Dengue Fever Disease Status Prediction

Author: Jinye Lu

Introduction

Dengue fever is caused by the dengus virus. And we have some data in Mexico about individuals who do not have the disease and those who do have the disease.

If we are able to develop generalized linear models to predict disease status based on characteristics of the individual and where they live, it might be easier for stakeholders to determine areas of the city and parts of population that are more at risk. And they can plan resource allocation to treat the disease. Also, they can try to assist vulnerable populations with preparation and preventative measures.

Dataset

The data contains disease population information collected in Mexico. The dataset contains the following variables:

Variable in dataset	**Variable**	**Explanation**
ID.num	ID number	Identifies each individual in the dataset with a number from 1 - 196
age	Age (years)	The age of the person in years
socioeconomic.status	Socioeconomic status	Socioeconomic status: upper, middle, lower
sector	City sector	Sector in the city where the person lives: sector 1, sector 2
disease.status	Disease status	Disease status of the individual: 1 if they have dengue fever, 0 if they do not
savings.account	Savings account	Whether or not the individual has a savings account: savings or no.savings

Preliminary Analysis

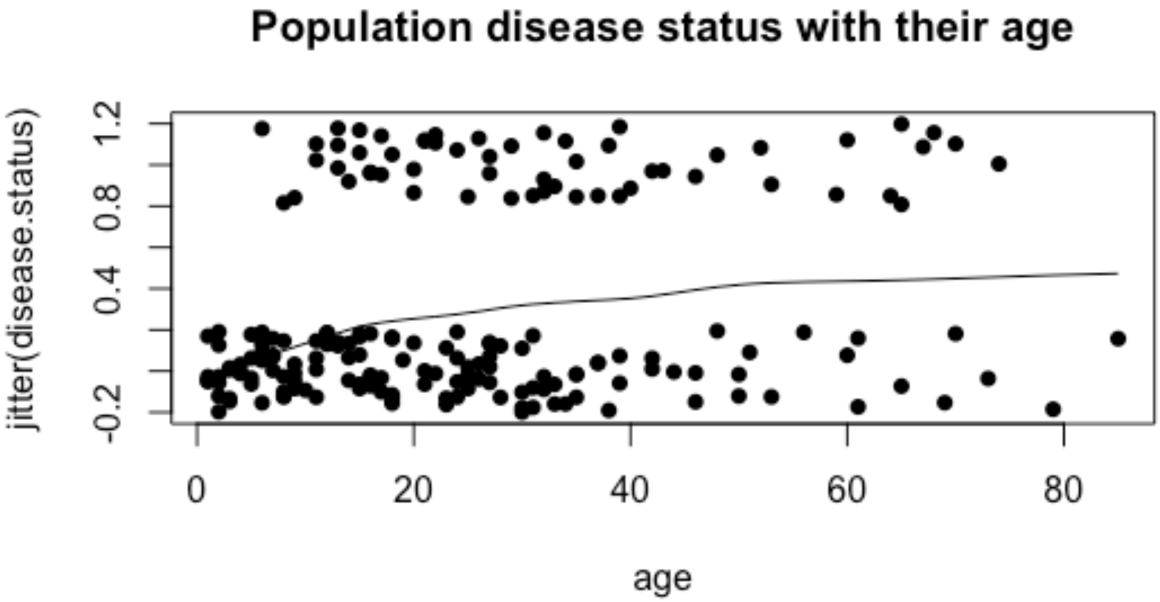
1. The proportion of individuals in the sample with the disease is 0.2908163. The variance of the dataset is 0.2072998.
2. The proportion of individuals with the disease for each of the category levels for socioeconomic status is shown in the table below:

Level	Upper	Middle	Lower
Proportion	0.3116883	0.2857143	0.2714286

3. The proportion of individuals with the disease who live in the different city sectors is shown in the table below:

Sector	1	2
Proportion	0.1880342	0.443038

4. From my point of view, **the probability of being infected with the disease increases as the age of the person decreases**. Because as people age their immunity systems become stronger than youngsters. So they would be able to resist dengue fever infection from the mosquitoes' transmission.
   Here is the plot of the disease status against age below.



From the plot I find the pattern that more points located in the y=1 line on the right side line whereas more points located in the y=0 line on the left side. So combining the lowess line, the pattern seems like that **older people have a larger proportion of disease than youngsters**. It is not consistent with my prediction.

