AssignmentIII

Log-Linear Models

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

2024-12-06

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(X)	Smoking habits(Y)	Gestational age(Z)	Child survival(V) - No	Child survival(V) - 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 (Akaike Information Criterion) to evaluate the goodness of fit and complexity of the models.
- 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	9.64339717002902e-13	0	0	123.973219039056
XYZ,XYV,XZV,YZV	0.35934950171435	1	0.548867727694106	122.33256854077
XYZ,XYV,YZV	0.809179394602827	2	0.667250529393751	120.782398433659
.	.	.	.	.
XYV,XZV,YZV	0.411332636961655	2	0.81410468265266	120.384551676019
XYZ,XYV	338.617237695092	4	0	454.590456734149
.	.	.	.	.
XYV,XZV	0.750759621405906	4	0.944924837430706	116.723978660462
XYZ	8414.23302576947	8	0	8522.20624480853
.	.	.	.	.
YZV	1372.28115330706	8	0	1480.25437234612
XY,XZ,XV,YZ,YV,ZV	1.72253567002695	5	0.886048506369316	115.695754709083
XY,XZ,XV,YZ,YV	339.32787127171	6	0	451.301090310765
.	.	.	.	.
XY,XZ,XV,YV,ZV	1.8221264731824	6	0.935309105567753	113.795345512238
XZ,XV,YZ,YV	357.737682852939	7	0	467.710901891994
.	.	.	.	.
XY,XV,YV,ZV	4.7486055625191	7	0.690610667502826	114.721824601576
XY,XZ,YZ	8414.38683205193	9	0	8520.36005109099
.	.	.	.	.
XY,XV,ZV	7.71973005859697	8	0.461315516913574	115.692949097653
XZ,YZ	8432.21182106576	10	0	8536.18504010482
.	.	.	.	.
XY,ZV	17.8452709339312	9	0.0370117018907137	123.818489972987
XY	13772.3740511227	12	0	13872.3472701618
.	.	.	.	.
ZV	6556.53891014951	12	0	6656.51212918857
X,Y,Z,V	377.787971852491	11	0	479.761190891547

Conclusion:

1. Trends in AIC Across Models:

- Models with only main effects or single two-way interactions have high AICs, indicating poor fit.
- Removing specific two-way and three-way interactions significantly increases the AIC, highlighting the importance of these interaction terms for explaining the data.
- 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.

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.
- The differences in the degrees of freedom in the same order interactions is caused due to the repeated use of one variable in the interaction terms and in turn, causing the missing effect of one or more variables in the model.
- The deviance of the models has an increasing trend as the number of interactions decreased, indicating that the model fit is getting worse as the interactions are removed.

Question2:

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

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 can infer from the model comparison results that higher-order interactions (e.g., 3-way and 4-way) don't 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 comparing models based on AIC, p-values, and goodness-of-fit, and p-values from Likelihood Ratio Test and identifying the model with the lowest AIC that balances fit and complexity.

LRT Output:

We have chosen the Likelihood Ratio Test due to the nested nature of all models, where each model is a subset of the next model. This test allows us to compare the fit of nested models and determine if the additional parameters in the saturated model significantly improve the fit.

To compare with the saturated model, we have chosen the model with the least AIC from the model comparison results with a p-value of 0.9353 and 6 degrees of freedom. We will compare the saturated model with the chosen model using the Likelihood Ratio Test.

Table 3: AIC Comparison Results

	AIC
msat	123.9732
m1	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
```

Conclusion:

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

The residual deviance of the reduced model (m1) is slightly higher than the saturated model, which is expected as the reduced model has fewer parameters.

P-value:

p-value = 0.9353 (1 - 0.9353 = 0.0647)

The p-value from the Likelihood Ratio Test is 0.0647, which is greater than the statistical significant level 0.05. This indicates that the reduced model (m1) is not significantly different from the saturated model. This suggests that the reduced model is adequate, provides a good fit without losing significant explanatory power and the additional complexity of the saturated model is not justified.

Question3:

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

Interpretation of the model:

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 significant two way interactions that provide interpretative results about 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.
- The model captures the most critical associations between variables while maintaining simplicity and interpretability.

Model coefficients:

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 across 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.

Odds Ratios and Confidence Intervals:

To quantify the associations between variables, we calculated the odds ratios and 95% confidence intervals for each significant interaction term in the model.

Compute Odds Ratios from Coefficients:

- The odds ratio for each coefficient in a logistic regression model represents the multiplicative change in the odds of the outcome for a one-unit increase in the predictor variable, holding all other variables constant.
- Exponentiating the coefficient (e^{β}) converts the log-odds into odds ratios.
- For each coefficient k, compute the confidence interval by exponentiating the bounds of the confidence interval for β_k.

Table 4: Odds Ratios and 95% Confidence Intervals

	Odds Ratio	Lower CI	Upper CI
(Intercept)	51.5792455	39.9295713	65.5213007
x	0.7467764	0.5326970	1.0400901
y	0.1802998	0.1087523	0.2834862
z	0.4619169	0.3203070	0.6558851
v	6.1363084	4.7455095	8.0416809
x:y	0.6627767	0.5436097	0.8031437
x:z	0.8474089	0.7031862	1.0246687
x:v	0.6282533	0.4425315	0.8973840
y:v	0.6416234	0.4066629	1.0667429
z:v	27.4220351	19.2297650	39.7066300

Interpretation of Associations:

The odds of child survival based on the odds ratios and confidence intervals for each significant interaction term are as follows:

- Mother's age and smoking habits (XY): The odds of survival decrease by 33.7% for a certain smoking habit when maternal age increases.
- Mother's age and gestational age (XZ): A 15.3% decrease in odds of survival is observed with longer gestational age for older mothers.
- Mother's age and child survival (XV): Older maternal age reduces the odds of survival by 37.2%.
- Smoking habits and child survival (YV): A 35.8% reduction in the odds of survival is observed with worsening smoking habits.
- Gestational age and child survival (ZV): Child survival odds increase by a factor of 27.4 with longer gestational age.

Conclusion:

- Quantitative Associations: Odds ratios quantify the strength and direction of associations for maternal age, smoking habits, gestational age, and child survival.
- Significant Predictors: Variables with confidence intervals that exclude 1 (x:y, x:v, z:v) are statistically significant predictors.
- Practical Implications: Smoking habits and gestational age have the strongest influence on child survival, as shown by their odds ratios.