Assignment III

Log-Linear Models

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

2024-12-20

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	0	0	1	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, indicating that the main effects or single interactions alone are insufficient to explain the data.
- Removing specific two-way and three-way interactions, such as ZV or XZV, significantly increases the AIC, highlighting the importance of these interaction terms of gestational age, mother's age and child survival 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 single or lower-order interactions (e.g., X,Y,Z,V) exhibit low p-values compared to the higher-order interactions (e.g., XYZ,XYV,YZV).
- 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 increasing trend of deviance and degrees of freedom (df) as the number of interactions decreases reflects the trade-off between model complexity and goodness-of-fit: simpler models fit the data less well but have more residual degrees of freedom.

Interpretation:

- 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 identifying the model with the lowest AIC that balances fit and complexity.

To compare with the saturated model, we have chosen the model with the least AIC from the model comparison results of ordered AIC with a p-value of 0.9353 and 6 degrees of freedom. We will compare the saturated model with the chosen model by their AIC values.

Table 3: AIC Comparison Results

	AIC
msat	123.9732
m1	113.7953

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.

P-value:

We reject the null hypothesis if the p-value is less than the significance level (0.05), indicating that the reduced model is not sufficient to explain the data.

p-value = 0.9353

The p-value of the reduced model is 0.9353. This indicates that the reduced model (m1) is a a very good fit. 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 two-way interactions XY, XZ, XV, YV, ZV. This interactions highlight the relationships between:

- Mother's age and smoking habits (XY): Suggests that smoking habits may vary across the age of mothers below and above 30 years.
- Mother's age and gestational age (XZ): Suggests that gestational age may be influenced by maternal age.
- Mother's age and child survival (XV): Suggests that child survival may depend if the maternal age is below or above 30 years.
- Smoking habits and child survival (YV): Suggests there could be a direct impact of smoking habits on child survival rates.
- Gestational age and child survival (ZV): Suggests that gestational age could be a factor in determining child survival rates.

The significance of these interactions is further evaluated by calculating by the p-values.

Significance of Interactions:

The p-values for the interactions in the reduced model (m1) are as follows:

Table 4: P-values for Interactions in the Reduced Model

	P-value
x:y	0.0000356
x:z	0.0845588
x:v	0.0098267
y:v	0.0697742
z:v	0.0000000

- The p-values for the interactions XY, XV, ZV are all less than 0.05, indicating that these interactions are statistically significant at the 5% significance level.
- This indicates that the interactions between Mother's age and smoking habits, Mother's age and child survival, and Gestational age and child survival are significant associations.
- The interactions XZ and YV have p-values greater than 0.05, suggesting that these interactions are statistically insignificant at the 5% significance level.
- This indicates that the interactions between Mother's age and gestational age, and Smoking habits and child survival are not significant associations.

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 coefficients of interaction terms in a log-linear model directly correspond to the logarithm of conditional odds ratio when a baseline level is defined for the variables involved in the interaction, assuming all other variables in the model are held constant.
- Exponentiation of the coefficient (e^{β}) converts the log-odds into odds ratios, while keeping all other variables constant.
- For each coefficient k, compute the confidence interval by exponentiating the bounds of the confidence interval for β_k.

Table 5: Odds Ratios and 95% Confidence Intervals

	Odds Ratio	Lower CI	Upper CI
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 ratios and confidence intervals for each significant interaction term are as follows, while keeping all other variables constant:

- Mother's age and smoking habits (XY): Older maternal age reduces the odds of smoking habits by 33.7%.

- Mother's age and gestational age (XZ): Older maternal age reduces the odds of a longer gestational age by 15.3%.
- 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: Mother's age 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.
  - Fit the full model with all main effects and interactions, then remove insignificant interactions to simplify the model.
- Interpret the Results:
  - Examine the coefficients to determine the direction and strength of the relationships.
  - The coefficients of this logistic regression model represent the change in the log-odds of the response variable when the explanatory variables changes from 0 to 1, holding other variables constant.
  - The odds ratio for a explanatory variable X_i with respect to the response variable is e^{β_i} where β_i is the coefficient of X_i in the model.
  - Interpretation:
    - If $\beta_i > 0$: The odds of the outcome are higher when $X_i = 1$ compared to $X_i = 0$.
    - If $\beta_i < 0$: The odds of the outcome are lower when $X_i = 1$ compared to $X_i = 0$.
    - If $\beta_i = 0$: X_i has no effect on the odds of the outcome.

