~ CHAPTER 5 ~

THE PREDICTION OF DENGUE SEROLOGY TEST RESULTS

5.1 Introduction

This chapter presents the multiple logistic regression model developed to predict and classify the dengue serology test results for patients suspected of dengue infection at the UMMC. For this purpose, the dependent variable is the *final serology test* results with either positive or negative result in confirming the dengue infection.

The following logistic regression analysis considered only the adult patients. A total of 145 adult cases were available for the analysis. Such low number of valid cases was due to the missing data as explained earlier in the chapter of data and methodology. Children were excluded for there were only 16 observations eligible for analysis. The independent variables considered in this analysis were listed in Table 5.1. Variable age, gender and ethnicity of patients were not included since they were not suggested as the factors of dengue infection as per the literature as reviewed in Chapter 2.

In the process of constructing the logistic model, the chapter first looks into the logistic relationship via the likelihood ratio test as well as the odds ratios for each independent variable as in Section 5.2. The final estimated model is presented in Section 5.3 along with the interpretation. The probability of positive dengue serology test result and the adjusted odds ratios are also computed in this section. The fit of the model is assessed in Section 5.4 followed by the diagnostic tests in Section 5.5. In closing the chapter, Section 5.7 provides some concluding comments in relation to the logistic model.

5.2 Model Estimation – Logistic Regression

In short-listing independent variables for the logistic regression model, the logistic relationship between each independent variable and the outcome variable (the final serology test) was assessed. This was accomplished via the likelihood ratio test which examined the change of the -2 log-likelihood to determine if each of these variables could significantly reduce the log-likelihood ratio and thus, significantly explain the outcome. The findings are summarized in Table 5.1.

Referring to Table 5.1, *skin rash*, *thrombocytopenia_100* and *platelet count at admission* are significant at 1% in predicting the dengue serology test outcome. The *fever duration* and *thrombocytopenia_50* are significant at 5% followed by *rash/petechiae* and *abdominal pain* at 10%. *Fever*, *eye pain* and *heart rate per minute* are significant at 25%. All these variables with at least 25% level of significance are the potential candidates for inclusion in the ensuing stepwise procedure of the logistic model. The 0.25 level is recommended by Hosmer and Lemeshow (2000) based on the findings from a few works that show the inadequacy of the traditional level of 0.05 that often fails to pick up variables known to be essential. They argued that the significance level at the univariate stage should be large enough to allow any suspected variables to become candidates for inclusion in the model since some variables may become an important predictor when taken collectively, despite the weak association with the outcome at the univariate level.

Table 5.1: Likelihood ratio test[a] and odds ratio for the dengue serology test result of the adult patients

Variable	Likelihood Ratio Test Statistic[b c]	Odds Ratio	95% Confidence Interval
Fever duration	3.845***	1.158	(0.996, 1.347)
Fever	1.511*	0.538	(0.192, 1.510)
Vomit	0.085	1.072	(0.672, 1.710)
Giddiness	0.129	0.894	(0.485, 1.647)
Headache	0.793	0.738	(0.378, 1.441)
Skin Rash	9.844****	2.944	(1.475, 5.878)
Eye Pain	2.370*	0.507	(0.213, 1.203)
Muscle & Joint Pain	0.148	1.148	(0.568, 2.323)
Bleeding	0.080	0.923	(0.534, 1.598)
Shock Evidence	0.000	1.022	(0.992, 1.053)
Hepatomegaly	0.036	1.111	(0.370, 3.337)
Rash / Petechiae	2.960**	1.803	(0.917, 3.546)
Abdominal Pain	3.478**	1.670	(0.964, 2.893)
Dehydration	0.186	1.111	(0.687, 1.797)
Haemoconcentration_20	0.362	0.830	(0.452, 1.521)
Haemoconcentration_50	0.172	1.409	(0.269, 7.381)
Thrombocytopenia_50	4.895***	1.784	(1.058, 3.006)
Thrombocytopenia_100	9.479****	2.316	(1.357, 3.954)
Platelet count at admission	8.688****	0.993	(0.988, 0.998)
Heart rate per minute	2.482*	0.985	(0.967, 1.004)
Hematocrit change	0.030	0.998	(0.980, 1.017)

[a] For comparing the based model with a constant only to the univariate logistic model. Independent variable is *Final Serology Test.*

[b] It is the change in the -2 log-likelihood from the based model to the univariate logistic model and is distributed as χ^2 with 1 degree of freedom under the hypothesis that the coefficient for the independent variable is zero.

