits, Z=Gestational age and V Child survival.
1.read the data

```
data3<-read.csv("data_ca3.csv")
head(data3)
```

```
##   x y z v   n
## 1 0 1 0 1  40
## 2 0 0 0 1 315
## 3 0 1 0 0   9
## 4 0 0 0 0  50
## 5 0 1 1 0   6
## 6 0 0 1 0  24
```

2. Fitting a loglinear model

```
msat<-glm(n~x*y*z*v, family=poisson(link=log), data=data3)
summary(msat)
```

2.1 First we start with the saturated model

```
##
## Call:
## glm(formula = n ~ x * y * z * v, family = poisson(link = log),
##     data = data3)
##
## Deviance Residuals:
##  [1]  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.91202    0.14142  27.662  < 2e-16 ***
## x           -0.19845    0.21069  -0.942  0.34624
## y           -1.71480    0.36209  -4.736 2.18e-06 ***
## z           -0.73397    0.24833  -2.956  0.00312 **
## v            1.84055    0.15223  12.090  < 2e-16 ***
## x:y         -0.61248    0.63679  -0.962  0.33614
## x:z         -0.34055    0.39684  -0.858  0.39082
## y:z          0.32850    0.58262   0.564  0.57286
## x:v         -0.56369    0.23317  -2.418  0.01563 *
## y:v         -0.34889    0.39911  -0.874  0.38201
## z:v          3.27844    0.25513  12.850  < 2e-16 ***
## x:y:z       -0.64028    1.29817  -0.493  0.62186
## x:y:v        0.08364    0.72896   0.115  0.90866
## x:z:v        0.17964    0.41029   0.438  0.66150
## y:z:v       -0.43281    0.60831  -0.711  0.47678
## x:y:z:v      0.78340    1.34991   0.580  0.56169
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 2.0311e+04  on 15  degrees of freedom
## Residual deviance: 2.3315e-14  on  0  degrees of freedom
## AIC: 123.97
##
## Number of Fisher Scoring iterations: 3
```

Saturated Model (`msat`)

- The saturated model includes all variables (x, y, z, v) and all interaction terms up to the fourth order.
- **Residual Deviance**: 2.3315×10^{-14} with 0 degrees of freedom, indicating a perfect fit. This is expected for a saturated model, as it uses all available degrees of freedom to capture the variability in the data.
- **AIC**: 123.97. AIC is primarily used for model comparison; for now, this value serves as a baseline.

Significant terms: - Significant main effects: y, z, v. - Significant interactions: $x : v$, $z : v$.

Non-significant terms: - Higher-order interactions ($x : y : z : v$) and some lower-order interactions ($x : y, y : z$, etc.) are not significant ($p > 0.05$).

Conclusion: The saturated model perfectly fits the data but is overly complex, containing many non-significant interactions. Simplification is necessary to improve interpretability.

```r
m3<-glm(n~(x*y*z+x*y*v+x*z*v+y*z*v), family=poisson(link=log), data=data3)
summary(m3)
```

2.2 Then all three-way interactions:

```
##
## Call:
## glm(formula = n ~ (x * y * z + x * y * v + x * z * v + y * z *
##     v), family = poisson(link = log), data = data3)
##
## Deviance Residuals:
##        1         2         3         4         5         6         7         8
##  0.07489  -0.02657  -0.15458   0.06693   0.19788  -0.09567  -0.02201   0.00745
##        9        10        11        12        13        14        15        16
## -0.14025   0.24584  -0.07339   0.03895  -0.01181  -0.41303   0.12752   0.04242
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   3.9025     0.1412  27.646  < 2e-16 ***
## x            -0.1775     0.2074  -0.856   0.3921
## y            -1.6542     0.3399  -4.867 1.13e-06 ***
## z            -0.7050     0.2422  -2.912   0.0036 **
## v             1.8515     0.1517  12.207  < 2e-16 ***
## x:y          -0.8100     0.5582  -1.451   0.1467
## x:z          -0.4152     0.3777  -1.099   0.2717
## y:z           0.1666     0.5189   0.321   0.7482
## x:v          -0.5893     0.2289  -2.575   0.0100 *
## y:v          -0.4228     0.3726  -1.135   0.2564
## z:v           3.2479     0.2485  13.067  < 2e-16 ***
## x:y:z         0.0793     0.3493   0.227   0.8204
## x:y:v         0.3397     0.5962   0.570   0.5688
## x:z:v         0.2594     0.3889   0.667   0.5047
## y:z:v        -0.2564     0.5344  -0.480   0.6314
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 2.0311e+04  on 15  degrees of freedom
## Residual deviance: 3.5935e-01  on  1  degrees of freedom
## AIC: 122.33
##
## Number of Fisher Scoring iterations: 4
```

