AssignmentIII

Log-Linear Models

Silpa Soni Nallacheruvu (19980824-5287) Hernan Aldana (20000526-4999)

2024-12-04

Exercise 3:1 (Higher Dimension Table)

Question1:

Fit several models in order to find a 'good' model for the given 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.

Table 1: Data from the study on the association of variables with child survival

Mother's age	Smoking habits	Gestational age	Child survival - No	Child survival - Yes
< 30	< 5	< 260	50	315
< 30	< 5	>= 260	24	4012
< 30	5+	< 260	9	40
< 30	5+	>= 260	6	459
30+	< 5	< 260	41	147
30+	< 5	>= 260	14	1594
30+	5+	< 260	4	11
30+	5+	>= 260	1	124

Approach:

1. **Read the data:**
   - The given dataset 'data_ca3.csv' contains the variables X, Y, Z, and V, along with their corresponding frequencies (n).
2. **Fit a saturated model:**
   - Fit a saturated model which includes all four variables X, Y, Z, and V and all their interactions. It fits the data perfectly and serves as the reference model.
3. **Reduced Models:**
   - We started by removing the 4 way interaction term from the saturated model.
   - Then, we removed the 3 way interaction terms, then 2 way interaction terms, and finally we fit a model with the only the main effects.
   - We removed interactions in a systematic way, where higher order interactions are removed before lower order interactions to evaluate the effect of each interaction term on the model.

4. **Model Comparison:**

- We compared the models using deviance, degrees of freedom, p-value, and AIC.
- We calculated the p-value using the chi-square distribution calculated from the deviance and degrees of freedom.

R Output:

Table 2: Model Comparison Results

Model	Deviance	df	p-value	AIC
XYZV	0.000000e+00	0	0.0000000	123.9732
XYZ,XYV,XZV,YZV	3.593495e-01	1	0.5488677	122.3326
XYZ,XYV,XZV	5.857524e-01	2	0.7461145	120.5590
XYZ,XYV,YZV	8.091794e-01	2	0.6672505	120.7824
XYZ,XZV,YZV	6.940357e-01	2	0.7067927	120.6673
XYV,XZV,YZV	4.113326e-01	2	0.8141047	120.3846
XYZ,XYV	3.386172e+02	4	0.0000000	454.5905
XYZ,XZV	3.383920e+00	4	0.4957465	119.3571
XYZ,YZV	7.677148e+00	4	0.1041468	123.6504
XYV,XZV	7.507596e-01	4	0.9449248	116.7240
XYV,YZV	3.796771e+00	4	0.4342079	119.7700
XZV,YZV	1.910308e+01	4	0.0007501	135.0763
XYZ	8.414233e+03	8	0.0000000	8522.2062
XYV	5.696758e+03	8	0.0000000	5804.7315
XZV	5.203361e+03	8	0.0000000	5311.3341
YZV	1.372281e+03	8	0.0000000	1480.2544
XY,XZ,XV,YZ,YV,ZV	1.722536e+00	5	0.8860485	115.6958
XY,XZ,XV,YZ,YV	3.393279e+02	6	0.0000000	451.3011
XY,XZ,XV,YZ,ZV	4.052266e+00	6	0.6696034	116.0255
XY,XZ,XV,YV,ZV	1.822126e+00	6	0.9353091	113.7953
XY,XZ,YZ,YV,ZV	8.174676e+00	6	0.2255834	120.1479
XY,XV,YZ,YV,ZV	4.697321e+00	6	0.5831776	116.6705
XZ,XV,YZ,YV,ZV	1.996134e+01	6	0.0028136	131.9346
XY,XZ,XV,YZ	3.422990e+02	7	0.0000000	452.2722
XY,XZ,XV,YV	3.400689e+02	7	0.0000000	450.0421
XY,XZ,XV,ZV	4.793251e+00	7	0.6851769	114.7665
XY,XZ,YZ,YV	3.500352e+02	7	0.0000000	460.0084
XY,XZ,YZ,ZV	1.009059e+01	7	0.1834979	120.0638
XY,XZ,YV,ZV	8.301895e+00	7	0.3067273	118.2751
XY,XV,YZ,YV	3.465578e+02	7	0.0000000	456.5310
XY,XV,YZ,ZV	7.116352e+00	7	0.4168669	117.0896
XY,XV,YV,ZV	4.748606e+00	7	0.6906107	114.7218
XY,YZ,YV,ZV	1.540461e+01	7	0.0311488	125.3778
XZ,XV,YZ,YV	3.577377e+02	7	0.0000000	467.7109
XZ,XV,YZ,ZV	2.187726e+01	7	0.0026672	131.8505
XZ,XV,YV,ZV	2.001263e+01	7	0.0055425	129.9858
XZ,YZ,YV,ZV	2.599966e+01	7	0.0005037	135.9729
XV,YZ,YV,ZV	2.288782e+01	7	0.0017830	132.8610
XY,XZ,XV	3.430400e+02	8	0.0000000	451.0132
XY,XZ,YZ	8.414387e+03	9	0.0000000	8520.3601
XY,XZ,YV	3.507761e+02	8	0.0000000	458.7494
XY,XZ,ZV	1.083157e+01	8	0.2114264	118.8048
XY,XV,YZ	3.495289e+02	8	0.0000000	457.5022