Descriptive and Predictive Analytics with model fit

Null Model

5. Create a null model. The equation for the model, including values for the intercept is shown below:

$$ln(\frac{p_i}{1-p_i}) = -0.8914 + \varepsilon_i$$

Link function: $g(u_i)=\ln\frac{u_i}{1-u_i}$

Model A

6. Create a model with age as an explanatory variable. The equation for the model, including values for the intercept and co-efficients is shown below:

$$ln(\frac{p_i}{1-p_i}) = -1.659043 + 0.028464*\text{age}_i + \varepsilon_i$$

Link function: $g(u_i)=\ln\frac{u_i}{1-u_i}$

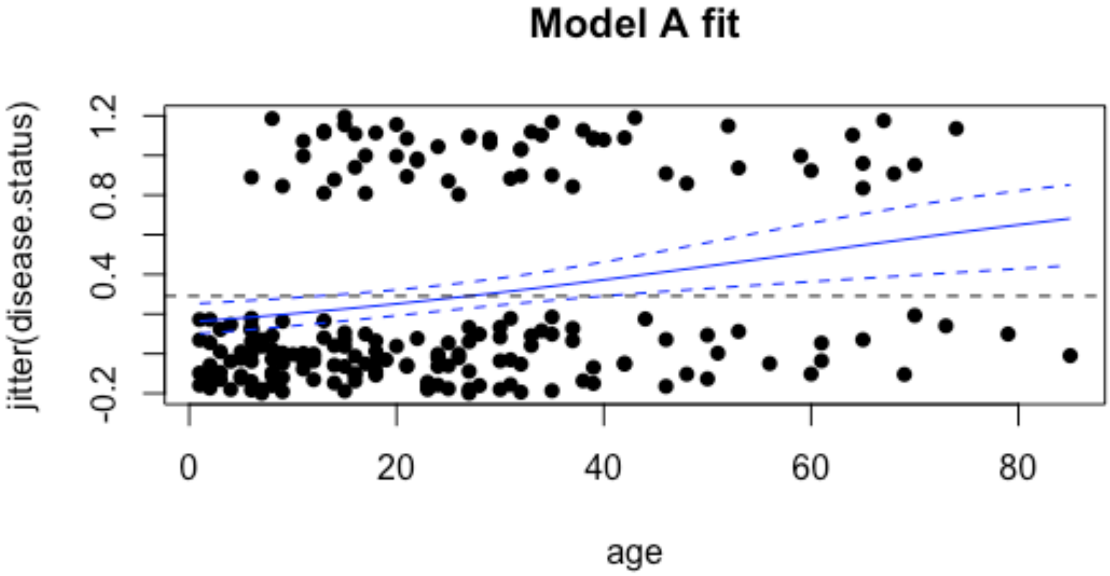
The residual deviance is 224.32 on 194 degrees of freedom.

According to the value of the residual deviance, it is a little large value, which indicates the model does not fit the saturated model well to some extent. So we can try other model to get a better fit of the data.

7.

Likelihood ratio test	
H_0: No significant difference between the model with age to the null model	Reduced model variables: ln L_R = -118.1647
H_1: The two models have significantly different likelihoods.	Full model variables: ln L_F =-112.1582
Calculation of likelihood ratio test statistic: G = -2 $\ln\left(\frac{L_R}{L_F}\right) = -2(\ln L_R - \ln L_F)$ = -2*(-118.1647+112.1582)=12.013	
p-value = 0.0005283	Do you reject H_0 or fail to reject H_0: Reject
What does this mean for your model? According to the result, the model with age have significantly different likelihoods. So this means that **the 'age' variable is good indicator for predicting the disease status**.	

8.