Model with all three-way interactions (`m3`)

- This model retains only up to three-way interactions, removing the four-way interaction ($x : y : z : v$).
- **Residual Deviance**: 0.35935 with 1 degree of freedom. The residual deviance is very low, indicating the model still fits the data well despite simplification.
- **AIC**: 122.33, which is lower than the saturated model's AIC (123.97). This suggests the three-way interaction model is more parsimonious while still providing a good fit.

Significant terms: - Significant main effects: y, z, v. - Significant interactions: $x:v$, $z:v$.

Non-significant terms: - Three-way interactions ($x:y:z, x:y:v, x:z:v, y:z:v$) and some lower-order interactions ($y:z, x:z$) remain non-significant.

Conclusion: The three-way interaction model provides a simpler, nearly equivalent fit compared to the saturated model. Further simplification can be considered by removing non-significant terms.

```
#Build the two-way interactions
m2 <- glm(n ~ x*y + x*z+ x*v + y*z + y*v + z*v, family = poisson(link = log), data = data3)
summary(m2)
```

2.3 Then we try the two-way interactions and the simplest model with only main effects

```
##
## Call:
## glm(formula = n ~ x * y + x * z + x * v + y * z + y * v + z *
##     v, family = poisson(link = log), data = data3)
##
## Deviance Residuals:
##        1         2         3         4         5         6         7         8
##  0.30931  -0.00233  -0.13614  -0.20568   0.83670   0.01227  -0.15519   0.02289
##        9        10        11        12        13        14        15        16
## -0.25007  -0.31318   0.40561  -0.08498  -0.00941  -0.61636  -0.29529   0.20409
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.94097    0.12653  31.146  < 2e-16 ***
## x           -0.29142    0.17035  -1.711  0.08713 .
## y           -1.69871    0.24723  -6.871 6.38e-12 ***
## z           -0.76542    0.18367  -4.167 3.08e-05 ***
## v            1.81173    0.13478  13.443  < 2e-16 ***
## x:y         -0.41194    0.09952  -4.139 3.48e-05 ***
## x:z         -0.16718    0.09613  -1.739  0.08203 .
## x:v         -0.46385    0.18005  -2.576  0.00999 **
## y:z         -0.04727    0.14907  -0.317  0.75119
## y:v         -0.41442    0.26173  -1.583  0.11333
## z:v          3.30940    0.18462  17.926  < 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: 20311.0677  on 15  degrees of freedom
## Residual deviance:     1.7225  on  5  degrees of freedom
## AIC: 115.7
##
## Number of Fisher Scoring iterations: 4
```

```
# Build the simplest model with only main effects
m1 <- glm(n ~ x + y + z + v, family = poisson(link = log), data = data3)
summary(m1)
```

```
##
## Call:
## glm(formula = n ~ x + y + z + v, family = poisson(link = log),
##     data = data3)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -8.0999  -2.0631  -0.3999   2.3575  11.3271
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.16423    0.09088   23.81   <2e-16 ***
## x           -0.93167    0.02683  -34.72   <2e-16 ***
## y           -2.24871    0.04111  -54.69   <2e-16 ***
## z            2.31290    0.04220   54.80   <2e-16 ***
## v            3.80621    0.08283   45.95   <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: 20311.07  on 15  degrees of freedom
## Residual deviance:   377.79  on 11  degrees of freedom
## AIC: 479.76
##
## Number of Fisher Scoring iterations: 6
```

