Assignment III

Log-Linear Models

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

2024-12-08

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 since the deviance is derived from likelihood ratio statistic and it asympotically follows a chi-square distribution.

R Output:

The following table contains the least and highest AIC models with their corresponding deviance, degrees of freedom, p-value, and AIC values for each combination of interactions.

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, such as ZV or XZV, 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.

Interpretation:

- Adding three-way interactions increases the number of parameters dramatically.
- For categorical variables like Mother's Age, Smoking Habits, and Gestational Age, a two-way interaction like Smoking Habits × Gestational Age may explain most of the variation in child survival, while the three-way interaction Mother's Age × Smoking Habits × Gestational Age contributes very little.
- In practice, three-way interactions between variables like Mother's Age, Smoking Habits, and Gestational Age often have small or negligible effects compared to the main effects and two-way interactions.

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.

The Null Hypothesis of a LR test (H_0) : the reduced model fits the data better than the saturated model.

Alternative Hypothesis (H_A) : the saturated model provides a better fit.

The p-value from the LR test will help us determine if the reduced model is significantly different from the saturated model where we reject the null hypothesis if the p-value is less than the significance level (0.05).

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

The p-value from the Likelihood Ratio Test is 0.9353, which is greater than the statistical significant level 0.05. This indicates that the reduced model (m1) is a better fit compared to the saturated model (msat). The reduced model captures the essential relationships between variables while maintaining simplicity and interpretability.

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.

Question 4:

Fit a logistic regression model for the probability of child survival as a function of the explanatory variables. Interpret the results.

Approach:

- Define the logistic regression Model:
  - Set Child Survival (V) as the response variable.
  - Use Mother's age (X), Smoking habits (Y), and Gestational age (Z) as explanatory variables.
  - Expanded the dataset into a binary response table based on the frequency of the child survival data to fit the logistic regression model.
- Fit the logistic Regression Model:
  - Use the glm() function with the family argument set to binomial(link = "logit") to fit the logistic regression model.
- Interpret the Results:
  - Examine the coefficients to determine the direction and strength of the relationships.

Interpretation of Results:

```
##
## Call:
## glm(formula = v ~ x + y + z + x:y + x:z + y:z, family = binomial(link = "logit"),
##     data = binary_data)
##
## Coefficients:
```

```
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   1.8515     0.1517  12.207   <2e-16 ***
## x            -0.5893     0.2289  -2.575    0.010 *
## y            -0.4228     0.3726  -1.135    0.256
## z             3.2479     0.2485  13.069   <2e-16 ***
## x:y           0.3397     0.5962   0.570    0.569
## x:z           0.2594     0.3888   0.667    0.505
## y:z          -0.2564     0.5344  -0.480    0.631
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1435.5  on 6850  degrees of freedom
## Residual deviance: 1083.6  on 6844  degrees of freedom
## AIC: 1097.6
##
## Number of Fisher Scoring iterations: 7
```

Table 5: Logistic Regression Results: Odds Ratios, Confidence Intervals, and P-Values

	Variable	Odds_Ratio	Lower_CI	Upper_CI	P_Value
(Intercept)	(Intercept)	6.3695314	4.7790724	8.6717751	0.0000000
x	x	0.5547043	0.3542961	0.8707851	0.0100298
y	y	0.6551941	0.3243852	1.4128268	0.2564300
z	z	25.7357787	15.9601680	42.4556146	0.0000000
x:y	x:y	1.4045305	0.4583668	4.9260954	0.5688339
x:z	x:z	1.2961905	0.6108206	2.8243137	0.5046623
y:z	y:z	0.7738491	0.2774818	2.3098850	0.6313799

-Model Summary:

- Null deviance: 1435.5
- Residual Deviance: 1083.6
  - a significant reduction in deviance from the null model, indicating a good fit.

-Significant Predictors:

- Intercept:
  - Odds ratio = 6.37: The baseline of child survival when all predictors are zero.
  - Highly significant ($p < 0.05$).
- Mother's age (X):
  - Estimate = -0.5893: Older mothers have decreased odds of their child survival.
  - Odds Ratio = 0.55: The odds of child survival decrease by 45% (for each year increase in maternal age(=1-0.55) for older mothers.
  - Significant ($p = 0.01 < 0.05$).
- Gestational age (Z):
  - Estimate = 3.2479: Higher gestational age strongly increases the odds of child survival.

