§7.2.7 A Strategy for Logistic Regression Analyses

Here is a suggested (and somewhat simplified) strategy for a logistic regression analysis:

- Descriptive/exploratory analysis
  - PROC FREQ and PROC UNIVARIATE as usual
  - Preview of bivariate associations:
    - Cross-tabulations of categorical predictors vs outcome (and, possibly, chi-square tests)
    - Side-by-side boxplots comparing average value of a continuous predictor for the two outcome categories (and, possibly, two-sample t-tests)
- Estimate the model and make inferences:
  - Overall test (likelihood ratio test)
  - If H_0 is rejected in the overall test, produce confidence intervals for odds ratios (and, possibly, perform partial tests such as the Wald test)
- Model fit statistics
  - concordance/discordance statistics
  - ROC curves

We will cover model inferences and model fit statistics; descriptive statistics were covered in BIOS 500.

§7.2.8 Logistic Regression Inferences: Overall and Partial Tests

Consider the logistic regression model

$$\Pr(Y=1 \mid X_1 = x_1, ..., X_k = x_k) = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+...+\beta_k x_k)}},$$

or, equivalently, the logit form of the model:

$$\text{logit}[\Pr(Y=1)] = \ln\left[\frac{\Pr(Y=1)}{1-\Pr(Y=1)}\right] = \beta_0 + \beta_1 X_1 + ... + \beta_k X_k$$

The first test that is usually performed for this model is the **overall test**, which is a test of:

H_0: $\beta_1 = ... = \beta_k = 0$
H_A: $\beta_1 \neq 0, and/or \beta_2 \neq 0, ..., and/or \beta_k \neq 0$
(i.e., the alternative is that at least one slope parameter is significantly different from 0 or that at least one predictor is significantly associated with the outcome).

The likelihood ratio test is typically used for the overall test. This test compares the likelihood function value when the predictors are included in the logistic model with the likelihood function value when they are not included. The actual test statistic is:

$$\chi^2 = -2\, ln\left[\frac{\text{likelihood function when predictors are not in the model}}{\text{likelihood function when predictors are in the model}}\right]$$

This statistic follows a chi-square distribution under the null hypothesis, with degrees of freedom equal to the number of terms (βs) being tested.

If the null hypothesis is rejected in the overall test, then partial tests can be conducted for the significance of each slope parameter (i.e., for the importance of each predictor), given that the other terms are in the model:

H0: $\beta_1 = 0$
HA: $\beta_1 \neq 0$, given that $X_2, ..., X_k$ are in the model

H0: $\beta_1 = 0$
HA: $\beta_2 \neq 0$, given that $X_1, X_3, \ldots, X_k$ are in the model

etc.

PROC LOGISTIC produces partial test results by default (in the 'Analysis of Maximum Likelihood Estimates' table and, sometimes, in a 'Type III Tests' table, depending on the form of the model). The tests are produced using the 'Wald' method, which assumes asymptotic normality of the estimators; other methods are available –e.g., the Score method or the likelihood ratio method –but either have to be specifically requested in the LOGISTIC step, or must be produced in another procedure such as PROC GENMOD). When sample sizes are large, the results will be similar across the methods; when sample sizes are small, the likelihood ratio test is more robust.

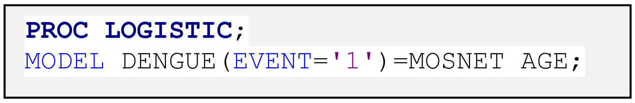
Example: Using the (heavily edited) SAS output on the next page, conduct the overall and partial tests for the following model:

$$Pr(Y = 1 | X_1 = x_1, X_2 = x_2) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2)}}$$

