Chapter 4

Statistical Modeling

In the preceding chapter we focused on crude and adjusted odds ratios for assessing the association between HIV status and the determinants of interest. Logistic regression provides an alternative approach to the Mantel-Haenszel methods for analysing contingency tables used in the preceding chapter. This method, in common with linear multiple regression analysis, enables a model to be built, which simultaneously takes into account all the determinant variables.

Model Building

In Chapter 3 we found that age, marital status, occupation, haemoptysis, fever, weakness, weight loss, and chest x-ray findings were the statistically significant determinants of HIV status. Some of these factors were confounded by age, and other factors, including religion, duration of cough symptoms, receiving BCG, and drinking, were statistically significant at 10% but not at the conventional 5% criterion.

The characteristics of disease are the main variables of interest in this study. From the preliminary data analysis in the preceding chapter as shown in Figure 3.1, we can see that, the sample size of variables are different due to some missing values. Thus, we should reduce these missing data as categorical variables in order to reduce bias before fitting the logistic regression model. The missing cases of symptoms of disease all have symptoms of unknown duration so they are grouped into two groups as "present" or "absent". For some categories of age groups, "40-49 years", "50-59 years" and "> 60 years" were combined with "> 40 years" category.

After coding the missing values, the first step involves selecting variables in each subset, namely demographic factors and intervening variables, and disease characteristics. Finding the variables in each subset that is statistical significant, using the logistic regression method.

The full model of demographic factors and intervening variables are presented in Figure 4.1.

Figure 4.1 Full model of demographic factors and intervening variables

TB patients with and without HIV infection

factor	coeff	St.Error	p-value	Odds ratio	95% CI	
HIV						
positive / negative	-3.03	0.4224	0	0.0483	0.0211	0.1106
age			0			
>= 40 years	(0)					
< 30 years	1.028	0.3259	0.0016	2.7955	1.4758	5.2951
30-39 years	1.5873	0.2411	0	4.8906	3.0485	7.8457
marital status			0.1958			
married/mis	(0)					
single	0.5229	0.3148	0.0967	1.6869	0.9103	3.1262
others	0.3525	0.3894	0.3653	1.4227	0.6632	3.052
occupation			0.0106			
agric	(0)					
others	0.7018	0.2497	0.0049	2.0173	1.2366	3.2908
unknown	1.8076	1.2463	0.147	6.0956	0.5299	70.1215
religion			0.1964			
bud	(0)					
islam	0.6675	0.4058	0.1	1.9493	0.8799	4.3184
unknown	-0.3327	1.26	0.7918	0.717	0.0607	8.4736
receiving BCG						
no/unknown	(0)					
yes	-0.3426	0.3386	0.3116	0.7099	0.3656	1.3785
drinking						
no/unknown	(0)					
yes	0.4855	0.2381	0.0414	1.6251	1.0191	2.5914
oth.diseases			0.0641			
no	(0)					
yes	-0.5689	0.3833	0.1377	0.5662	0.2671	1.2
unknown	-1.5205	0.6496	0.0192	0.2186	0.0612	0.7809
family hx.			0.0068			
absent	(0)					
present	-0.5108	0.3307	0.1224	0.6	0.3138	1.1471
unknown	0.9613	0.3741	0.0102	2.6151	1.2562	5.444

df: 1065 deviance: 686.249 number of iterations: 4

Using the backward methods reduces variables that have a large p-value in turn, odds ratios and p-values were found to slightly change. However, there is little confounding. When gender, smoking, receiving BCG, marital status, religion, other diseases, and drinking are eliminated, then we obtain the significant variables as shown in Figure 4.2.

Figure 4.2 Significant demographic factors and intervening variables

TB patients with and without HIV infection

factor	coeff	St.Error	p-value	Odds ratio	95% CI	
HIV						
positive / negative	-3.2673	0.2382	0	0.0381	0.0239	0.0608
age			0			
>= 40 years	(0)					
< 30 years	1.0565	0.2658	0.0001	2.8764	1.7084	4.8428
30-39 years	1.5385	0.2283	0	4.6575	2.9774	7.2856
occupation			0.0015			
agric	(0)					
others	0.7863	0.2394	0.001	2.1952	1.373	3.5097
unknown	0.883	0.3012	0.0034	2.4182	1.3401	4.3639
family hx.			0.0423			
absent	(0)					
present	-0.4328	0.3246	0.1824	0.6487	0.3434	1.2256
unknown	0.6842	0.3487	0.0497	1.9822	1.0007	3.9263

df: 1073 deviance: 701.994 number of iterations: 4

Figure 4.2 shows that the demographic variables, age and occupation are most strongly associated with HIV status. It is interesting to note that family history with TB is associated with HIV infection in the unknown group. This means that the officer tended to not ask about family history among TB patients with HIV infection.

