Study of the Relationship between a Particular Type of Heart Disease, Cholesterol Level and Blood Pressure

Introduction

The data given were obtained in study of the relationship between a particular type of heart disease, cholesterol level and blood pressure, where the total number of individuals is fixed in the study. The individuals are cross classified according to three variables: the serum cholesterol (mg/100 cc) at 4 levels, systolic blood pressure (mm Hg) at 4 levels and the heart disease status (Present, Absent). I was interested in examining the nature of the association among these 3 variables.

Since individuals are cross classified with respect to three factors serum cholesterol, blood pressure and heart disease status with 4, 4 and 2 levels, and I was interested in the association among all three variables, I will then consider methods based on log-linear models for 3-way table with the number of individuals with V = i, W = j and X = k, Y_{ijk}, follows

$$Y_{ijk} \sim \text{Poisson}(\mu_{ijk}) \quad (1)$$

I will begin modeling with a saturated log-linear model for 3-way table of form

$$\log \mu_{ijk} = \gamma + \gamma_i^V + \gamma_j^W + \gamma_k^X + \gamma_{ij}^{VW} + \gamma_{ik}^{VX} + \gamma_{jk}^{WX} + \gamma_{ijk}^{VWX} \quad (2)$$

where V denote cholesterol, W denote blood pressure, and Z denote heart disease, with corner point constraints. We know that the saturated model provides prefect fit to data, and we will be looking for a simpler model to describe the relationship among V, W and X.

Further, we will then assume the heart disease factor (X) as the response variable, cholesterol level factor (V) and blood pressure factor (W) as the covariates. Under this assumption, we then build a logistic regression model as form

$$\log \frac{\pi_n}{1-\pi_n} = \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \beta_3 x_{n3} + \beta_4 x_{n4} + \beta_5 x_{n5} + \beta_6 x_{n6} \quad (3)$$

where X_{n1} =1 if cholesterol level of group n is between 200 - 219 (mg/100 cc) and 0 otherwise, X_{n2} =1 if cholesterol level of group n is between 220 - 259 (mg/100 cc) and 0 otherwise, X_{n3} = 1 if cholesterol level of group n is $\geq$ 260 (mg/100 cc) and 0 otherwise, X_{n4} = 1 if blood pressure of group n is between 127 – 146 and 0 otherwise, X_{n5} = 1 if blood pressure of group n is between 147 – 166 and 0 otherwise and X_{n6} = 1, blood pressure of group n is 167+ and 0 otherwise. Then given the number of individuals had heart disease in group n Y_n follows

$$Y_n \sim \text{Binomial}(m_n, \pi_n) \quad (4)$$

and this model allows us to explore the association between the response variable and the two explanatory variables.

Model selection and performance

To seek a suitable model describes the relationship between the serum cholesterol, blood pressure and heart disease status, my approach was to use model comparison by analysis of deviance. This is, we seek to a ΔD that leads to the inability to reject the null hypothesis with asymptotic distribution $\Delta D \sim \chi^2_{v2-v1}$, where v_1, v_2 are the degrees of freedom (d.f.) of two comparison models. By deviance statistic

$$D = 2 \sum_i \sum_j \sum_k y_{ijk} log \frac{y_{ijk}}{\widehat{\mu_{ijk}}} \quad (5)$$

with asymptotic distribution $D \sim \chi^2_{IJK-q}$, where q is the number of parameters under H_0, we can obtain the residual deviance for each log-linear model. And, we followed the rule that always build hierarchical log-linear models for contingency tables. I performed model selection procedure as follow:

1. Start with the saturated log-linear model as the best model.
2. Construct a new model by dropping a higher-order interaction and set a null hypothesis H_0 (e.g. Model 1 is as good as Model 0).
3. Compute deviance for each model from step 2 and obtain ΔD. Then, find its p-value.
4. If null hypothesis H_0 reject, then the new model is different in fitting the data than the model with more terms. STOP.
5. If null hypothesis H_0 cannot reject, use the new model instead, and then repeat steps 2-4.