Model	Deviance	df	p-value	AIC
XY,XV,YV	5.697315e+03	9	0.0000000	5803.2883
XY,XV,ZV	7.719730e+00	8	0.4613155	115.6929
XY,YZ,YV	3.572651e+02	8	0.0000000	465.2383
XY,YZ,ZV	1.732052e+01	8	0.0269393	125.2937
XY,YV,ZV	1.545590e+01	8	0.0508639	123.4291
XZ,XV,YZ	3.601240e+02	8	0.0000000	468.0972
XZ,XV,YV	3.582594e+02	8	0.0000000	466.2326
XZ,XV,ZV	5.203875e+03	9	0.0000000	5309.8486
XZ,YZ,YV	3.678602e+02	8	0.0000000	475.8334
XZ,YZ,ZV	2.791558e+01	8	0.0004906	135.8888
XZ,YV,ZV	2.605095e+01	8	0.0010294	134.0242
XV,YZ,YV	3.647483e+02	8	0.0000000	472.7215
XV,YZ,ZV	2.480373e+01	8	0.0016782	132.7770
XV,YV,ZV	2.293911e+01	8	0.0034433	130.9123
YZ,YV,ZV	1.372625e+03	9	0.0000000	1478.5981
XY,XZ	8.415128e+03	10	0.0000000	8519.1010
XY,XV	5.700286e+03	10	0.0000000	5804.2594
XY,YZ	8.421617e+03	10	0.0000000	8525.5900
XY,YV	5.708022e+03	10	0.0000000	5811.9956
XY,ZV	1.784527e+01	9	0.0370117	123.8185
XZ,XV	5.542122e+03	10	0.0000000	5646.0953
XZ,YZ	8.432212e+03	10	0.0000000	8536.1850
XZ,YV	3.683849e+02	9	0.0000000	474.3581
XZ,ZV	5.209914e+03	10	0.0000000	5313.8869
XV,YZ	3.671377e+02	9	0.0000000	473.1109
XV,YV	5.715506e+03	10	0.0000000	5819.4788
XV,ZV	5.206802e+03	10	0.0000000	5310.7751
YZ,YV	1.714485e+03	10	0.0000000	1818.4586
YZ,ZV	1.374541e+03	10	0.0000000	1478.5140
YV,ZV	1.372676e+03	10	0.0000000	1476.6494
XY	1.377237e+04	12	0.0000000	13872.3473
XZ	1.361421e+04	12	0.0000000	13714.1832
XV	1.089937e+04	12	0.0000000	10999.3416
YZ	9.778837e+03	12	0.0000000	9878.8103
YV	7.065243e+03	12	0.0000000	7165.2159
ZV	6.556539e+03	12	0.0000000	6656.5121
X,Y,Z,V	3.777880e+02	11	0.0000000	479.7612

Conclusion:

1. **Trends in AIC Across Models:**
   - Models with only main effects or single two-way interactions have high AICs, indicating poor fit.
   - Adding specific two-way and three-way interactions significantly improves the AIC, highlighting the importance of these interaction terms for explaining the data.