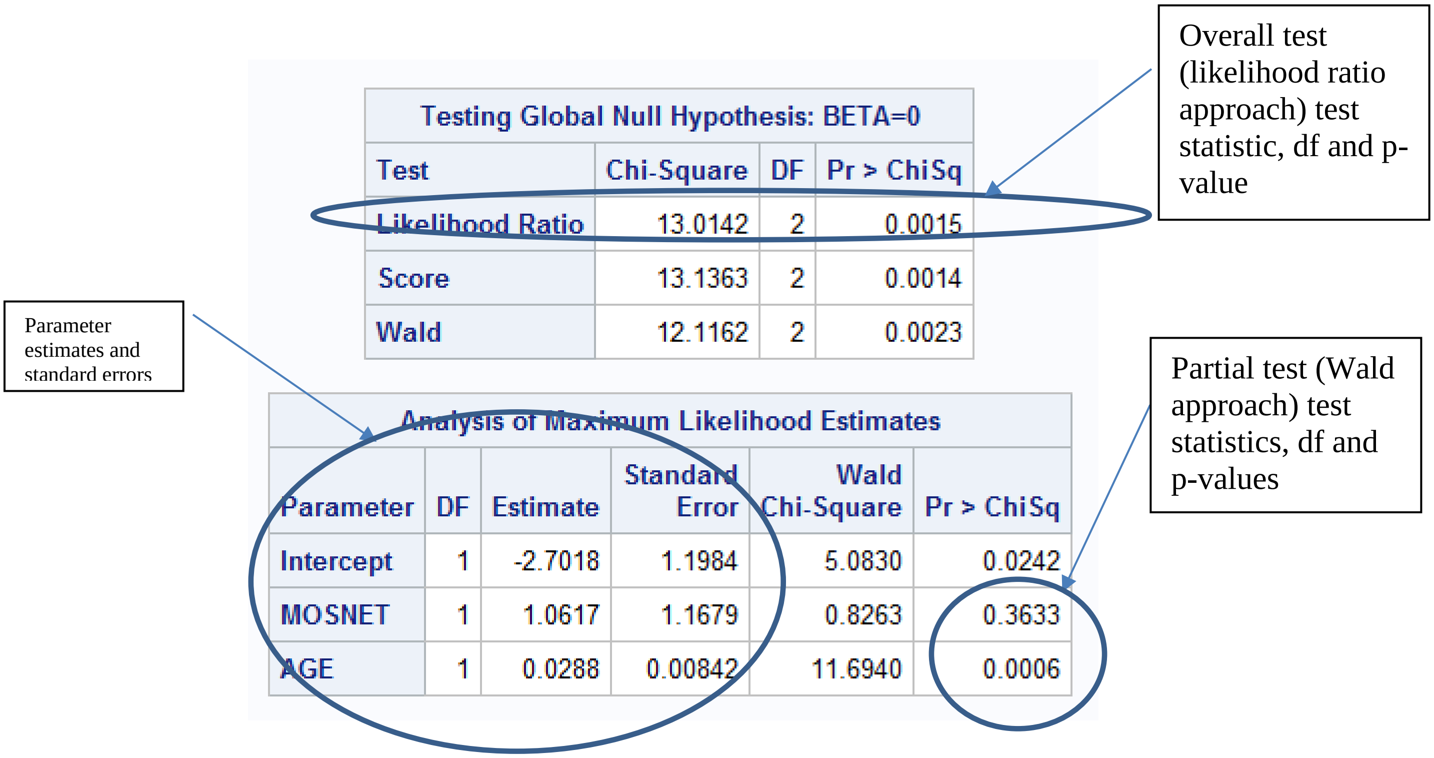
Where Y = 1 if dengue fever, 0 else; X_1 = 1 if mosquito net not used, 0 else; X_2 = age

Overall Test:

Partial Tests:



```
PROC LOGISTIC;
MODEL DENGUE(EVENT='1')=MOSNET AGE;
```



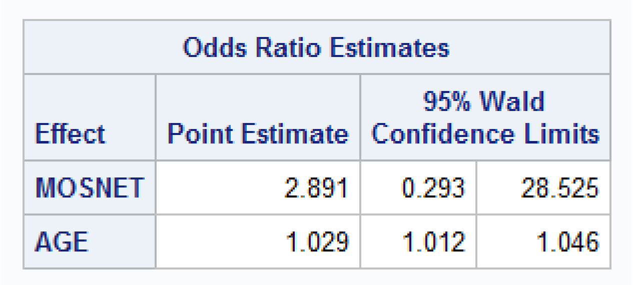
Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	13.0142	2	0.0015
Score	13.1363	2	0.0014
Wald	12.1162	2	0.0023

Overall test (likelihood ratio approach) test statistic, df and p-value

Parameter estimates and standard errors

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-2.7018	1.1984	5.0830	0.0242
MOSNET	1	1.0617	1.1679	0.8263	0.3633
AGE	1	0.0288	0.00842	11.6940	0.0006

Partial test (Wald approach) test statistics, df and p-values



Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
MOSNET	2.891	0.293	28.525
AGE	1.029	1.012	1.046

§7.2.9 Logistic Regression Inferences: Odds Ratio Calculations Using SAS

Often, after conducting the overall test, analysts will report confidence intervals on odds ratios rather than performing partial tests. If such a confidence interval includes 1, then the relevant predictor is not important (i.e., the slope parameter for that predictor is not significantly different from zero) given that the other predictors are in the model.