[c] Critical Value, $\chi^2_{(1)}$:

1.323	25%	*
2.706	10%	**
3.841	5%	***
6.635	1%	****

The odds ratio shows that adults with skin rashes are about 3 times more likely to be tested positive of dengue infection. Likewise, those with low platelet count (*thrombocytopenia_100 and thrombocytopenia_50*), *abdominal pain* or positive *rash/petechiae* test are about 2 times more likely. Nonetheless, the 95% confidence interval for the odds ratio of *abdominal pain* encompasses the value of one implying that not all patients with abdominal pain turn out positive in the serology test. *Fever, eye pain*

and *heart rate per minute*, though significant in the likelihood ratio test, have odds ratios of close to one and confidence intervals that include one, suggesting that these symptoms may not be significant predictors of the serology test outcome for adults.

In the stepwise procedure, *Thrombocytopenia_50* and *Thrombocytopenia_100* were eliminated due to collinearity with platelet count. This is expected because the former two variables are in effect the dichotomized versions of *platelet count at admission*. For the same reason, *fever duration* and *fever* are also highly related since the latter was derived from the former numerical variable (in any event, both were not retained in the final model). The same goes for *rash/petechiae* test and *skin rash* which are highly correlated as they essentially measure the same symptom and because of that, only *skin rash* was retained in the final model due to the stronger association with the outcome variable. However, it should be noted that even though collinearity among variables can make discriminatory power redundant, it does not make the variables irrelevant from a perspective of explanation.

Once a model with the main effects is obtained, possible interactions among the variables in the model are checked one at a time. In this instance, the inclusion of the interactions between the different dengue symptoms in the model produces inappropriate coefficients and renders all the main effects insignificant. Given that and due to the lack of scientific basis between the pairs of variables that are clinically plausible, no interaction term is added to the model. The final logistic model containing only the main effects is presented in the following section along with the interpretation.

5.3 Model Interpretation

Referring to Table 5.2, the estimated signs of the effect are in accordance to expectation. Skin rash and abdominal pain are positively related to dengue infection while platelet count is the opposite. The logistic model essentially hinges on three vital clinical features of dengue infection. The first being skin rash which is a rather common indicator of haemorrhagic tendency in dengue. The general abdominal pain may be related to gastrointestinal bleeding and ascites (abnormal accumulation of serous fluid in the abdominal cavity) that are frequently observed in such infection. Platelet count can indicate and detect thrombocytopenia, an abnormal decrease of platelets in the circulatory blood, which has been proven to be a consistent indicator of the severity of dengue infection (Chin, 1993).

Table 5.2: Estimated logistic regression model for the dengue serology test result of adult patients

	Variable	β	Std. Error	Wald	Sig.
X_1	Skin rash	0.9958	0.379	6.915	0.009
X_2	Abdominal Pain	0.8302	0.390	4.535	0.033
X_3	Platelet count at admission	-0.0143	0.005	7.610	0.006
	Constant	0.6313	0.453	1.939	0.164

Y = *Final Serology Test* (1: Positive, 0: Negative)
N = 145 adult cases

The Wald statistics for all the estimated coefficients are significant at 5%, except for the constant. The coefficients for skin rash and platelet count are significant even at 1%.

The adjusted odds ratios for the estimated logistic model are presented in Table 5.3. These sample odds ratios provide indirect estimates of the population relative risks.

Controlling for other effect, the odds ratio of 2.707 for skin rashes translates to about 3 times the risk of being tested positive of dengue serologically compared to those without any rashes. Those with general abdominal pain have about 2 times the odds of being confirmed serologically of such infection compared to those without. With every additional 50 units (in thousand) of platelet during admission, the odds of the disease reduce by about 0.489 times ($e^{(50 * \beta)}$, where β is -0.0143), assuming the logit is linear[3]. Nevertheless, the odds ratio and the 95% confidence interval of platelet count is very close to one, implying that a small change in this risk factor may not substantially affect the odds of the disease. The 95% confidence intervals for skin rash and abdominal pain suggest that the odds of being dengue positive could be as low as 1 time (no relative risk) or as much as 5 times compared to those without the said symptoms.