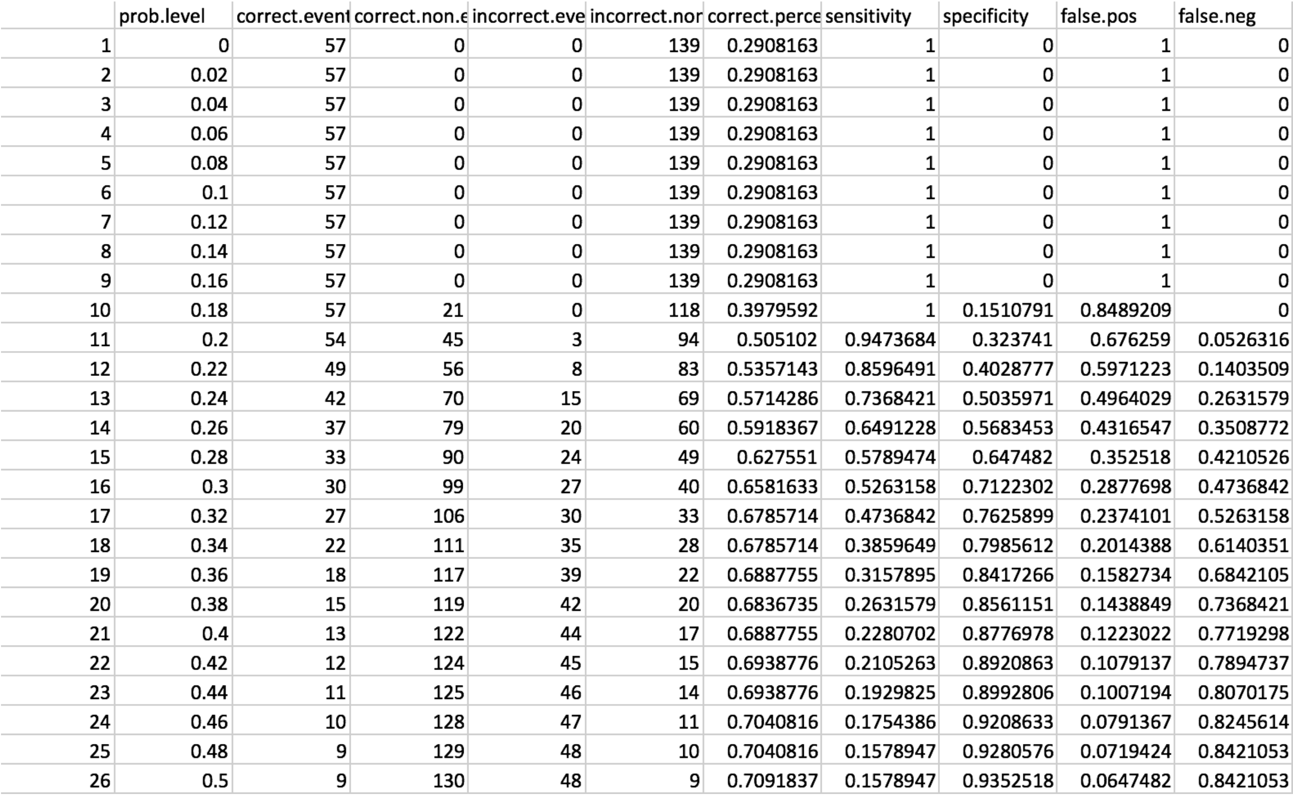
On the plot, the pattern of the model fit is clear that as the age of population increases, the proportion of the people who have the disease enhances. And the prediction interval become larger as the age of the population increases, which means that the uncertainly of the prediction increases.

While the average disease status of the sample is 0.2908163, the age that the probability of having the disease become higher than the average is calculated as below:

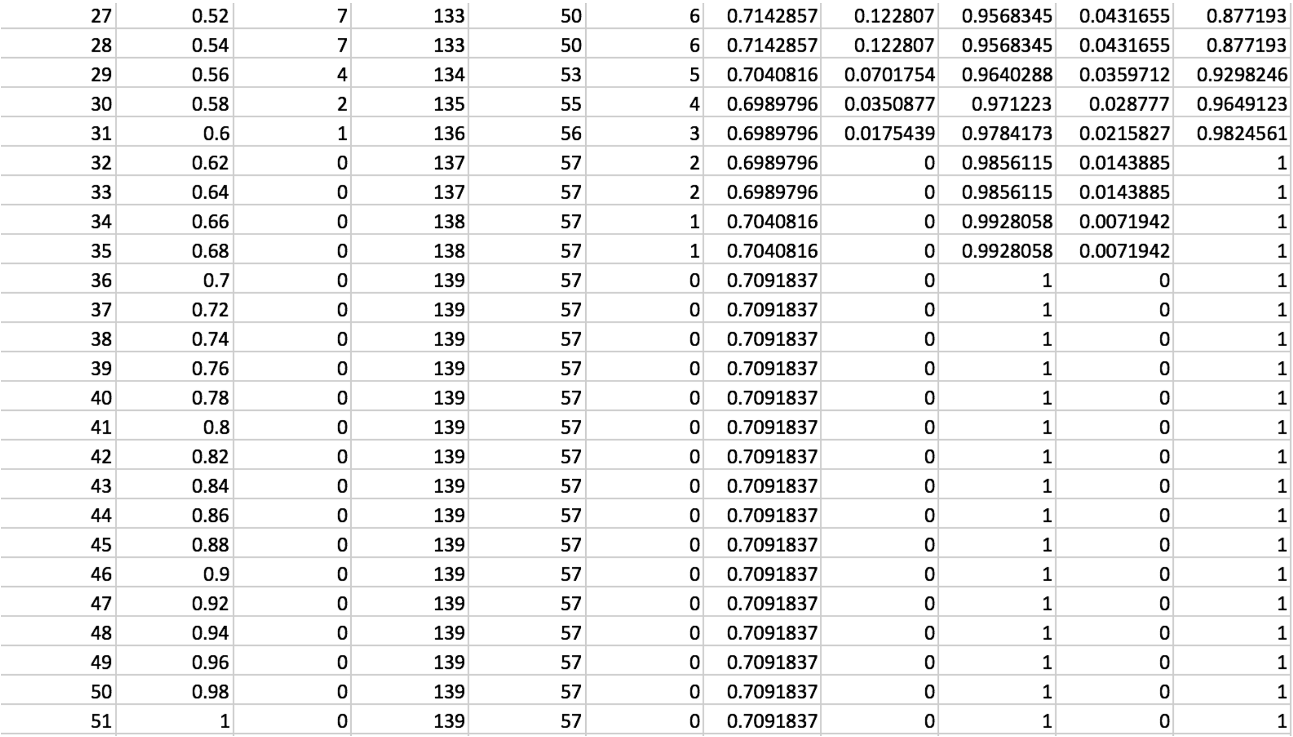
ln(0.2908163/(1-0.2908163)+1.659043)/0.028464=26.96811

After rounding up the value, the age is 27 year olds.