3. Stepwise model selection, using AIC

Perform Stepwise Selection Based on AIC: - Using the `step()` function, we performed stepwise selection to find a balance between model complexity and fit: - **Starting Point**: The most complex model (`saturated`). - **Scope**: From the simplest model (`m1`) to the most complex model (`saturated`). - **Criteria**: AIC (Akaike Information Criterion), which evaluates the trade-off between goodness-of-fit and model complexity. - **Direction**: Both forward and backward, allowing variables to be added or removed iteratively.

```
mstep_AIC_msatm1_forward<-step(m1, direction="forward", trace=TRUE, scope = list(upper = msat, lower = m
```

3.1 First we try from m1 to saturated model, adding a variable each time

```
## Start:  AIC=479.76
## n ~ x + y + z + v
##
##        Df Deviance    AIC
## + z:v   1    35.45 139.43
## + x:y   1   360.18 464.15
## + x:v   1   367.66 471.64
## + x:z   1   370.77 474.75
## + y:v   1   375.40 479.37
## <none>      377.79 479.76
## + y:z   1   377.26 481.24
```

```
##
## Step:  AIC=139.43
## n ~ x + y + z + v + z:v
##
##        Df Deviance    AIC
## + x:y   1   17.845 123.82
## + x:v   1   25.328 131.30
## + x:z   1   28.440 134.41
## + y:v   1   33.065 139.04
## <none>      35.454 139.43
## + y:z   1   34.929 140.90
##
## Step:  AIC=123.82
## n ~ x + y + z + v + z:v + x:y
##
##        Df Deviance    AIC
## + x:v   1   7.7197 115.69
## + x:z   1  10.8316 118.81
## + y:v   1  15.4559 123.43
## <none>     17.8453 123.82
## + y:z   1  17.3205 125.29
##
## Step:  AIC=115.69
## n ~ x + y + z + v + z:v + x:y + x:v
##
##        Df Deviance    AIC
## + y:v   1   4.7486 114.72
## + x:z   1   4.7933 114.77
## <none>      7.7197 115.69
## + y:z   1   7.1164 117.09
##
## Step:  AIC=114.72
## n ~ x + y + z + v + z:v + x:y + x:v + y:v
##
##         Df Deviance    AIC
## + x:z    1   1.8221 113.80
## <none>       4.7486 114.72
## + x:y:v  1   4.1918 116.17
## + y:z    1   4.6973 116.67
##
## Step:  AIC=113.8
## n ~ x + y + z + v + z:v + x:y + x:v + y:v + x:z
##
##         Df Deviance    AIC
## <none>       1.8221 113.80
## + x:y:v  1   1.2653 115.24
## + x:z:v  1   1.3076 115.28
## + y:z    1   1.7225 115.70
```

```
summary(mstep_AIC_msatm1_forward)
```

```
##
## Call:
## glm(formula = n ~ x + y + z + v + z:v + x:y + x:v + y:v + x:z,
```

```
##     family = poisson(link = log), data = data3)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.65265  -0.22147  -0.09177   0.07486   0.77533
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.94312    0.12624  31.236  < 2e-16 ***
## x           -0.29199    0.17033  -1.714  0.08648 .
## y           -1.71313    0.24310  -7.047 1.83e-12 ***
## z           -0.77237    0.18237  -4.235 2.28e-05 ***
## v            1.81422    0.13442  13.497  < 2e-16 ***
## z:v          3.31135    0.18452  17.945  < 2e-16 ***
## x:y         -0.41132    0.09950  -4.134 3.56e-05 ***
## x:v         -0.46481    0.18003  -2.582  0.00983 **
## y:v         -0.44375    0.24471  -1.813  0.06977 .
## x:z         -0.16557    0.09599  -1.725  0.08456 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 20311.0677  on 15  degrees of freedom
## Residual deviance:     1.8221  on  6  degrees of freedom
## AIC: 113.8
##
## Number of Fisher Scoring iterations: 4
```