We have performed some odds ratio calculations by hand. In SAS (version 9.4), confidence intervals for odds ratios are automatically produced for each predictor, but only for the situation where the difference in predictor variable values is 1 (i.e., only for a "one unit increase" in the predictor).

To illustrate, consider the previous example. The odds ratio for age in the SAS output is 1.029. This is the odds ratio comparing the odds of dengue for a person who is who is one year older with the odds of dengue for a person who is one year younger (e.g., a 40-year-old compared with a 39-year-old). This odds ratio is simply e^{β_2} (why?)

SAS *also* presents a 95% confidence interval for this odds ratio ($e^{\hat{\beta}_2 \pm 1.96 SE(\hat{\beta}_2)}$): we are 95% confident that the odds of dengue fever for the person who is older by one year are between 1.012 and 1.046 times the odds for the person who is younger, adjusted for mosquito net use. (another way of saying this is that the odds for the older person are between 1.2% and 4.6% higher than for the younger person).

Interpret the odds ratio confidence interval for mosquito net use that is found in the SAS output:

Often, we need confidence intervals for odds ratios which involve something other than a one-unit difference in the predictor. In such situations, the UNITS statement can be used in SAS.

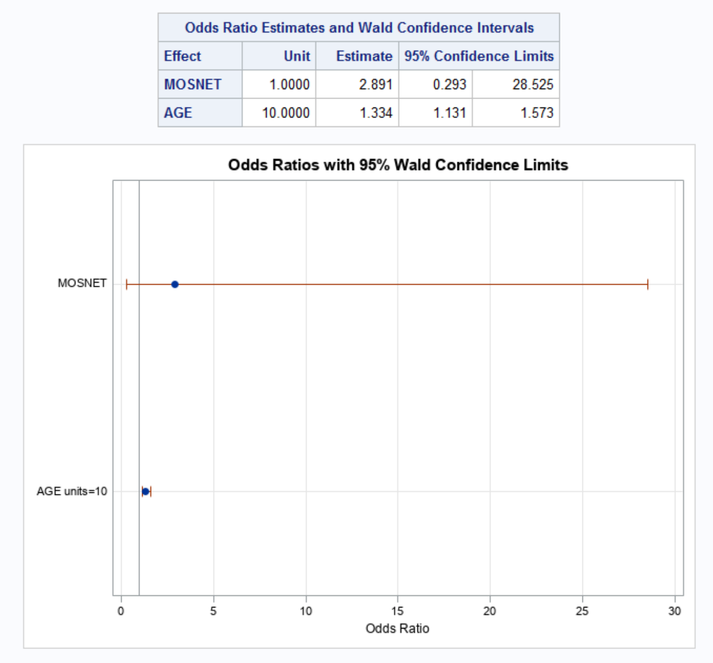
Example: In the previous model, find a 95% confidence interval for the odds ratio comparing the odds of dengue fever for 40-year-olds with the odds for 30-year-olds (adjusted for mosquito net use). The odds ratio here is $e^{10\beta_2}$ (why? See the example on page 87.)

By Hand:

exp[10 $\hat{\beta}_2$ ± $Z_{1-\alpha/2}$ ·10·SE($\hat{\beta}_2$)]

Using SAS:

```
PROC LOGISTIC;
  MODEL dengue(EVENT='1') = mosnet age /CLODDS=wald;
  UNITS age=10 / DEFAULT=1;
RUN;
```



Odds Ratio Estimates and Wald Confidence Intervals				
Effect	Unit	Estimate	95% Confidence Limits	
MOSNET	1.0000	2.891	0.293	28.525
AGE	10.0000	1.334	1.131	1.573

Odds Ratios with 95% Wald Confidence Limits

§7.2.10 Creating Dummy Variables Using the CLASS statement in PROC LOGISTIC

Consider the model below:

Y = 1 if dengue fever, 0 else
X_1 = geographic sector (=1, 2 ,3, 4 or 5).
X_2 = 1 if mosquito net not used, 0 else
X_3 = age (years)

The categorical predictor for sector needs to be re-coded using dummy variables:
Z_1 = 1 if sector 1, 0 else
Z_2 = 1 if sector 2, 0 else
Z_3 = 1 if sector 3, 0 else
Z_4 = 1 if sector 4, 0 else

The model (without interactions) will then be:

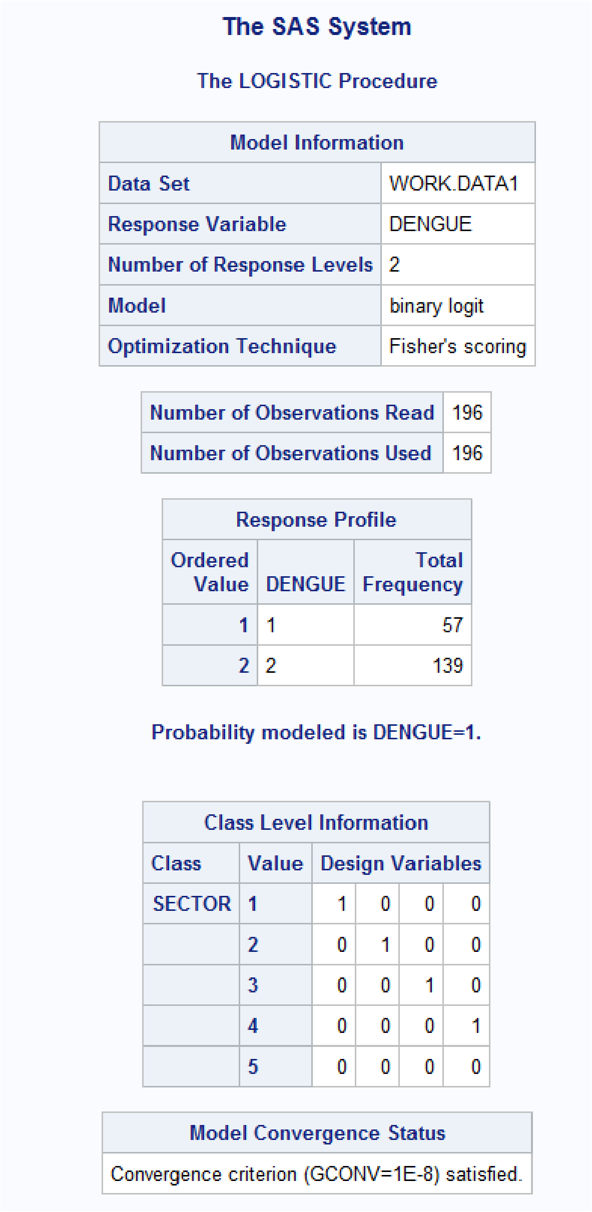
The CLASS statement in PROC LOGISTIC can be used to create these dummy variables:

```
PROC LOGISTIC;
  CLASS sector (PARAM=ref REF='5');
  MODEL dengue(EVENT='1') = sector mosnet age /CLODDS=wald;
  UNITS age=10 / DEFAULT=1
RUN;
```

The CLASS statement will cause the four dummy variables for sector to be created during the PROC LOGISTIC run (note: the dummy variables will not be added to the main data set itself), with reference-cell coding as the method of creating the dummy variables (‘PARAM=ref’) and sector 5 as the reference sector (REF=’5’).

Note: If the CLASS statement is omitted, then sector will effectively be treated as a pseudo-continuous variable, with the implicit assumption being that the difference (if any) in the odds of dengue fever between two sectors is the same regardless of which two sectors are being considered.

Example: The code above produces the output shown on the next page. Use it to perform the overall test and to examine the odds ratio estimates and confidence intervals.



The SAS System

The LOGISTIC Procedure

Model Information	
Data Set	WORK.DATA1
Response Variable	DENGUE
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring

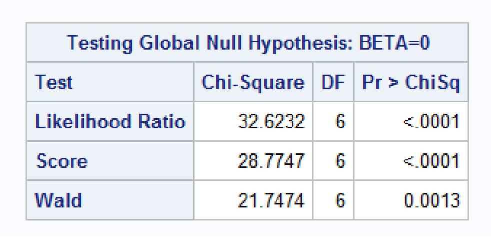
Number of Observations Read	196
Number of Observations Used	196