9. Create a classification table for this model with 51 rows.



	prob.level	correct.event	correct.non.e	incorrect.eve	incorrect.nor	correct.perce	sensitivity	specificity	false.pos	false.neg
1	0	57	0	0	139	0.2908163	1	0	1	0
2	0.02	57	0	0	139	0.2908163	1	0	1	0
3	0.04	57	0	0	139	0.2908163	1	0	1	0
4	0.06	57	0	0	139	0.2908163	1	0	1	0
5	0.08	57	0	0	139	0.2908163	1	0	1	0
6	0.1	57	0	0	139	0.2908163	1	0	1	0
7	0.12	57	0	0	139	0.2908163	1	0	1	0
8	0.14	57	0	0	139	0.2908163	1	0	1	0
9	0.16	57	0	0	139	0.2908163	1	0	1	0
10	0.18	57	21	0	118	0.3979592	1	0.1510791	0.8489209	0
11	0.2	54	45	3	94	0.505102	0.9473684	0.323741	0.676259	0.0526316
12	0.22	49	56	8	83	0.5357143	0.8596491	0.4028777	0.5971223	0.1403509
13	0.24	42	70	15	69	0.5714286	0.7368421	0.5035971	0.4964029	0.2631579
14	0.26	37	79	20	60	0.5918367	0.6491228	0.5683453	0.4316547	0.3508772
15	0.28	33	90	24	49	0.627551	0.5789474	0.647482	0.352518	0.4210526
16	0.3	30	99	27	40	0.6581633	0.5263158	0.7122302	0.2877698	0.4736842
17	0.32	27	106	30	33	0.6785714	0.4736842	0.7625899	0.2374101	0.5263158
18	0.34	22	111	35	28	0.6785714	0.3859649	0.7985612	0.2014388	0.6140351
19	0.36	18	117	39	22	0.6887755	0.3157895	0.8417266	0.1582734	0.6842105
20	0.38	15	119	42	20	0.6836735	0.2631579	0.8561151	0.1438849	0.7368421
21	0.4	13	122	44	17	0.6887755	0.2280702	0.8776978	0.1223022	0.7719298
22	0.42	12	124	45	15	0.6938776	0.2105263	0.8920863	0.1079137	0.7894737
23	0.44	11	125	46	14	0.6938776	0.1929825	0.8992806	0.1007194	0.8070175
24	0.46	10	128	47	11	0.7040816	0.1754386	0.9208633	0.0791367	0.8245614
25	0.48	9	129	48	10	0.7040816	0.1578947	0.9280576	0.0719424	0.8421053
26	0.5	9	130	48	9	0.7091837	0.1578947	0.9352518	0.0647482	0.8421053