Final Model (`mstep_AIC`)

1. **Formula**:
$$n \sim x + y + z + v + x:y + x:z + x:v + y:v + z:v$$
   - Includes main effects: x, y, z, v.
   - Retains significant two-way interactions: $x:y, x:z, x:v, y:v, z:v$.
   - Excludes all three-way and higher-order interactions.
2. **Fit and Complexity**:
   - **Residual Deviance**: 1.8221 with 6 degrees of freedom (excellent fit).
   - **AIC**: 113.8 (lowest among tested models).
   - **Significance**: All retained terms are significant or borderline significant, ensuring interpretability.

```
mstep_AIC_msatm1_backward<-step(msat, direction="backward", trace=TRUE, scope = list(upper = msat, lowe
```

3.2 Next try from saturated model to m1, removing a variable each time

```
## Start:  AIC=123.97
## n ~ x * y * z * v
##
```

```
##           Df Deviance    AIC
## - x:y:z:v  1  0.35935 122.33
## <none>        0.00000 123.97
##
## Step:  AIC=122.33
## n ~ x + y + z + v + x:y + x:z + y:z + x:v + y:v + z:v + x:y:z +
##     x:y:v + x:z:v + y:z:v
##
##         Df Deviance    AIC
## - x:y:z  1  0.41133 120.39
## - y:z:v  1  0.58575 120.56
## - x:y:v  1  0.69404 120.67
## - x:z:v  1  0.80918 120.78
## <none>      0.35935 122.33
##
## Step:  AIC=120.38
## n ~ x + y + z + v + x:y + x:z + y:z + x:v + y:v + z:v + x:y:v +
##     x:z:v + y:z:v
##
##         Df Deviance    AIC
## - y:z:v  1  0.63838 118.61
## - x:z:v  1  0.85318 118.83
## - x:y:v  1  0.91813 118.89
## <none>      0.41133 120.39
##
## Step:  AIC=118.61
## n ~ x + y + z + v + x:y + x:z + y:z + x:v + y:v + z:v + x:y:v +
##     x:z:v
##
##         Df Deviance    AIC
## - y:z    1  0.75076 116.72
## - x:z:v  1  1.16271 117.14
## - x:y:v  1  1.20573 117.18
## <none>      0.63838 118.61
##
## Step:  AIC=116.72
## n ~ x + y + z + v + x:y + x:z + x:v + y:v + z:v + x:y:v + x:z:v
##
##         Df Deviance    AIC
## - x:z:v  1  1.26530 115.24
## - x:y:v  1  1.30759 115.28
## <none>      0.75076 116.72
##
## Step:  AIC=115.24
## n ~ x + y + z + v + x:y + x:z + x:v + y:v + z:v + x:y:v
##
##         Df Deviance    AIC
## - x:y:v  1     1.82 113.80
## <none>         1.27 115.24
## - x:z    1     4.19 116.16
## - z:v    1   339.51 451.49
##
## Step:  AIC=113.8
## n ~ x + y + z + v + x:y + x:z + x:v + y:v + z:v
```

```
##
##        Df Deviance    AIC
## <none>        1.82 113.80
## - x:z   1     4.75 114.72
## - y:v   1     4.79 114.77
## - x:v   1     8.30 118.28
## - x:y   1    20.01 129.99
## - z:v   1   340.07 450.04
```

```
summary(mstep_AIC_msatm1_backward)
```

```
##
## Call:
## glm(formula = n ~ x + y + z + v + x:y + x:z + x:v + y:v + z:v,
##     family = poisson(link = log), data = data3)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.65265  -0.22147  -0.09177   0.07486   0.77533
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.94312    0.12624  31.236  < 2e-16 ***
## x           -0.29199    0.17033  -1.714  0.08648 .
## y           -1.71313    0.24310  -7.047 1.83e-12 ***
## z           -0.77237    0.18237  -4.235 2.28e-05 ***
## v            1.81422    0.13442  13.497  < 2e-16 ***
## x:y         -0.41132    0.09950  -4.134 3.56e-05 ***
## x:z         -0.16557    0.09599  -1.725  0.08456 .
## x:v         -0.46481    0.18003  -2.582  0.00983 **
## y:v         -0.44375    0.24471  -1.813  0.06977 .
## z:v          3.31135    0.18452  17.945  < 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: 20311.0677  on 15  degrees of freedom
## Residual deviance:     1.8221  on  6  degrees of freedom
## AIC: 113.8
##
## Number of Fisher Scoring iterations: 4
```