- Odds Ratio = 25.74: For a longer gestational age, the odds of child survival are 25.74 higher for a one-unit increase in gestational age.
- Highly significant (p < 0.05).

Conclusion:

The logistic regression analysis provides valuable insight into the factors influencing child survival. The significant predictors in the model are Mother's age and Gestational age, which have a fairly strong impact on child survival odds.

- Mother's age: Older mothers have lower odds of child survival, with a 45% decrease in odds for each year increase in maternal age.
- Gestational age: Longer gestational age significantly increases the odds of child survival, with a 2474% increase in odds for each additional day of gestation.
- Model fit: The model shows a significant reduction in deviance from the null model, indicating a good fit. The AIC value of 1097.6 suggests that the model is relatively simple and provides a good balance between fit and complexity.

While smoking habits were not statistically significant in this analysis, further research may be needed to explore the impact of smoking on child survival in more detail.

Question 5:

Illustrate the relationship between the logistic regression model from question 4 and a corresponding log-linear model. Confirm that the two models gives identical estimates and standard errors of the corresponding parameters.

Approach:

- **Log-Linear Model:**
  - Fit a log-linear model to analyze the relationship between the categorical variables Mother's age, Smoking habits, Gestational age, and Child survival to capture the relationships corresponding to the logistic regression model where Child survival is the response variable.
  - Use the glm() function with the family argument set to poisson(link = "log") to fit the log-linear model with interaction variables as $x * y * v, x * z * v, y * z * v$.
- **Relationship between Log-Linear and Logistic Regression Models:**
  - The log-linear model is used to analyze the relationship between categorical variables, while the logistic regression model is used to model the relationship between categorical response variables and explanatory variables.
  - Both models are generalized linear models (GLMs) that use the log link function to model the relationship between variables.
- **Comparison of Estimates and Standard Errors:**
  - We will compare the estimates and standard errors of the corresponding parameters from the log-linear and logistic regression models to confirm that they are identical.
  - We will extract the coefficients and standard errors from both models and compare them to verify the consistency of the results.

Comparison of Estimates and Standard Errors:

Table 6: Comparison of Estimates and Standard Errors between Log-Linear and Logistic Regression Models

	Variable	Log_Linear_Est	Logistic_Est	Log_Linear_SE	Logistic_SE
x:v	Mother's age	-0.5928716	-0.5893201	0.2284349	0.2288798
y:v	Smoking habits	-0.4330442	-0.4228238	0.3709920	0.3725749
v:z	Gestational age	3.2492387	3.2478822	0.2486651	0.2485129
x:y:v	Mother's age:Smoking habits	0.3881626	0.3397031	0.5573876	0.5962111
x:v:z	Mother's age:Gestational age	0.2570992	0.2594296	0.3888486	0.3888497
y:v:z	Smoking habits:Gestational age	-0.2570438	-0.2563783	0.5349728	0.5343602

Interpretation:

While the estimation process is similar, the interpretation of parameters differs:

- Logistic Regression: Coefficients represent changes in the log-odds of the child survival (binary outcome).
- Log-Linear Models: Coefficients represent changes in the log of expected counts.

Why Both Models Yield Consistent Estimates:

- Both models rely on MLE, which is consistent and asymptotically efficient under regularity conditions.
- The estimates and standard errors are derived using the Fisher information matrix, which is valid for both models.
- Shared statistical frameworks (GLMs) ensure similar estimation methodologies, leading to consistent results.

Justification:

- Both the log-linear model and logistic regression model could be used to analyze the relationship between categorical variables and child survival.
- The comparison of estimates and standard errors between the two models confirms that they yield identical results, providing consistent and reliable estimates of the relationships between variables.

Strengths of Log-Linear Models:

- Ideal for exploring associations and interactions in multidimensional contingency tables.
- Can handle higher-order interactions (e.g., three-way or four-way interactions) effectively.

Strengths of Logistic Regression:

- Better suited for predicting probabilities or understanding direct effects on a binary outcome.
- More interpretable when the focus is on one specific outcome (e.g., child survival).