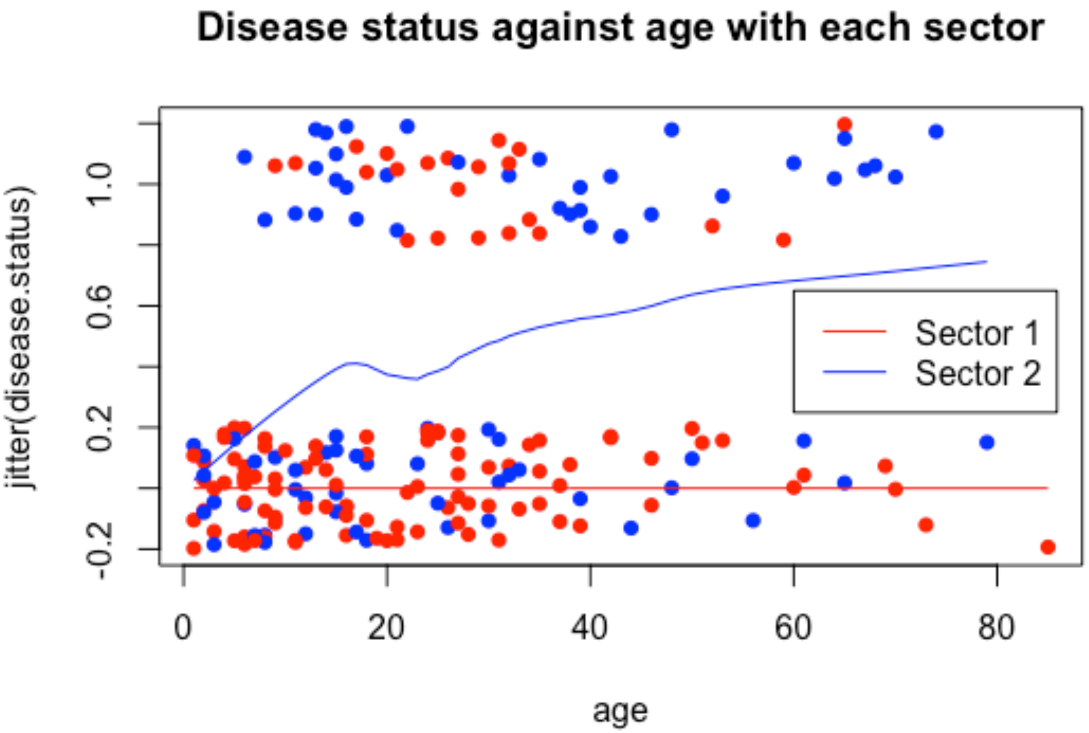


27	0.52	7	133	50	6	0.7142857	0.122807	0.9568345	0.0431655	0.877193
28	0.54	7	133	50	6	0.7142857	0.122807	0.9568345	0.0431655	0.877193
29	0.56	4	134	53	5	0.7040816	0.0701754	0.9640288	0.0359712	0.9298246
30	0.58	2	135	55	4	0.6989796	0.0350877	0.971223	0.028777	0.9649123
31	0.6	1	136	56	3	0.6989796	0.0175439	0.9784173	0.0215827	0.9824561
32	0.62	0	137	57	2	0.6989796	0	0.9856115	0.0143885	1
33	0.64	0	137	57	2	0.6989796	0	0.9856115	0.0143885	1
34	0.66	0	138	57	1	0.7040816	0	0.9928058	0.0071942	1
35	0.68	0	138	57	1	0.7040816	0	0.9928058	0.0071942	1
36	0.7	0	139	57	0	0.7091837	0	1	0	1
37	0.72	0	139	57	0	0.7091837	0	1	0	1
38	0.74	0	139	57	0	0.7091837	0	1	0	1
39	0.76	0	139	57	0	0.7091837	0	1	0	1
40	0.78	0	139	57	0	0.7091837	0	1	0	1
41	0.8	0	139	57	0	0.7091837	0	1	0	1
42	0.82	0	139	57	0	0.7091837	0	1	0	1
43	0.84	0	139	57	0	0.7091837	0	1	0	1
44	0.86	0	139	57	0	0.7091837	0	1	0	1
45	0.88	0	139	57	0	0.7091837	0	1	0	1
46	0.9	0	139	57	0	0.7091837	0	1	0	1
47	0.92	0	139	57	0	0.7091837	0	1	0	1
48	0.94	0	139	57	0	0.7091837	0	1	0	1
49	0.96	0	139	57	0	0.7091837	0	1	0	1
50	0.98	0	139	57	0	0.7091837	0	1	0	1
51	1	0	139	57	0	0.7091837	0	1	0	1

The maximum of the percent correct is 0.7143. And when the probability level is at 0.52 or 0.54, the percent correct is at the highest level. So I would use the probability level in the range of 0.52 to 0.54 as the cut-off.
The sensitivity associated with this cut-off is 0.1228. And the specificity associated with this cut-off is 0.9568.

10. In this particular situation, our objective is to identify the parts of population that are more at risk. So we would focus on the correctness of the event (sensitivity) based on the prediction results instead of the non-event. Since the false negative is the percentage of the incorrect false event and equal to 1-sensitivity, we would like to minimize the error to gain a clear look of the population who have the disease. Even if the predicted disease person turns out not be infected, the circumstance is less serious than the actual disease person being predicted to be healthy. So the false negative is more serious.

11.