The full model of disease characteristics is shown in Figure 4.3. Using the same method to reduce redundant variables as for the demographic factors, we eliminate the insignificant variables including degree of sputum, dyspnea, fever and cough respectively.

Figure 4.3 Full model of characteristics of disease

TB patients with and without HIV infection

symptoms	coeff	St.Error	p-value	Odds ratio	95% CI	
HIV positive / negative	-0.5938	0.8904	0.5049	0.5522	0.0964	3.1628
cough absent present	(0) -1.5626	0.8733	0.0736	0.2096	0.0378	1.1607
haemoptysis absent present	(0) -0.976	0.3031	0.0013	0.3768	0.208	0.6826
chest pain absent present	(0) 0.5011	0.2201	0.0228	1.6505	1.0722	2.5409
dyspnea absent present	(0) 0.257	0.251	0.3058	1.2931	0.7906	2.1149
fever absent present	(0) 0.2663	0.2127	0.2105	1.3051	0.8602	1.98
weakness absent present	(0) 0.9676	0.2761	0.0005	2.6315	1.5317	4.5211
weight loss absent present	(0) 0.7463	0.2385	0.0018	2.1092	1.3216	3.3663
other symptoms absent present	(0) -0.9545	0.2741	0.0005	0.385	0.225	0.6588
chest x-ray others cavity	(0) -0.7145	0.2053	0.0005	0.4894	0.3273	0.7319
degree of sputum negative positive	(0) -0.0769	0.3115	0.805	0.926	0.5028	1.7053

df: 1067 deviance: 713.548 number of iterations: 4

The odds ratios and p-values change only slightly, so there is no confounding. Finally, we obtain the significant variables of characteristics as shown in Figure 4.4.

Figure 4.4 Significant characteristics of disease

TB patients with and without HIV infection

symptoms	coeff	St.Error	p-value	Odds ratio	95% CI	
HIV positive / negative	-1.9006	0.2253	0	0.1495	0.0961	0.2325
haemoptysis absent present	(0) -0.9777	0.2989	0.0011	0.3762	0.2094	0.6758
chest pain absent present	(0) 0.5544	0.2169	0.0106	1.7409	1.1381	2.663
weakness absent present	(0) 0.9699	0.2732	0.0004	2.6376	1.544	4.5058
weight loss absent present	(0) 0.7336	0.2362	0.0019	2.0826	1.3107	3.3089
other symptoms absent present	(0) -0.897	0.269	0.0009	0.4078	0.2407	0.6909
chest x-ray others cavity	(0) -0.7142	0.1979	0.0003	0.4896	0.3322	0.7215

df: 1071 deviance: 718.536 number of iterations: 4

Figure 4.4 shows weakness, other symptoms and chest x-ray findings are most strongly associated with HIV infection. Haempotysis, other symptoms and chest x-ray are negatively associated with HIV infection. In contrast, chest pain, weakness and weight loss are positively associated with HIV infection.

Fitting the Model

At the next step, based on these findings, and also on subject matter knowledge, we decided to include the following determinants in the initial logistic model.

age	occupation	family history with TB
haemoptysis	chest pain	weakness
weight loss	other symptoms	chest x-ray findings
smoking	drinking	other diseases

Including all variables of interest, the full model of logistic regression is shown in Figure 4.5.

Figure 4.5 Full logistic model with selected variables of interest

TB patients with and without HIV infection

factor	coeff	St.Error	p-value	Odds ratio	95% CI	
HIV positive / negative	-3.078	0.3283	0	0.0461	0.0242	0.0876
age			0			
>=40 years	(0)					
<30 years	1.1425	0.2759	0	3.1347	1.8253	5.3833
30-39 years	1.5517	0.2407	0	4.7195	2.9442	7.5653
occupation			0.0032			
agriculture	(0)					
others	0.7985	0.2474	0.0012	2.2222	1.3683	3.609
unknown	0.7922	0.3169	0.0124	2.2083	1.1865	4.1099
family hx.			0.0635			
absent	(0)					
present	-0.4463	0.3352	0.183	0.64	0.3318	1.2345
unknown	0.6365	0.362	0.0787	1.8898	0.9295	3.8425
haemoptysis						
absent	(0)					
present	-1.0816	0.3098	0.0005	0.3391	0.1848	0.6223
chest pain						
absent	(0)					
present	0.4563	0.2287	0.046	1.5782	1.0082	2.4706
weakness						
absent	(0)					
present	0.9834	0.2803	0.0004	2.6736	1.5436	4.6308
weight loss						
absent	(0)					
present	0.6215	0.2492	0.0126	1.8617	1.1424	3.0341
other symptoms						
absent	(0)					
present	-0.7328	0.2732	0.0073	0.4805	0.2813	0.8209
chest x-ray						
others	(0)					
cavity	-0.7443	0.2081	0.0003	0.4751	0.316	0.7142

df: 1065 deviance: 653.598 number of iterations: 4

Figure 4.5 shows that age, haemoptysis, weakness and chest x-ray are most strongly associated with HIV infection. However, occupation, chest pain, weight loss and other symptoms are strongly associated with HIV status too, except family history with TB.