2. **Impact of Higher-Order Interactions:**
   - Models including higher-order interactions (e.g., XYZV) exhibit extremely low p-values, suggesting overfitting and unnecessary complexity. The saturated model achieves perfect fit but at the cost of increased complexity, as reflected in its AIC.
   - Models with XYV, XZV as interactions have the highest p-values, indicating that removing some of the three-way interactions does not significantly reduce the model fit. These models are candidates for simpler yet statistically robust options.

Question2:

Approach:

To answer question 2, we aim to identify a good model that balances simplicity and fit. A good model should:

- Explain the Relationships: Capture significant associations between variables.
- Avoid Overfitting: Include only necessary interactions to prevent overfitting.
- Optimize Fit: Minimize AIC and retain a good fit as assessed by likelihood ratio tests.

We assume that higher-order interactions (e.g., 3-way and 4-way) might not significantly contribute to explaining the relationships, as they may be too complex and difficult to interpret, as well as prone to overfitting. Therefore our focus will be on identifying the best model with two-way interactions. To do this we have to follow this steps:

- Start with the Saturated Model: Fit a saturated model with all interactions to establish a baseline.
- Fit Reduced Models: Gradually remove higher-order interactions (3-way and 4-way) and evaluate the impact on model fit.
- Evaluate models: Compare models based on AIC, p-values, and goodness-of-fit, and p-values from LRT.
- Select the Best Model: Identify the model with the lowest AIC that balances fit and complexity.

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

```
##      df      AIC
## msat 16 123.9732
## m1   10 113.7953
```

```
## [1] "Likelihood Ratio Test Results:"
```

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

```
##
## Call:
## glm(formula = n ~ x * y + x * z + x * v + y * v + z * v, family = poisson(link = "log"),
##     data = data3)
##
## 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
```

Results

AIC Results:

Saturate model (msat): AIC = 123.9732 Reduced model (m1): AIC = 113.7953

The reduced model (m1) has a lower AIC value compared to the saturated model, indicating a better balance between goodness-of-fit with fewer parameters.

Residual Deviance:

Saturation model (msat): 0.0000 (perfect fit) Reduced model (m1): 1.8221

Model coefficients

All coefficients in the reduced model (m1) are statistically significant ($p < 0.05$), with the exception of a few borderline cases.

The chosen model(m1) includes significant two-way interactions XY, XZ, XV, YV, ZV. This interactions highlight the relationships between:

- Mother's age and smoking habits (XY): Indicates that smoking habits vary significantly accross maternal age groups.
- Mother's age and gestational age (XZ): Suggests that gestational age may be influenced by maternal age.
- Mother's age and child survival (XV): Indicates that child survival is critically dependent on maternal age.
- Smoking habits and child survival (YV): Shows the direct impact of smoking habits on child survival rates.
- Gestational age and child survival (ZV): Reinforces that gestational age is a key factor in determining child survival rates.

Conclusion:

The reduced model (m1) with interactions XY, XZ, XV, YV, ZV was selected as the best model because:

- It has the lowest AIC.
- It retains statistically significan two way interactions that provide interpretable results inot the relationships between variables.
- Removing higher-order interactions (3-way and 4-way) did not significantly impact the model fit, as evidenced by the residual deviance and p-values.

this fits better in question 2 rather as a conclusion in question 1

The model with the lowest AIC is the one with interactions XY, XZ, XV, YV, ZV, suggesting it balances goodness-of-fit and model complexity most effectively. This model includes all two-way interactions except YZ, capturing significant dependencies among variables while avoiding overfitting.

Recommendation for a ‘Good’ Model: Based on the results, the model with XY, XZ, XV, YV, ZV is recommended for its optimal balance of simplicity and fit (lowest AIC). However, models like XYV, XZV might also be considered if further parsimony is desired, given their statistical significance.