Computer Assignment 3

Gabriel Lindqvist, Jiahui Li

2024-12-09

Exercise 3:1

Task 1

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

We start by reading the data.

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

2. Fitting a loglinear Model

The goal is to fill the table provided in the assignment. We start by fitting 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:

- 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.

Next, we fit the model (XYZ, XYV, XZV, YVZ), calculate the AIC and perform a LR-test to find the deviance and the p-value

```
## [1] "AIC: 122.33256854077"
```

```
## Analysis of Deviance Table
##
## Model 1: n ~ (x * y * z + x * y * v + x * z * v + y * z * v)
## Model 2: n ~ x * y * z * v
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1         1    0.35935
## 2         0    0.00000  1  0.35935   0.5489
```

Model with all three-way interactions:

- **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.
- **p-value**: The p-value is non-significant, when testing the submodel against the saturated model. That is, we retain the submodel.
- **Conclusion**: The three-way interaction model provides a simpler, nearly equivalent fit compared to the saturated model. Further simplification can be considered by removing potential non-significant terms.

Model with up to two-way interactions and the simplest model:

```
## [1] "AIC: 115.695754709083"
```

```
## Analysis of Deviance Table
##
## Model 1: n ~ x * y + x * z + x * v + y * z + y * v + z * v
## Model 2: n ~ x * y * z * v
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1         5     1.7225
## 2         0     0.0000  5   1.7225    0.886
```

```
## [1] "AIC: 479.761190891547"
```

```
## Analysis of Deviance Table
##
## Model 1: n ~ x + y + z + v
## Model 2: n ~ x * y * z * v
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1        11     377.79
## 2         0       0.00 11   377.79 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

3. Stepwise Model Selection Using AIC

We start by implementing the `step()` function to find the best model between the model with all three-way interactions and the model with all the two-way interactions.

```
mstep <- step(m3, direction="both", trace=TRUE, scope = list(upper = m3, lower = m2))
```

```
## Start:  AIC=122.33
## n ~ (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
## + x:y:z  1  0.35935 122.33
##
## 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
## - x:z:v  1  1.16271 117.14
## - x:y:v  1  1.20573 117.18
## <none>      0.63838 118.61
## + y:z:v  1  0.41133 120.39
## + x:y:z  1  0.58575 120.56
##
## Step:  AIC=117.14
## n ~ x + y + z + v + x:y + x:z + y:z + x:v + y:v + z:v + x:y:v
##
##         Df Deviance    AIC
## - x:y:v  1  1.72254 115.70
## <none>      1.16271 117.14
## + x:z:v  1  0.63838 118.61
## + y:z:v  1  0.85318 118.83
## + x:y:z  1  1.11631 119.09
##
## Step:  AIC=115.7
## n ~ x + y + z + v + x:y + x:z + y:z + x:v + y:v + z:v
##
##         Df Deviance    AIC
## <none>       1.7225 115.70
## + x:y:v  1   1.1627 117.14
## + x:z:v  1   1.2057 117.18
## + y:z:v  1   1.3843 117.36
## + x:y:z  1   1.4875 117.46
```

From the output above, the best model (relative to AIC) between (XYZ, XYV, XZV, YZV) and (XY, XZ, XV, YZ, YV, ZV) is

$$(XYV, XZ, YZ, ZV),$$

with an AIC of 117.14. A LR-test between the simpler model (stated above) against the saturated model, yield the p-value $p = 0.8842$ and a deviance of 117.14.

Next, we implement the same function but change the scope argument `scope = list(upper = m2, lower = m1)` and the starting model.

```
mstep <- step(m2, direction="both", trace=TRUE, scope = list(upper = m2, lower = m1))
```