In the results that follow, the observed deviance of Model 1 is 8.0762, and the p-value is P(D>d) = P (D > 8.0762) = 0.526 (Table 2.1) so we concluded that Model 1 is as good as Model 0 (saturated). We saw that each of Models 2, 3, 4 is significantly different (worse) in fitting the data than Model 1 so that we cannot drop any of the 2nd order interaction terms from Model 1 (Table 2.1). Therefore, it suggested that the final model should have cholesterol level, blood pressure, heart disease status and their 2nd order interaction terms of the form

$$\text{Model 1: } \log \mu_{ijk} = \gamma + \gamma_i^V + \gamma_j^W + \gamma_k^X + \gamma_{ij}^{VW} + \gamma_{ik}^{VX} + \gamma_{jk}^{WX} \quad (6)$$

Further, we observed cell counts and the fitted values (expected cell counts) agree reasonably well (Table 2.2). The residuals behave well as all residuals within [−2,2] and are roughly symmetric about 0, with no obvious patterns when plotted against the fitted value and the log fitted value (Figure 2.a). Residual analysis confirms that Model 1 is appropriate for the data.

Then, think of heart disease factor (X) at the response variable. We built a logistic regression model with cholesterol level and blood pressure of the form

$$\text{Model 5: } \log \frac{\pi_i}{1-\pi_i} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \beta_6 x_{i6} \quad (7)$$

The observed deviance statistic (“residual deviance” in R output) for Model 5 is d = 8.0762 and p-value = P (D>d) = P (D>8.0762) = 0.526, which indicates the model fits the data well. We saw that each of Models 6, 7 is significantly different (worse) in fitting the data than Model 5 so that we cannot drop any of explanatory variables. The residuals are well-behaved; all residuals within [−2,2] and are roughly symmetric about 0, with no obvious patterns when plotted against the fitted value. If we model the cholesterol level and blood pressure interaction terms, this adds 9 terms. And total number of parameters become 16, which gives a saturated model that we are not interested.

Table 2.1: Analysis of deviance table for models for heart disease study

Model	Form	Deviance	D.F.	p-value
0	(VWX)	0	0	NA
1	(VW, VX, WX)	8.0762	9	0.526 (vs Model 0)
2	(VW, VX)	26.805	12	0.000311 (vs Model 1)
3	(VW, WX)	35.163	12	5.645e-06 (vs Model 1)
4	(VX, WX)	27.666	18	0.0206 (vs Model 1)
5	Cholesterol + Blood pressure	8.0762	9	0.526 (vs saturated)
6	Cholesterol	26.805	12	0.000311 (vs Model 5)
7	Blood pressure	35.163	12	5.646e-06 (vs Model 5)

Table 2.2: Cell counts and the fitted values

y	Fitted values
117	115.450170
121	120.446940
47	47.512189
22	23.590702
85	85.855829
98	97.660083
43	41.245841
20	21.238247
119	120.499154
209	209.173541
68	67.773485
43	41.553820
67	66.194847
99	99.719436
46	47.468485
33	31.617231