Response Profile		
Ordered Value	DENGUE	Total Frequency
1	1	57
2	2	139

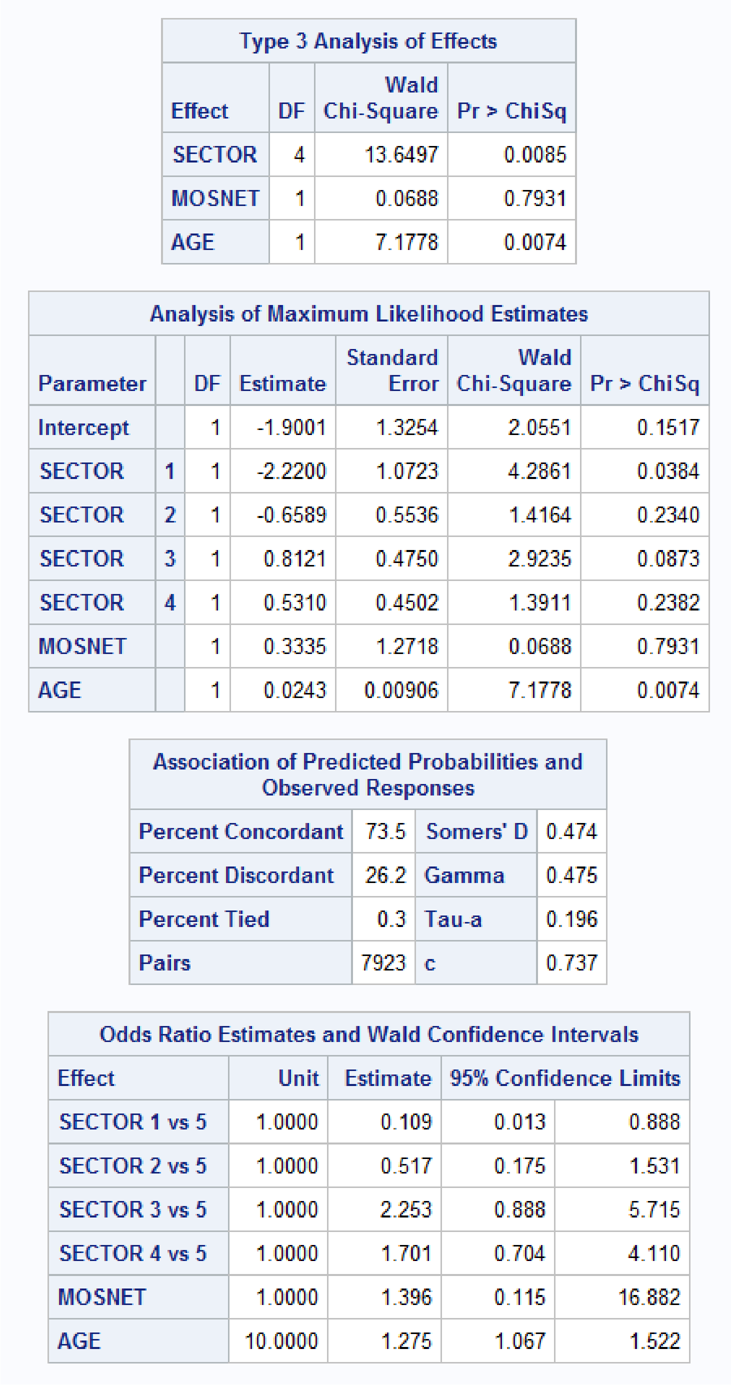
Probability modeled is DENGUE=1.

Class Level Information					
Class	Value	Design Variables			
SECTOR	1	1	0	0	0
	2	0	1	0	0
	3	0	0	1	0
	4	0	0	0	1
	5	0	0	0	0

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.



Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	32.6232	6	<.0001
Score	28.7747	6	<.0001
Wald	21.7474	6	0.0013



Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
SECTOR	4	13.6497	0.0085
MOSNET	1	0.0688	0.7931
AGE	1	7.1778	0.0074

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-1.9001	1.3254	2.0551	0.1517
SECTOR	1	1	-2.2200	1.0723	4.2861	0.0384
SECTOR	2	1	-0.6589	0.5536	1.4164	0.2340
SECTOR	3	1	0.8121	0.4750	2.9235	0.0873
SECTOR	4	1	0.5310	0.4502	1.3911	0.2382
MOSNET		1	0.3335	1.2718	0.0688	0.7931
AGE		1	0.0243	0.00906	7.1778	0.0074