Table 5.3: Adjusted odds ratios and 95% confidence intervals for the estimated logistic model in Table 5.2

		95% Confidence Interval	
	e^{β}	Lower	Upper
X_1 Skin rash	2.707	1.289	5.686
X_2 Abdominal Pain	2.294	1.068	4.925
X_3 Platelet count at admission	0.986	0.976	0.996

[3] The plot of lowess smoothed univariable logit versus platelet count at admission supports treating the latter continuous variable as linear in the logit. This method of determining whether the model is linear in the logit for continuous variable is discussed by Hosmer and Lemeshow (2000).

The logit, which gives the log odds of dengue infection, can be fitted as:

$$\text{Logit}(\hat{P}) = \ln\left(\frac{P}{1-P}\right) = 0.6313 + 0.9958\,X_1 + 0.8302\,X_2 - 0.0143\,X_3 \qquad (1)$$

For interpretation, the log odds of serologically confirmed dengue infection would increase by 0.9958 for those with skin rash compared to those without, holding other effect constant. For those with abdominal pain, the log odds of positive dengue serology test result should see an increase of 0.8302 vis-à-vis those without such symptom, holding other variable constant. Conversely, the log odds should decrease by 0.0143 for a unit increase in the platelet count at admission (in thousand), others being constant.

The estimated multiple logistic model can be written as:

$$\hat{P}(Y=1 \mid X_1, X_2, X_3) = \left[\ 1 + e^{-(0.6313 + 0.9958\,X_1 + 0.8302\,X_2 - 0.0143\,X_3\ )}\ \right]^{-1} \qquad (2)$$

which basically provides the probability of positive dengue serology test given the values of the independent variables. The probability differs for patients with different set of conditions, which means different values of the independent variables. Using equation (2), Table 5.4 tabulates the probabilities under a few general conditions. Note that three levels of platelet count – 150, 100 and 50 thousand per mm^3 – are medically meaningful thresholds. The platelet count of 100,000 per mm^3 is recognized as the threshold for thrombocytopenia in diagnosing DHF (WHO, 1997a) while the 50,000 per mm^3 is the threshold for admission into UMMC (Chin, 1993). Anything above 100,000 per mm^3 is

considered normal. As shown in the said table, in general, as the platelet count drops, the probability of being tested positive of dengue infection increases rapidly. The probability is relatively high for those with positive skin rash, abdominal pain and thrombocytopenia of 100,000 platelet cells per mm^3 or less.

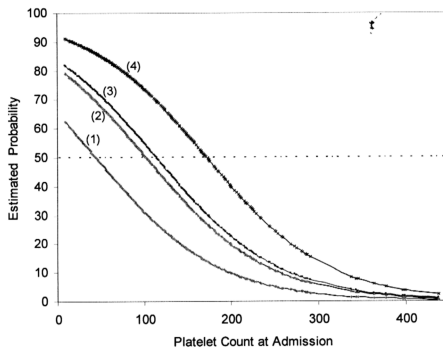
Table 5.4: Estimated probability of positive dengue serology test under different conditions based on equation (2)

		Abdominal Pain ($X_2 = 1$)			No Abdominal Pain ($X_2 = 0$)		
Platelet Count [a] (X_3)		150	100	50	150	100	50
Skin Rash	($X_1 = 1$) Yes	57.7%	73.6%	85.1%	37.3%	54.9%	71.3%
	($X_1 = 0$) No	33.5%	50.8%	67.8%	18.0%	31.0%	47.9%

[a] Measured in thousand per mm^3 at the point of admission. According to WHO (1997a), platelet count of less than 100,000/mm^3 is deemed low.