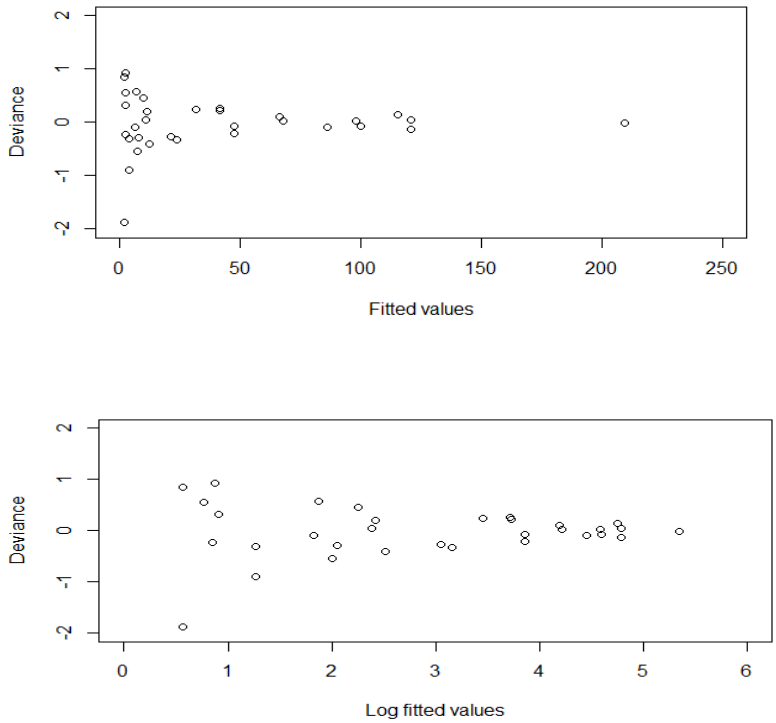

Figure 2.a: Deviance residuals vs. fitted value, and vs. log fitted value for Model 1

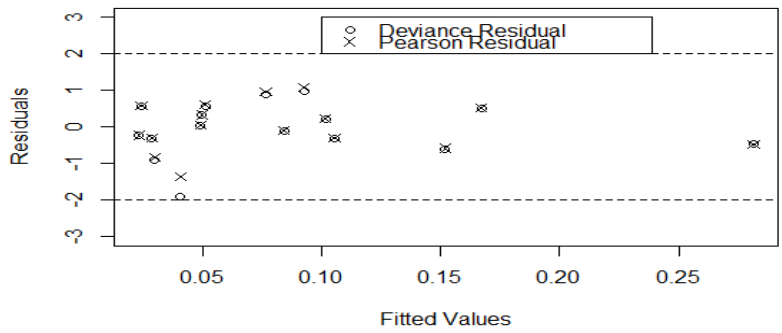

Figure 2.b: Deviance residuals and Pearson residuals vs. fitted value for Model 5

Parameter estimate and interpretation

We chose the best fit models form log-linear model for 3-way table and logistic regression model respectively and fit to the data. Using the log-likelihood function for the logistic regression model

$$l(\beta) = \sum_{i=1}^{n}[y_i x_i^T \beta - m_i \log(1+e^{x_i^T \beta})] \quad (8)$$

provided the estimate for a fit of Model 5.

In lay term of Model 1, γ_{ik}^{VX} and γ_{jk}^{WX} respectively describe the relationship between cholesterol and heart disease, blood pressure and heart disease. By comparing different γ_{ik}^{VX}, it implies that individuals with higher cholesterol level have a higher probability of heart disease than those with cholesterol level at baseline. In the much same

way, individuals with higher blood pressure have a higher probability of heart disease than those with blood pressure at baseline.

Fit of Model 5 (Table 2.4) suggests that at a given blood pressure, individuals with cholesterol level ≥ 260 have a different (higher) probability of heart disease than cholesterol level < 200. In a given cholesterol level group, individuals with blood pressure is between 147 – 166, or 167+ have a different (higher) probability of heart disease than blood pressure < 127. To describe the effects of these two explanatory variables, I calculated odds ratios of their estimates and 95% CIs in Table 2.5, and these values can be calculated from $\hat{\beta}$ obtained from equation (8). Individuals with two higher cholesterol groups and two higher blood pressure groups are more likely have heart disease than those with baseline categories (Table 2.5).

From Model 1 and Model 5, both γ_{22}^{VX} and β_1 represent the log odds ratio of the occurrence of heart disease for those with cholesterol level 200 - 219 vs those with cholesterol level < 200, among those with blood pressure < 127. This indicates that the parameters have the same representation should have same estimate value. And, we observed estimate of γ_{22}^{VX} and β_1, γ_{32}^{VX} and β_2, γ_{42}^{VX} and β_3, γ_{22}^{WX} and β_4, γ_{32}^{WX} and β_5, and γ_{42}^{WX} and β_6 are agreed with each other respectively (Table 2.3 and Table 2.4).