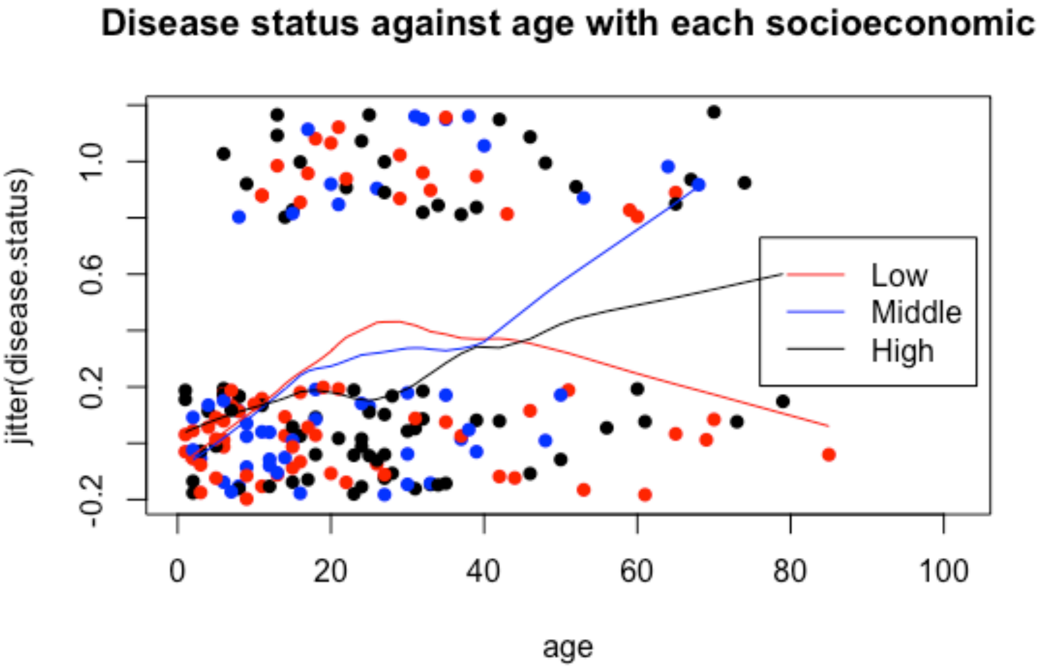


From the plot, we can find that more people in sector 2 have the disease than sector 1. So the sector 2 is more at risk than sector 1.

For people in sector 2, elders have higher disease risk than youngsters.

For people in sector 1, people who have the disease are mainly clustered in the age of 20 to 40. But since a lot of people in this age do not have the disease similarly, the lowess line's slope does not change.

12.



From the plot, we can find the most risk population is the elders in middle socioeconomic status. For middle and upper socioeconomic population, more elders have the disease than youngsters. However, for low socioeconomic population, youngsters whose age is between 20 and 40 have higher risk of being infected.

Model B with age, sector and socioeconomic status

13. Create the model with age, sector and socioeconomic status. The model is shown as below:

$y_i = b_0 + b_1$*$\text{age}_i + b_2$ * $\text{socioeconomic.status.middle}_i + b_3$ * $\text{socioeconomic.status.upper}_i + b_4$ * $\text{sector.sector2}_i + b_5$ * $\text{socioeconomic.status.middle}_i$ * $\text{age}_i + b_6$ * $\text{socioeconomic.status.upper}_i$ * $\text{age}_i + b_7$ * sector.sector2_i * age_i

14. The full model is that the model contains age, sector, socioeconomic status and the interactions between the age and sector, socioeconomic status.

Likelihood ratio test	
H_0: No significant difference between the full model and the null model	Null model variables: ln L_R = -118.1647
H_1: The two models have significantly different likelihoods.	Full model variables: ln L_F = -104.5699
Calculation of likelihood ratio test statistic: $G = -2\ \ln\left(\frac{L_R}{L_F}\right) = -2(\ln L_R - \ln L_F)$ = -2*(-118.1647+104.5699)= 27.1896	
p-value = 0.0003081	Do you reject H_0 or fail to reject H_0: Reject
What does this mean for your model? According to the result, the model with age, sector and socioeconomic status have significantly different likelihoods. So this means that **the explanatory variable either 'age' or 'sector' or 'socioeconomic status' are good indicators to predict the disease status**.	

15.

(1) Test the 'age' variable

In order to test the significance of the "age" variable, we would build a reduced model that only contains the sector and socioeconomic status as the explanatory variables.

Likelihood ratio test	
H_0: No significant difference between the full model to the reduced model.	Reduced model variables: ln L_R = -110.6313
H_1: The two models have significantly different likelihoods.	Full model variables: ln L_F = -104.5699
Calculation of likelihood ratio test statistic: $G = -2\ \ln\left(\frac{L_R}{L_F}\right) = -2(\ln L_R - \ln L_F)$ = -2*(-110.6313+104.5699)= 12.123	
p-value = 0.01646	Do you reject H_0 or fail to reject H_0: Reject
What does this mean for your model? According to the result, the model with age, sector and socioeconomic status have significantly different likelihoods from the model with sector and socioeconomic status.	

So this means that **the explanatory variable 'age' is a good indicator to predict the disease status**.

(2) Test the 'sector' variable

Since the significance of the "age" have been confirmed, the next step is to test the "sector". In order to test the significance of it, we would build a reduced model that only contains the age, socioeconomic status and the interaction between them as the explanatory variables.

Likelihood ratio test	
H_0: No significant difference between the full model to the reduced model	Reduced model variables: ln L_R = -110.825
H_1: The two models have significantly different likelihoods.	Full model variables: ln L_F = -104.5699
Calculation of likelihood ratio test statistic: G = -2 $\ln\left(\frac{L_R}{L_F}\right) = -2(\ln L_R - \ln L_F)$ = -2*(-110.825 +104.5699)= 12.51	
p-value = 0.001921	Do you reject H_0 or fail to reject H_0: Reject
What does this mean for your model? According to the result, the model with age, sector and socioeconomic status have significantly different likelihoods from the model with age and socioeconomic status. So this means that **the explanatory variable 'sector' is a good indicator to predict the disease status**.	