Using the backward method to eliminate redundant variables (family history with TB), the model with all significant variables is shown in Figure 4.6.

Figure 4.6 The model with all significant variables

TB patients with and without HIV infection

factor	coeff	St.Error	p-value	Odds ratio	95% CI	
HIV positive / negative	-2.999	0.3142	0	0.0498	0.0269	0.0923
age			0			
>=40 years	(0)					
<30 years	1.1271	0.2701	0	3.0866	1.818	5.2404
30-39 years	1.5041	0.2386	0	4.5001	2.8194	7.1827
occupation			0.0043			
agriculture	(0)					
others	0.725	0.2399	0.0025	2.0648	1.2903	3.3043
unknown	0.8245	0.3126	0.0084	2.2807	1.2358	4.2091
haemoptysis						
absent	(0)					
present	-1.1102	0.3082	0.0003	0.3295	0.1801	0.6028
chest pain						
absent	(0)					
present	0.4183	0.2262	0.0644	1.5193	0.9753	2.3668
weakness						
absent	(0)					
present	1.0174	0.2822	0.0003	2.7661	1.591	4.8093
weight loss						
absent	(0)					
present	0.6277	0.2487	0.0116	1.8732	1.1505	3.0499
other symptoms						
absent	(0)					
present	-0.7816	0.2742	0.0044	0.4577	0.2674	0.7833
chest x-ray						
others	(0)					
cavity	-0.7194	0.2067	0.0005	0.487	0.3248	0.7303

df: 1067 deviance: 659.024 number of iterations: 4

After omitting family history with TB, the odds ratios and p-values change slightly. Thus family history is not a confounder in this model. Age, haemoptysis, weakness and chest x-ray findings are still most strongly associated with HIV status. While chest pain became insignificant (p-value = 0.0644), the occurrence of significance in the model as shown in Figure 4.5 may have occurred by chance. Thus it could be eliminated and reconstructed in a new model as shown in Figure 4.7.

Figure 4.7 The penultimate model

TB patients with and without HIV infection

factor	coeff	St.Error	p-value	Odds ratio	95% CI	
HIV						
positive / negative	-2.7389	0.2753	0	0.0646	0.0377	0.1109
age			0			
>=40 years	(0)					
<30 years	1.1525	0.2695	0	3.1662	1.867	5.3695
30-39 years	1.5647	0.2365	0	4.7811	3.0075	7.6007
occupation			0.0058			
agriculture	(0)					
others	0.704	0.2394	0.0033	2.0217	1.2645	3.2324
unknown	0.7954	0.3106	0.0104	2.2154	1.2051	4.0726
haemoptysis						
absent	(0)					
present	-1.1167	0.3077	0.0003	0.3273	0.1791	0.5983
weakness						
absent	(0)					
present	1.0445	0.2814	0.0002	2.8421	1.6372	4.9338
weight loss						
absent	(0)					
present	0.6028	0.2477	0.0149	1.8272	1.1245	2.9691
other symptoms						
absent	(0)					
present	-0.7699	0.2733	0.0048	0.463	0.271	0.7912
chest x-ray						
others	(0)					
cavity	-0.7119	0.2062	0.0006	0.4907	0.3276	0.7351

df: 1068 deviance: 662.566 number of iterations: 4

After eliminating chest pain, all variables are statistically significant. We can see age, haemoptysis, weakness and chest x-ray finding are most strongly associated with HIV infection. The odds ratios and p-values only change slightly, so chest pain is not a confounder in this model.

Using the model as shown in Figure 4.7, we can see that the risk of the "others" and the "unknown" group in occupation are similar. Thus, they may be grouped together for the final model as shown in Figure 4.8.