Final Model (`mstep_AIC`)

1. **Formula**:
$$n \sim x + y + z + v + x:y + x:z + x:v + y:v + z:v$$
   - Includes main effects: x, y, z, v.
   - Retains significant two-way interactions: $x:y, x:z, x:v, y:v, z:v$.
   - Excludes all three-way and higher-order interactions.
2. **Fit and Complexity**:

- **Residual Deviance**: 1.8221 with 6 degrees of freedom (excellent fit).
- **AIC**: 113.8 (lowest among tested models).
- **P value**: 0.94, which means we could accept this model

4. The final sheet——from m1 to saturated model

Model	Deviance	df	AIC	p_value
x * y * z * v (saturated model)	0	0	123.97	NA
x + y + z + v + x:y + x:z + y:z + x:v + y:v + z:v + x:y:z + x:y:v + x:z:v + y:z:v	0.35935	1	122.33	0.55
x + y + z + v + x:y + x:z + y:z + x:v + y:v + z:v + x:y:v + x:z:v + y:z:v	0.41133	2	120.39	0.81
x + y + z + v + x:y + x:z + y:z + x:v + y:v + z:v + x:y:v + x:z:v	0.63838	3	118.61	0.89
x + y + z + v + x:y + x:z + x:v + y:v + z:v + x:y:v + x:z:v	0.75076	4	116.72	0.94
x + y + z + v + x:y + x:z + x:v + y:v + z:v + x:z:v	1.30759	5	115.28	0.93
x + y + z + v + z:v + x:y + x:v + y:v + x:z	1.8221	6	113.8	0.94
x + y + z + v + z:v + x:y + x:v + y:v	4.7486	7	114.72	0.69
x + y + z + v + z:v + x:y + x:v	7.7197	8	115.69	0.46
x + y + z + v + z:v + x:y	17.845	9	123.82	0.037
x + y + z + v + z:v	35.45	10	139.43	1.046e-4
x + y + z + v	377.79	11	479.76	0

Task2

Choose from your table a model with few parameters and a good fit.Describe the procedure to compare different models.

The best model chosen from task1

```
mbest <- glm(n ~ x + y + z + v + x:y + x:z + x:v + y:v + z:v, family = poisson(link = log), data = data3
summary(mbest)
```

```
##
## Call:
## glm(formula = n ~ x + y + z + v + x:y + x:z + x:v + y:v + z:v,
##     family = poisson(link = log), data = data3)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.65265  -0.22147  -0.09177   0.07486   0.77533
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.94312    0.12624  31.236  < 2e-16 ***
## x           -0.29199    0.17033  -1.714  0.08648 .
## y           -1.71313    0.24310  -7.047 1.83e-12 ***
## z           -0.77237    0.18237  -4.235 2.28e-05 ***
## v            1.81422    0.13442  13.497  < 2e-16 ***
## x:y         -0.41132    0.09950  -4.134 3.56e-05 ***
```

```
## x:z         -0.16557    0.09599  -1.725  0.08456 .
## x:v         -0.46481    0.18003  -2.582  0.00983 **
## y:v         -0.44375    0.24471  -1.813  0.06977 .
## z:v          3.31135    0.18452  17.945  < 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: 20311.0677  on 15  degrees of freedom
## Residual deviance:     1.8221  on  6  degrees of freedom
## AIC: 113.8
##
## Number of Fisher Scoring iterations: 4
```