(3) Test the 'socioeconomic status' variable

Since the significance of the "sector" have been confirmed, the next step is to test the "*socioeconomic status*". In order to test the significance of it, we would build a reduced model that only contains the age, sector and the interaction between them as the explanatory variables.

Likelihood ratio test	
H_0: No significant difference between the full model and the reduced model.	Reduced model variables: ln L_R = -105.6882
H_1: The two models have significantly different likelihoods.	Full model variables: ln L_F = -104.5699
Calculation of likelihood ratio test statistic: G = -2 $\ln\left(\frac{L_R}{L_F}\right) = -2(\ln L_R - \ln L_F)$ = -2*(-105.6882+104.5699)= 2.2365	
p-value = 0.6923	Do you reject H_0 or fail to reject H_0: Fail to reject
What does this mean for your model? According to the result, the model with age, sector and socioeconomic status does not have significantly different likelihood from the model with age and sector. So this means that **the explanatory variable 'socioeconomic status' is not a good indicator to predict the disease status**. It should be removed from the full model.	

16. Since the "sector" categorical variable is significant, test if the interaction is significant using another likelihood ratio test.

Likelihood ratio test	
H_0: No significant difference between **the model with the interaction between age and sector** to **the model without the interaction between them**	Reduced model variables: ln L_R = -105.8196
H_1: The two models have significantly different likelihoods.	Full model variables: ln L_F = -105.6882
Calculation of likelihood ratio test statistic: G = -2 $\ln\left(\frac{L_R}{L_F}\right) = -2(\ln L_R - \ln L_F)$ = -2*(-105.8196+105.6882)= 0.26293	
p-value = 0.6081	Do you reject H_0 or fail to reject H_0: Fail to reject
What does this mean for your model? According to the result, the model without the **the interaction between age and sector** does not have significantly different likelihood from the model with **the interaction between them**. So this means that the explanatory variable which is **the interaction between them is not a good indicator to predict the disease status**. It should be removed from the full model.	

Final Model

17. The equation for the final model, including values for the intercept and co-efficients is shown below:

$$ln(\frac{p_i}{1-p_i})= \text{-2.15966 + 1.18169*sector}_i + \text{0.02681*age}_i + \varepsilon_i$$

Link function: $g(u_i)=\ln\frac{u_i}{1-u_i}$

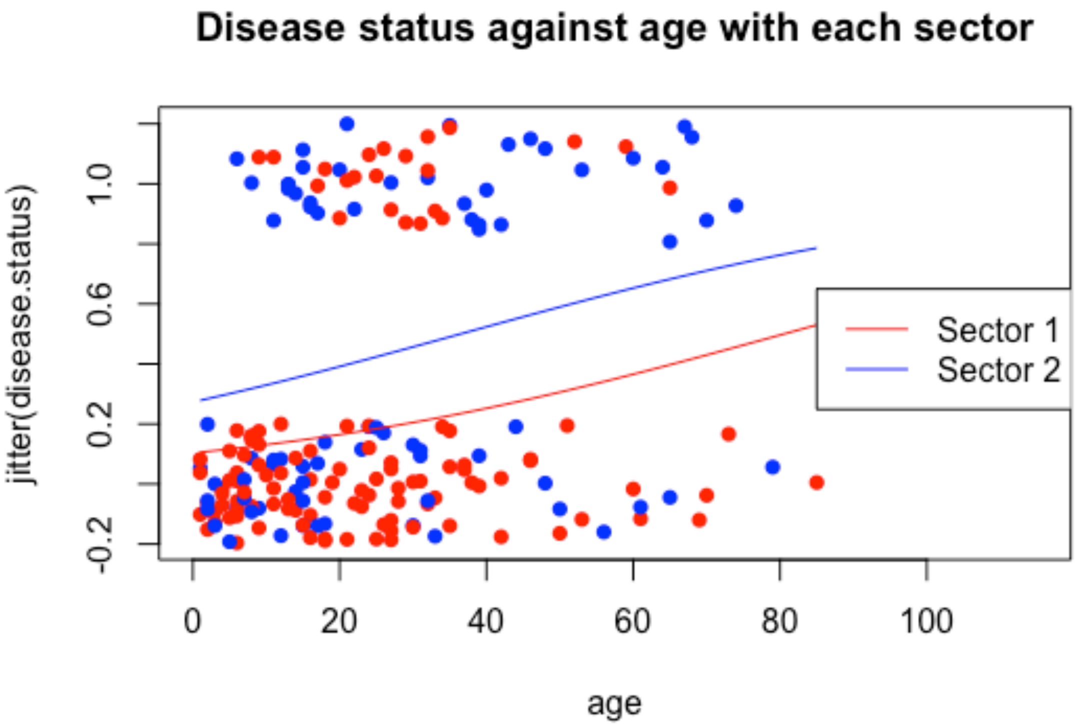
The intercept determines the minimum of the log odds ratio. The co-efficient of the sector determines the intercept adjustment when the sector is “sector 2”. And the co-efficient of the age determines the slope of the log odds ratio change as the age changes.