Table 2.3: Fit of Model 1 for heart disease data

	Estimate	SE (estimate)	P-value
(Intercept)	4.74884	0.09226	< 2e-16
cholesterolf2	-0.29617	0.14115	0.035879
cholesterolf3	0.0428	0.12835	0.738759
cholesterolf4	-0.55624	0.15033	0.000216
pressuref2	0.04237	0.12864	0.741866
pressuref3	-0.88785	0.1692	1.54E-07
pressuref4	-1.58799	0.21841	3.58E-13
heartdiseasef2	-3.48194	0.34865	< 2e-16
cholesterolf2:pressuref2	0.08645	0.19451	0.656707
cholesterolf3:pressuref2	0.50915	0.17008	0.002758
cholesterolf4:pressuref2	0.36739	0.19876	0.064541
cholesterolf2:pressuref3	0.15473	0.2512	0.537903
cholesterolf3:pressuref3	0.31238	0.2235	0.162203
cholesterolf4:pressuref3	0.55532	0.24624	0.024125
cholesterolf2:pressuref4	0.19112	0.32004	0.550388
cholesterolf3:pressuref4	0.52333	0.27572	0.057688
cholesterolf4:pressuref4	0.84909	0.29356	0.003824
cholesterolf2:heartdiseasef2	-0.20798	0.46642	0.655668
cholesterolf3:heartdiseasef2	0.56223	0.3508	0.108998
cholesterolf4:heartdiseasef2	1.34412	0.34297	8.89E-05
pressuref2:heartdiseasef2	-0.04146	0.30365	0.891394
pressuref3:heartdiseasef2	0.53236	0.3324	0.109252
pressuref4:heartdiseasef2	1.20042	0.32689	0.00024

Table 2.4: Fit of Model 5 for heart disease data

Covariate	Category	Estimate	SE (estimate)	P-value
	(Intercept)	-3.48194	0.34865	< 2e-16
Cholesterol level	200 - 219 (mg/100 cc)	-0.20798	0.46641	0.65566
	220 - 259 (mg/ 100 cc)	0.56223	0.35080	0.10900
	≥ 260 (mg/100 cc)	1.34412	0.34297	8.89e-05
blood pressure	127 - 146	-0.04146	0.30365	0.89139
	147 - 166	0.53236	0.33240	0.10925
	167+	1.20042	0.32689	0.00024

Table 2.5 Odds ratio: their estimates and 95% Cis based on Model 5

Covariate	Category	Estimate	Lower Limit	Upper Limit
Cholesterol level	200 - 219 (mg/100 cc)	0.81222539	0.32558629	2.02622195
	220 - 259 (mg/100 cc)	1.75457876	0.88221261	3.48957450
	≥ 260 (mg/100 cc)	3.83481305	1.95799219	7.51064852
Blood pressure	127 - 146	0.95938689	0.52908532	1.73964988
	147 - 166	1.70293991	0.88769321	3.26689931
	167+	3.32151870	1.75020631	6.30353486

Conclusion

Based on the above results, Model 1, denoted by (VW, VX, WX), provides the best fit to the data. This indicates that all pairs of variables are conditionally independent. That is, all variables are associated in a pairwise fashion, but the degree of association between two of them does not depend on the level of the third variable. Model 5 explains log odds of heart disease by covariates, cholesterol (V) and blood pressure (W). Both cholesterol and blood pressure play highly significant roles in the occurrence of heart disease. Individuals with high cholesterol and high blood pressure are more like to have heart disease.

According to data analysis, the logistic model (Model 5) corresponds to the final log-linear model (Model 1). The parameters represent the same log odds ratio have the same estimate values. Additionally, the residual deviance statistic and the corresponding degrees of freedom are the same for these two models, Model 1 and Model 5.