Model Fitting and Interpretation:

```
## Full Logistic Model Summary:
##
## 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
##
## Logistic Model after removing x:y interaction:
##
## Call:
## glm(formula = v ~ x + y + z + x:z + y:z, family = binomial(link = "logit"),
##     data = binary_data)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   1.8369     0.1486  12.357   <2e-16 ***
## x            -0.5551     0.2208  -2.514   0.0119 *
## y            -0.3235     0.3342  -0.968   0.3330
## z             3.2475     0.2463  13.182   <2e-16 ***
## x:z           0.2682     0.3873   0.692   0.4887
## y:z          -0.2707     0.5324  -0.508   0.6112
## ---
## 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: 1084.0  on 6845  degrees of freedom
## AIC: 1096
##
## Number of Fisher Scoring iterations: 7
```

```
##
## Logistic Model after removing x:z interaction:
##
## Call:
## glm(formula = v ~ x + y + z + y:z, family = binomial(link = "logit"),
##     data = binary_data)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   1.7997     0.1369  13.145  < 2e-16 ***
## x            -0.4661     0.1803  -2.585  0.00973 **
## y            -0.3120     0.3332  -0.936  0.34903
## z             3.3503     0.1994  16.801  < 2e-16 ***
## y:z          -0.2977     0.5307  -0.561  0.57482
## ---
## 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: 1084.4  on 6846  degrees of freedom
## AIC: 1094.4
##
## Number of Fisher Scoring iterations: 7
##
## Final Logistic Model after removing y:z interaction:
##
## Call:
## glm(formula = v ~ x + y + z, family = binomial(link = "logit"),
##     data = binary_data)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   1.8139     0.1351  13.430  < 2e-16 ***
## x            -0.4675     0.1803  -2.592  0.00954 **
## y            -0.4228     0.2624  -1.611  0.10710
## z             3.3098     0.1846  17.929  < 2e-16 ***
## ---
## 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: 1084.8  on 6847  degrees of freedom
## AIC: 1092.8
##
## Number of Fisher Scoring iterations: 7
```

Table 6: Logistic Regression Results: Odds Ratios, Confidence Intervals, and P-Values

	Variable	Odds_Ratio	Lower_CI	Upper_CI	P_Value
(Intercept)	(Intercept)	6.1343230	4.7376430	8.0487645	0.0000000
x	x	0.6265871	0.4410444	0.8955055	0.0095399
y	y	0.6552208	0.3997588	1.1236121	0.1071032
z	z	27.3783553	19.1955622	39.6504282	0.0000000

- We have chosen the final logistic regression model after removing the insignificant interactions based on the p-values and lowest AIC.

-Model Summary:

- Null deviance: 1435.5
- Residual Deviance: 1084.8
  - a significant reduction in deviance from the null model, indicating a good fit.

-Significant Predictors:

- Mother's age (X):
  - Odds ratio: 0.63
  - The odds of child survival decrease by 37%. (1-0.63) when maternal age is more than 30 years.
  - P-value: 0.009, indicating a statistically significant effect at the 5% significance level.
- Gestational age (Z):
  - Odds ratio: 27.38
  - Longer gestational age significantly increases the odds of child survival by 27 times.
  - P-value: < 0.0001, indicating a highly significant effect at the 5% significance level.

-Non-Significant Predictors:

- Smoking habits (Y):
  - Odds ratio: 0.655
  - P-value: 0.107, indicating a non-significant effect at the 5% significance level.
  - The odds of child survival is suggested to decrease by 34.5% when smoking habits are more than 5 cigarettes per day, but cannot be concluded for sure, since the p-value suggests that the odds ratio could have occurred due to random chance.

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.

- Model fit: The model shows a significant reduction in deviance from the null model, indicating a good fit. The AIC value of 1092.8 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.
  - The log-linear model includes these additional terms to correspond to the logistic regression model - $v \sim x + y + z$: all the interactions between the explanatory variables and each of the interaction between the response variable and the explanatory variables : $x*y, y*z, x*z, x*y*z, x*v, y*v, z*v$.
- **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.

Table 7: Comparison of Estimates and Standard Errors between Log-Linear and Logistic Regression Models

Variable	Log_Linear_Est	Logistic_Est	Log_Linear_SE	Logistic_SE
Mother's age	-0.4674675	-0.4674675	0.1803510	0.1803449
Smoking habits	-0.4227830	-0.4227830	0.2623864	0.2623779
Gestational age	3.3097528	3.3097527	0.1846168	0.1846043

Table 8: Differences in Estimates and Standard Errors between Log-Linear and Logistic Regression Models

Variable	Est_Difference	SE_Difference
Mother's age	1.54884e-09	6.09227e-06
Smoking habits	1.41250e-09	8.46700e-06
Gestational age	2.16363e-08	1.25600e-05

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) with respect to the change in explanatory variables.

- Log-Linear Models: Here, coefficients are derived as logarithm of conditional odds of child survival given the values of the explanatory variables with a defined baseline level, while keeping other variables constant.

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).