The estimated risk under the different scenarios tabulated in Table 5.4 is further illustrated in Figure 5.1 which provides the graphs of estimated probability of positive dengue serology test outcome in the presence of different risk factors in adults. The figure shows that at platelet count of 50,000 per mm^3 or less, the estimated probability is always about 50% or more for all conditions (using 0.50 as the cut-point in classifying the test results). At 100,000 platelet count per mm^3, the estimated risk is at least 50% for adults with either skin rash, abdominal pain or both. Those with both skin rash and abdominal pain will have at least 50% chance of being tested positive even if their platelet count is higher than the recommended threshold of 100,000 per mm^3, up to about 170,000 per mm^3 before the probability drops under 50%.

Figure 5.1: Estimated probability of positive dengue serology test for adult patients under different condition of platelet count at admission, skin rash and abdominal pain



Note: (1) Skin Rash = 0; Abdominal Pain = 0
(2) Skin Rash = 0; Abdominal Pain = 1
(3) Skin Rash = 1; Abdominal Pain = 0
(4) Skin Rash = 1; Abdominal Pain = 1

5.4 Model Fit Assessment

By means of the above logistic model, the classification matrix shown in Table 5.5 records an overall classification accuracy of 69% in classifying observations into positive and negative serology test results (using cut-point of 0.50). About 82.8% of the observed positive dengue cases are correctly classified (sensitivity), whereas only about 48.3% of the observed negative cases are correct (specificity).

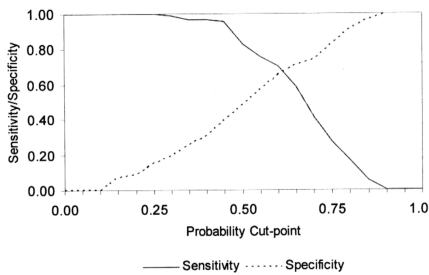
Table 5.5: Classification performance of the estimated logistic model in Table 5.2

		Observed		
		Positive	Negative	Total
Classified	Positive	72	30	102
	Negative	15	28	43
Total		87	58	145

Cut point is 0.50
Sensitivity = 82.8% (72 / 87)
Specificity = 48.3% (28 / 58)
Overall accuracy = 69.0% ((72 + 28) / 145)

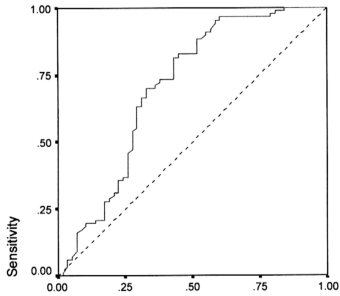
Instead of using the typical cut-point of 0.50 to classify observations as in the preceding table, it is possible to select an optimal cut-point for the purpose of classification. This optimal cut-point will maximize both sensitivity and specificity (Hosmer and Lemeshow, 2000). Figure 5.2 shows the plot of sensitivity and specificity for all possible cut-points. The optimal cut-point is where the sensitivity and specificity curves cross at approximately 0.62 (higher than 0.50). With this cut-point, the sensitivity is 64.4% while specificity is 69.0%.

Figure 5.2: Plot of sensitivity and specificity for all possible cut-points using the estimated logistic model in Table 5.2



Unlike sensitivity and specificity which rely on a single cut-point to classify a test result as positive, a better illustration of classification accuracy is provided by the area under the Receiver Operating Characteristic (ROC) curve (Hosmer and Lemeshow, 2000). The ROC curve is generated by plotting sensitivity (true signal) versus 1 minus specificity (false signal) over all possible cut-points as depicted in Figure 5.3. The area under the curve is essentially the likelihood that a subject with observed positive dengue serology outcome will have higher probability of positive test outcome than a subject who was found negative. In this case, the area under the curve is 0.704, which according to Hosmer and Lemeshow (2000) is considered acceptable discrimination.

Figure 5.3: Receiver Operating Characteristic (ROC) curve for the estimated logistic model in Table 5.2



Note: Area under the curve is 0.704

Based on the chance-based criteria of classification (Table 5.6), the highest accuracy attainable is about 60% based on the maximum chance criterion (87 positive cases divided by a total of 145 cases) and 52% by the proportional chance criterion. The hit ratio of 69% is relatively higher than both measures, implying that the classification via the logistic model is better than chance. The Press's Q statistic, which compares the number of correct classifications with the total sample size and number of groups, concludes that the predictions were significantly better than chance. Hence, the use of the logistic model has certainly improved the predictive accuracy, though by a small magnitude.