```
## Start:  AIC=115.7
## n ~ x * y + x * z + x * v + y * z + y * v + z * v
##
##        Df Deviance    AIC
## - y:z   1     1.82 113.80
## <none>        1.72 115.70
## - y:v   1     4.05 116.03
## - x:z   1     4.70 116.67
## - x:v   1     8.17 120.15
## - x:y   1    19.96 131.93
## - z:v   1   339.33 451.30
##
## 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
## + y:z   1     1.72 115.70
## - x:v   1     8.30 118.28
## - x:y   1    20.01 129.99
## - z:v   1   340.07 450.04
```

From the output above, we see that the best model (relative to AIC) between (X, Y, Z, V) and (XY, XZ, XV, YZ, YV, ZV) is the submodel when removing the interaction term `y:z`. That is, using compact notation

$$(XY, XZ, XV, YV, ZV),$$

with an AIC of 113.80. A LR-test between the simpler model (stated above) against the saturated model, yield the p-value $p = 0.94$ and a deviance of 1.8221.

4. The Final Table

Model	Deviance	df	AIC	p_value
$(XYZV)$	0	0	123.97	NA
(XYZ, XYV, XZV, YZV)	0.35935	1	122.33	0.55
(XYV, XZ, YZ, ZV)	1.1627	4	117.14	0.88
(XY, XZ, XV, YZ, YV, ZV)	7.7197	5	115.69	0.89
(XY, XZ, XV, YV, ZV)	1.8221	6	113.80	0.94
(X, Y, Z, V)	377.79	11	479.76	0

4.1 Conclusion Based on the Table Most notably, the simplest model (X, Y, Z, V) shows the worst fit with a relative high deviance and AIC. In contast, the best fitted model (relative to satured model) is the model found in row 2, with the lowest deviance. However, the lowered AIC value is not relative substantial, meaning an increase in goodness of fit does not justify an increase in complexity. Which is why the best model is (XY, XZ, XV, YV, ZV). Though a small increase in deviance, the model has the lowest AIC and largest p-value. That is, the trade-off in 'slightly' worse fit is not significant.

Task 2

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

From the conclusions drawn at the end of Task 1, the best model to choose is (XY, XZ, XV, YV, ZV).

Procedure to Compare Models

1. **Define Candidate Models**:
   - **Saturated model**: 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 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.

Task 3

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

Interpret the model we chose

```
##
## 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.
- **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.

Significant 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.

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

```
##                     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.

Task 4

To find a good model, we systematically try all possible combinations of interaction terms starting with the full model. After checking all the models, we found that no interaction terms were significant for each combination. Thus, we landed on the model which only includes the main effects. That is,

$$\text{logit}\big[P(V=1)\big] = \beta_0 + \beta_1 X + \beta_2 Y + \beta_3 Z.$$

For easier readability, we generate a table with relevant results.

```
##             Coefficient Std_Error Odds_Ratio CI_Lower CI_Upper P_value
## (Intercept)   1.8138997 0.1350625     6.1343   4.7076   7.9934  0.0000
## x            -0.4674675 0.1803507     0.6266   0.4400   0.8923  0.0095
## y            -0.4227830 0.2623859     0.6552   0.3918   1.0958  0.1071
## z             3.3097528 0.1846161    27.3784  19.0660  39.3148  0.0000
```

We are now ready to interpret the model.

Interpretation of the Table

1. **Intercept**: The estimated odds ratio is roughly 6.13 and is significant. With the baseline levels for our predictors being $x < 30, y < 5, z < 260$, the odds of child survival when the mother is younger than 30, smokes less than 5 cigarettes/day and the gestational age is less than 260 days approximately 6 times higher than the odds of not surviving.
2. **Mother's Age** (x): The estimated odds ratio is roughly 0.63 and is significant. Thus being 30+ years old is associated with approximately 37% lower odds of child survival compared to mothers under the age of 30.
3. **Smoking Habits** (y): 5+ cigarettes/day is associated with 34% lower odds of child survival compared to smoking fewer than 5 cigarettes/day. However, this effect is not statistically significant ($p = 0.1071$), suggesting insufficient evidence to confirm a direct impact of smoking on sural in this dataset.
4. **Gestational Age** (z): Gestational age is the most significant predictor of child survival. Births at ≥ 260 days are associated with odds of survival that are approximately 27 times greater than the odds for births at < 260 days.

The non-significance of smoking habits are rather surprising, since smoking is usually associated with health risks. The non-significance could be due to some limitations such as sparse counts; which could limit the statistical power to detect an effect.

Task 5

The equivalent log-linear model is (XYZ, XV, YV, ZV), where the two-way interactions `x:v`, `y:v`, `z:v` are of main interest. This is due to logistic models not assuming anything regarding relationships among the predictors.

We generate a table a table that includes the coefficients (estimates) and standard error for these three interaction terms.

```
##     Interaction_Term Coefficient Std_Error
## x:v              x:v  -0.4674675 0.1803510
## y:v              y:v  -0.4227830 0.2623864
## z:v              z:v   3.3097528 0.1846168
```

Comparing the table with the table presented under Task 4, we observe that the standard error for `z:v` is 'identical' to the standard error for the main effect `z` in the logistic model. The estimates and standard errors for the other terms are approximately the same, likely due to differences in numerical precision and the parameterization of the models.