Figure 4.8 The final model

TB patients with and without HIV infection

factor	coeff	St.Error	p-value	Odds ratio	95% CI	
HIV						
positive / negative	-2.7431	0.2751	0	0.0644	0.0375	0.1104
age			0			
>=40 years	(0)					
<30 years	1.1464	0.2688	0	3.1469	1.8583	5.3293
30-39 years	1.5684	0.2363	0	4.7989	3.0202	7.6252
occupation						
agriculture	(0)					
others/unknown	0.7272	0.2278	0.0014	2.0693	1.324	3.2343
haemoptysis						
absent	(0)					
present	-1.1275	0.3059	0.0002	0.3238	0.1778	0.5898
weakness						
absent	(0)					
present	1.0471	0.2812	0.0002	2.8494	1.6421	4.9441
weight loss						
absent	(0)					
present	0.6081	0.247	0.0138	1.837	1.132	2.9811
other symptoms						
absent	(0)					
present	-0.771	0.2732	0.0048	0.4625	0.2708	0.7901
chest x-ray						
others	(0)					
cavity	-0.7059	0.2052	0.0006	0.4937	0.3302	0.7381

df: 1069 deviance: 662.669 number of iterations: 4

Figure 4.8 shows the model after collapsing the categories of occupation. The p-value for occupation changes from 0.0058 to 0.0014. Small p-values show the most strongly associated factors for HIV infection include age, occupation, haemoptysis, weakness and chest x-ray findings.

This final model may be confounded by intervening variables such as smoking, drinking and other diseases. We should confirm the association with these variables. After adjustment for smoking, drinking and other diseases, the odds ratios and p-values are found to change only slightly, so these factors are not confounders in this model.

It is clear that age occupation, weakness and weight loss are positively associated with HIV infection and haemoptysis, other symptoms and chest x-ray findings are negatively associated with HIV infection.

In the univariate analysis, adjusted odds ratios (as clearly shown by the graphical method) show age, occupation, haemoptysis, weakness, weight loss and chest x-ray findings are associated with HIV status. In the multivariate analysis, the factors associated with HIV infection are all determinants in the univariate analysis and include other symptoms. The other symptoms are not significant in the univariate analysis after adjusting for age. The association may be diluted by other factors.

The logistic regression model could be fitted to the proportions, providing an alternative approach to the Mantel-Haenszel methods for analysing contingency tables. But due to zero cell counts for some variables, this analysis runs into numerical difficulties. So the residuals and normal scores plot cannot be used in this case.

After obtaining a model that contains the significant variables, we considered their interaction among these variables. When biological knowledge was taken into account, it was found that there was no interaction between these variables.

Logistic Modeling

The final model as shown in Figure 4.8 shows that the seven factors age, occupation, haemoptysis, weakness, weight loss, other symptoms, and chest x-ray findings are associated with HIV infection. These factors are used to assess the probability of HIV infection in TB patients in terms of log odds of disease with their β-coefficients as a linear function of explanatory variables.

1. Model of Log Odds of HIV Infection

The mathematical model from logistic regression print out as shown in Figure 4.8, where Y is the log odds of HIV infection, and Xs are the explanatory variables is described as follows.

$$Y = \{-2.74 + 1.15X_1 + 1.57X_2 + 0.73X_3 - 1.13X_4 + 1.05X_5 + 0.61X_6 - 0.77X_7 - 0.71X_8\}$$

where X_1 = (age <30 yrs.), X_2 = (age 30-39 yrs.), X_3 = (not agriculture worker),
X_4 = (presence of haemoptysis), X_5 = (presence of weakness),
X_6 = (presence of weight loss), X_7 = (presence of other symptoms),
X_8 = (chest x-ray with cavity)

This equation is modeled in terms of the log odds of disease (HIV infection). The log odds of HIV infection is not widely used in practice. Thus, given that the study is cross-sectional, this equation may be inverted to give the probability of HIV infection.

2. Model of Probability of HIV Infection

The modeled estimates of these measures derived from the regression coefficients as shown in Figure 4.8, are obtained by using equation (14) in Chapter 2 to give expressions for the probability of HIV infection as shown.

$$P = \frac{1}{1+\exp[2.74-1.15X_1-1.57X_2-0.73X_3+1.13X_4-1.05X_5-0.61X_6+0.77X_7+0.71X_8]}$$

For example, suppose that patient 1 is a 35 years old male, having agriculture occupation, who came into the Zonal TB Center with symptoms of cough without haemoptysis, chest pain, weakness, weight loss, anorexia, and chest x-ray finding with no cavitary.

We thus estimate the probability of HIV infection for this subject by using the above equation as follows.

$$\Pr ob[HIV^{+}] = \frac{1}{1 + \exp[2.74 - 1.57(X_2^{(1)}) - 1.05(X_5^{(1)}) - 0.61(X_6^{(1)}) + 0.77(X_7^{(1)})]}$$

$$= 0.4304$$

The conclusion is that this patient has a probability of HIV infection of about 43%. Thus, we have obtained a probability model to evaluate the demographic factors and disease characteristics to discriminate individuals with and without HIV infection in pulmonary TB patients. This model provides useful estimation of probabilities of being with and without HIV infection for simplified interpretation. The advantage of this model is that the physicians can be used as a screening test for decision making to confirm HIV infection in laboratory.