loglinear

Dong Luo

22 October 2018

Data Cleaning

We exclude the obervations with age less than 18 since our research questions only focus on adult.

We cut each of the continuous variables `CHOPER1`, `FATPER1`, and `PROPER1`, which stannd for the percentage of energy comes from carbohydrate, fat,and protein, into three distinct levels, from low, medium, to high.

	Carbohydrate	Fat	Protein
low(%)	[0,45]	[0,20]	[0,15]
medium(%)	(45,65]	(20,35]	(15,25]
high(%)	(65,100]	(35,100]	(25,100]

By dividing we are interested in the mean proportion of each diet types, so as to identify what are the most popular diet types.

Table 2: Proportion of the top 6 diet types

proportion	fat	carb	protein
0.1902491	medium	medium	medium
0.1785904	high	low	medium
0.1383148	medium	low	medium
0.1267621	medium	medium	low
0.0714361	medium	low	high
0.0557499	high	low	low

We may be interested to fit log-linear models to analyse this three-way contingency tables to see if there is any independence underlying.
We start with the addtivie table, which stands for the complete independence.

We found that the deviance for the additive model (including all three factors) is 4366.3600542. We compare it with the models with one two way interaction. There are three such models, and their residual deviance is shown as below.

Carb.Fat	Carb.Protein	Fat.Protein
1858.377	3216.99	4289.725

The model with `carb:fat` interaction has the lowest deviance. The difference of deviance between this model and the additive model is 2507.9829844. The two models are nested and when we compare them, the H_0 is the additive model (the smaller model).

$$\text{M1}\,(H_0): \hat{\mu} = \beta_0 + \alpha_i + \beta_j + \gamma_k,$$

where α_i, β_j, and γ_k denote the i^{th}, j^{th} and k^{th} group of carbohydrate, fat, and protein levels

$$\text{M2}\,(H_A) : \hat{\mu} = \beta_0 + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij}$$

Under H_0, the difference in deviance follows a χ^2 distribution whose degrees of freedom equals the difference in residual degrees of freedom of the two models.
The difference in degrees of freedom is 4, since each factor has 3 levels, so adding an interaction term would increase the number of parameters by $(3-1) \times (3-1) = 4$ in the model. The p-value for the test is close to 0, so we would reject the null hypothesis and prefer the model with `carb:fat` interaction. This model with one interaction term stands for the block independence.