LR-test between two nested models

```
anova(mbest,msat,test="LRT")
```

```
## Analysis of Deviance Table
##
## Model 1: n ~ x + y + z + v + x:y + x:z + x:v + y:v + z:v
## Model 2: n ~ x * y * z * v
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1         6     1.8221
## 2         0     0.0000  6   1.8221   0.9353
```

From the result, we can see that:

Fit and Complexity: - **Residual Deviance**: 1.8221 with 6 degrees of freedom (excellent fit). - **P value**: 0.94, which means we could accept this model with a good fit.

Procedure to Compare Models

1. **Define Candidate Models**:
   - **Most Complex Model (`saturated`)**: Includes all main effects and interactions:
   $$n \sim (x \times y \times z \times v)$$
   - **Simplest Model (`m1`)**: Includes only the main effects:
   $$n \sim x + y + z + v$$
2. **Perform Stepwise Selection Based on AIC**:
   - Using the `step()` function, we performed stepwise selection to find a balance between model complexity and fit:
     - **Starting Point**: The most complex model (`saturated`).
     - **Scope**: From the simplest model (`m1`) to the most complex model (`saturated`).
     - **Criteria**: AIC (Akaike Information Criterion), which evaluates the trade-off between goodness-of-fit and model complexity.

- **Direction**: Both forward and backward, allowing variables to be added or removed iteratively.

3. **Evaluate Models Using Residual Deviance and AIC**:
   - For each model, the residual deviance (measuring the lack of fit) and AIC were calculated.
   - Models with lower residual deviance fit the data better, and models with lower AIC balance fit and simplicity.
4. **Compare Models Using Statistical Tests**:
   - **Likelihood Ratio Test (LRT)**:
     - Compared nested models (e.g., `saturated` vs. `m1`) to assess whether removing higher-order terms significantly worsens the fit.
     - Significant p-values ($p < 0.05$) indicate that removing terms leads to a significant loss of fit. $P = 1 - \text{chi2.cdf(Deviance, Df)}$
5. **Select the Final Model**:
   - The final model (`mstep_AIC`) was chosen based on its low AIC, good residual deviance, and simplicity. It balances fit and complexity effectively. In this example, the saturated model is the most complex model that can be considered and m1 is the simplest.

Task3

Interpret the model you chose. Which associations are significant? Quantify the associations with odds ratios together with confidence intervals.

Interpret the model we chose

```
summary(mbest)
```

```
##
## Call:
## glm(formula = n ~ x + y + z + v + x:y + x:z + x:v + y:v + z:v,
##     family = poisson(link = log), data = data3)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.65265  -0.22147  -0.09177   0.07486   0.77533
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.94312    0.12624  31.236  < 2e-16 ***
## x           -0.29199    0.17033  -1.714  0.08648 .
## y           -1.71313    0.24310  -7.047 1.83e-12 ***
## z           -0.77237    0.18237  -4.235 2.28e-05 ***
## v            1.81422    0.13442  13.497  < 2e-16 ***
## x:y         -0.41132    0.09950  -4.134 3.56e-05 ***
## x:z         -0.16557    0.09599  -1.725  0.08456 .
## x:v         -0.46481    0.18003  -2.582  0.00983 **
## y:v         -0.44375    0.24471  -1.813  0.06977 .
## z:v          3.31135    0.18452  17.945  < 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: 20311.0677  on 15  degrees of freedom
## Residual deviance:     1.8221  on  6  degrees of freedom
## AIC: 113.8
##
## Number of Fisher Scoring iterations: 4
```

1. Chosen Model

$$n \sim x + y + z + v + x : y + x : z + x : v + y : v + z : v$$

Variables:

- x: Mother's age.
- y: Smoking habits.
- z: Gestational age.
- v: Child survival.
- Significant two-way interactions:
  - $x : y$: Interaction between mother's age and smoking habits.
  - $x : z$: Interaction between mother's age and gestational age.
  - $x : v$: Interaction between mother's age and child survival.
  - $y : v$: Interaction between smoking habits and child survival.
  - $z : v$: Interaction between gestational age and child survival.