Table 5.6: Measures of classification accuracy for the estimated logistic model in Table 5.2

Measures	Value / Statistic
Maximum Chance Criterion	60.0%
Proportional Chance Criterion	52.0%
Press's Q Statistic	20.86 [a]

[a] Critical value $\chi^2_{0.01\,(1)} = 6.635$

The Hosmer and Lemeshow goodness-of-fit test, which tests the null hypothesis of no difference in the distribution of the actual and predicted outcome, yields a Chi-square of 16.52 with p-value of 0.036, indicating a borderline significance at 5% (but non-significance at 1%), which is suggestive of a mediocre model fit.

The -2 log-likelihood value of the final logistic model is reduced significantly by about 21.95 (Chi-square at 5% with 3 degrees of freedom is 7.815) from the base model of 195.173 (with constant only), indicative of an improved fit. The Cox & Snell R-square

and Nagelkerke R-square are low at 14% and 19% respectively. Such low R-squares relate to much of the unexplained variability in the dependent variable. This could be due to the study design that includes not only patients with dengue, but also those with non-dengue viral infection and other illnesses. The latter patients, though do not have dengue, may exhibit dengue-like symptoms that further complicate the clinical features of dengue infection. For that reason, independent variables that explain positive dengue infection may also be explaining the negative cases to certain extend, in so doing causing low explanation of the variation in dengue serology test results, which translates to low R-squares.

However, it must be acknowledged that these three independent variables in the final logistic model can significantly provide a differential diagnosis of dengue infection in the presence of other non-dengue illnesses in the context of UMMC.

5.5 Model Diagnostic

Influential Analysis using the studentized residuals, leverage hat values, Cook's Distance and DFBETA did not reveal any persistent influential observation for removal. None of the observations appears influential in more than two diagnostic measures as depicted in Table 5.7. Hence, no removal was made to the original data in this instance.

Table 5.7: Summary of diagnostic tests for influential observations

Measure		Threshold Value	Influential Observation*
Studentized Residuals		± 1.96	369, 621
Hat Values		0.0552	374, 608, 662, 708, 714, 715
Cook's Distance		0.0284	344, 350, 352, 360, 368, 369, 374, 379, 407, 439, 581, 608, 618, 621, 624, 636, 637, 638, 639, 644, 650, 652, 654, 657, 662, 667, 690, 694, 697, 699, 702, 706, 714, 715, 717, 718, 721, 722, 724, 725
DFBETA	Intercept	0.1661	None
	X_1		
	X_2		
	X_3		

* Original case number

5.6 Concluding Remarks

The final logistic model consists of three covariates, namely skin rash, abdominal pain and platelet count at admission. With these variables, the model is able to correctly classify about 69% of the suspected dengue cases into groups of positive and negative dengue cases based on the serology tests. The area under the ROC curve of 0.704 suggests acceptable discrimination by the model. It must be recognized that the sample in this study is made up of dengue and non-dengue patients of whom the latter may exhibit dengue-like symptoms, despite being tested negative of dengue infection. Such phenomenon possibly explains the low R-Squares of the logistic model. If the sample consists of only dengue patients and healthy subjects, the logistic model would then be expected to include all covariates known to be the vital symptoms of dengue infection and the resulted R-Square should be higher for that matter.

It is worth noting that the said three independent variables in the final model, as well as those at the univariate level (i.e. fever, eye pain, rash/petechiae, heart rate per minute and thrombocytopenia), are shown to be significant in the differential diagnosis of dengue in the presence of other non-dengue illnesses. These differential symptoms are likely to be unique to the patient mix and the dengue management at UMMC. While the above logistic model might not be used explicitly to predict the outcome of the serology test, the findings suggest that adult patients with skin rash, abdominal pain and low platelet count at admission (thrombocytopenia of 100,000 cells per mm^3 or less) should be sent for dengue serology test for they are likely to be positive of such infection.