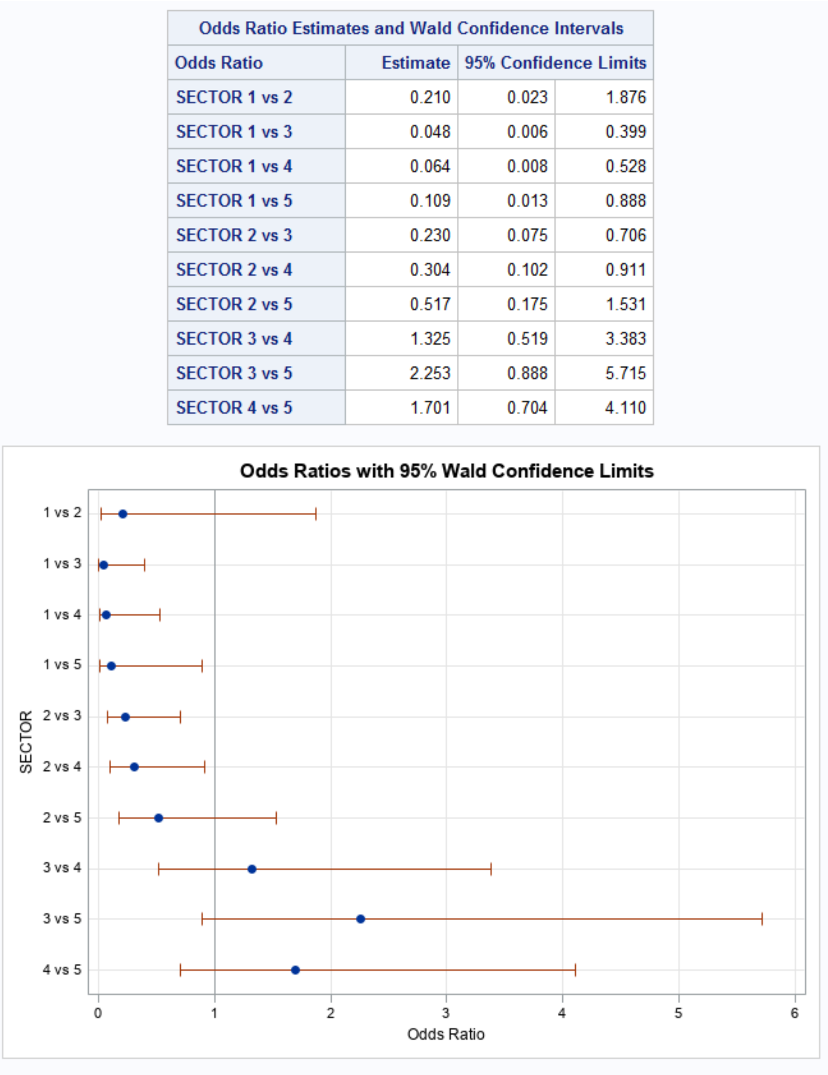
Association of Predicted Probabilities and Observed Responses			
Percent Concordant	73.5	Somers' D	0.474
Percent Discordant	26.2	Gamma	0.475
Percent Tied	0.3	Tau-a	0.196
Pairs	7923	c	0.737

Odds Ratio Estimates and Wald Confidence Intervals				
Effect	Unit	Estimate	95% Confidence Limits	
SECTOR 1 vs 5	1.0000	0.109	0.013	0.888
SECTOR 2 vs 5	1.0000	0.517	0.175	1.531
SECTOR 3 vs 5	1.0000	2.253	0.888	5.715
SECTOR 4 vs 5	1.0000	1.701	0.704	4.110
MOSNET	1.0000	1.396	0.115	16.882
AGE	10.0000	1.275	1.067	1.522

Note that odds ratio point estimates (and confidence intervals) comparing non-reference sectors with each other are not produced. The ODDSRATIO statement can be added to the PROC LOGISTIC step to produce them:

```
PROC LOGISTIC;
  CLASS sector (param=ref ref='5');
  MODEL dengue(EVENT='1') = sector mosnet age /CLODDS=wald;
  UNITS age=10 / default=1;
  ODDSRATIO sector / diff=all;
RUN;
```

Edited SAS output:



Odds Ratio Estimates and Wald Confidence Intervals

Odds Ratio	Estimate	95% Confidence Limits	
SECTOR 1 vs 2	0.210	0.023	1.876
SECTOR 1 vs 3	0.048	0.006	0.399
SECTOR 1 vs 4	0.064	0.008	0.528
SECTOR 1 vs 5	0.109	0.013	0.888
SECTOR 2 vs 3	0.230	0.075	0.706
SECTOR 2 vs 4	0.304	0.102	0.911
SECTOR 2 vs 5	0.517	0.175	1.531
SECTOR 3 vs 4	1.325	0.519	3.383
SECTOR 3 vs 5	2.253	0.888	5.715
SECTOR 4 vs 5	1.701	0.704	4.110