---

2. Significant Associations

Main Effects:

1. x **(Mother's Age)**:
   - Coefficient: -0.29199, $p = 0.08648$ (marginal significance).
   - Interpretation: Increasing mother's age slightly decreases the likelihood of the outcome.
2. y **(Smoking Habits)**:
   - Coefficient: -1.71313, $p < 0.001$.
   - Interpretation: Smoking significantly decreases the likelihood of the outcome.
3. z **(Gestational Age)**:
   - Coefficient: -0.77237, $p < 0.001$.
   - Interpretation: Shorter gestational age significantly decreases the likelihood of the outcome.
4. v **(Child Survival)**:
   - Coefficient: 1.81422, $p < 0.001$.
   - Interpretation: Child survival significantly increases the likelihood of the outcome.

Two-Way Interactions:

1. $x : y$:
   - Coefficient: -0.41132, $p < 0.001$.
   - Interpretation: The negative effect of smoking is stronger for older mothers.
2. $x : z$:
   - Coefficient: -0.16557, $p = 0.08456$ (marginal significance).
   - Interpretation: The relationship between gestational age and outcome weakens slightly for older mothers.
3. $x : v$:
   - Coefficient: -0.46481, $p = 0.00983$.
   - Interpretation: The positive effect of child survival is weaker for older mothers.
4. $y : v$:
   - Coefficient: -0.44375, $p = 0.06977$ (marginal significance).
   - Interpretation: The positive effect of child survival is weaker for smokers.
5. $z : v$:
   - Coefficient: 3.31135, $p < 0.001$.
   - Interpretation: The interaction between gestational age and child survival is highly significant, indicating that survival outcomes improve strongly with longer gestational age.

Quantify the associations with odds ratios together with confidence intervals.

```
# Calculate Odds Ratios and Confidence Intervals
exp_coef <- exp(coef(mbest)) # Odds Ratios
confint_vals <- confint(mbest) # Confidence Intervals
```

```
## Waiting for profiling to be done...
```

```
exp_confint <- exp(confint_vals) # Exponentiate to get OR CIs

# Combine results into a table
results <- data.frame(
  Term = names(exp_coef),
  Odds_Ratio = exp_coef,
  CI_Lower = exp_confint[, 1],
  CI_Upper = exp_confint[, 2]
)

# Display results
print(results)
```

```
##                    Term Odds_Ratio   CI_Lower   CI_Upper
## (Intercept) (Intercept) 51.5792455 39.9295713 65.5213007
## x                     x  0.7467764  0.5326970  1.0400901
## y                     y  0.1802998  0.1087523  0.2834862
## z                     z  0.4619169  0.3203070  0.6558851
## v                     v  6.1363084  4.7455095  8.0416809
```

```
## x:y                x:y  0.6627767  0.5436097  0.8031437
## x:z                x:z  0.8474089  0.7031862  1.0246687
## x:v                x:v  0.6282533  0.4425315  0.8973840
## y:v                y:v  0.6416234  0.4066629  1.0667429
## z:v                z:v 27.4220351 19.2297650 39.7066300
```

Key Findings

1. **Main Effects**:
   - Smoking (y), gestational age (z), and child survival (v) are strongly associated with the outcome, with significant odds ratios.
   - Mother's age (x) has a weaker, marginally significant association.
2. **Interactions**:
   - Significant interactions ($x : y, x : v, z : v$) suggest complex relationships between variables:
     - Smoking's negative effect increases with maternal age.
     - The positive effect of child survival decreases with maternal age and smoking.
     - Gestational age strongly amplifies the effect of child survival, highlighting its critical role.
3. **Statistical Significance**:
   - Terms with confidence intervals that do not include 1 are statistically significant. Marginal associations ($x, x : z, y : v$) should be interpreted cautiously.