18. The residual deviance is 211.64 on 193 degrees of freedom. So the residual deviance / residual df is equal to 211.64/193=1.09658.
This value indicates that the model is underdispersion.

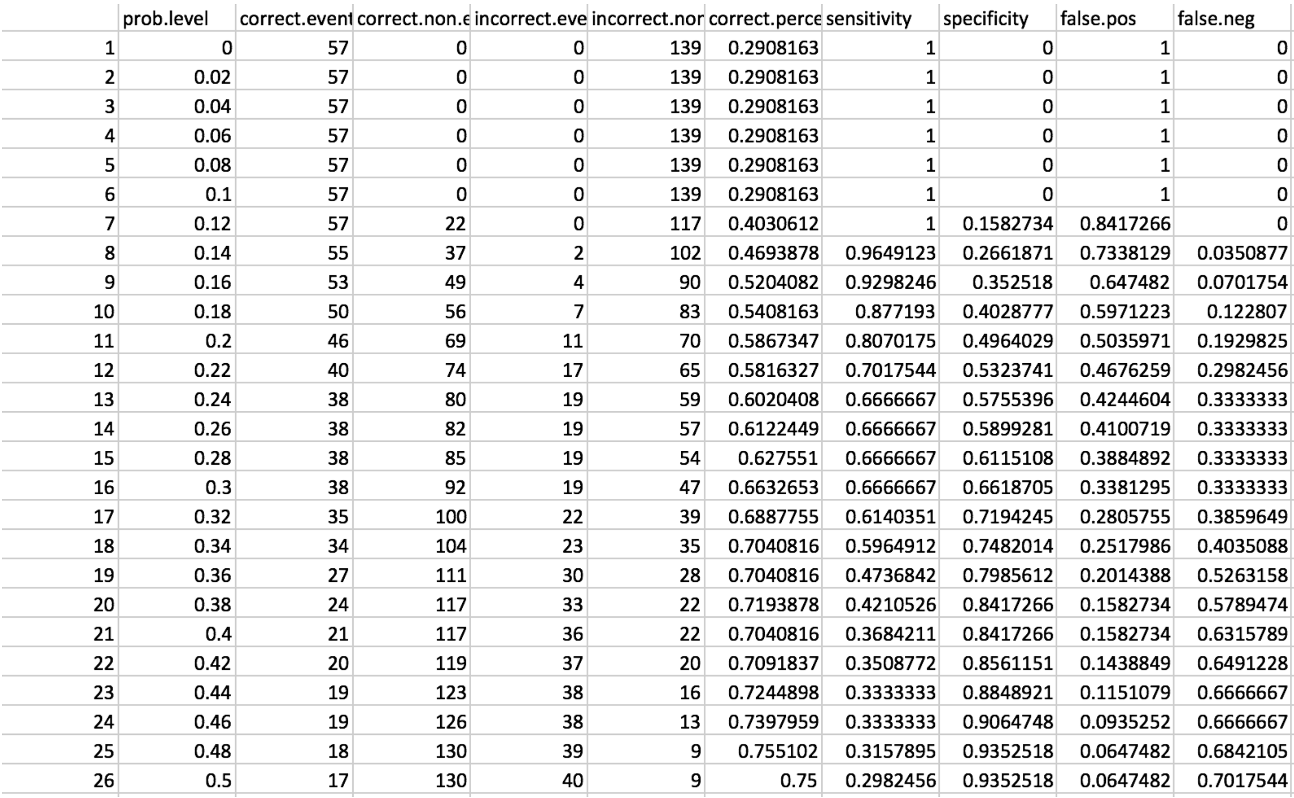
- The statistical independence of observations would be met if all the data is collected at same period. So there would be independent of time.
- The correct specification of link function is also met. Since the response data is binary, the link function ($g(u_i)=\ln\frac{u_i}{1-u_i}$) is appropriate for the type of data.
- The variance of the responses is 0.2072998. Moreover, what is expected from the link function is 0.2908163*(1-0.2908163)=0.20624219. The variance is close to the expected value, which means that the assumption is also met.

19. AIC = -2*lnL+2*(k+s)
Where k=level of y-1, s=number of predictor variables
So AIC = -2 * (-105.8196) + 2 * (1 + 2) = 217.6392

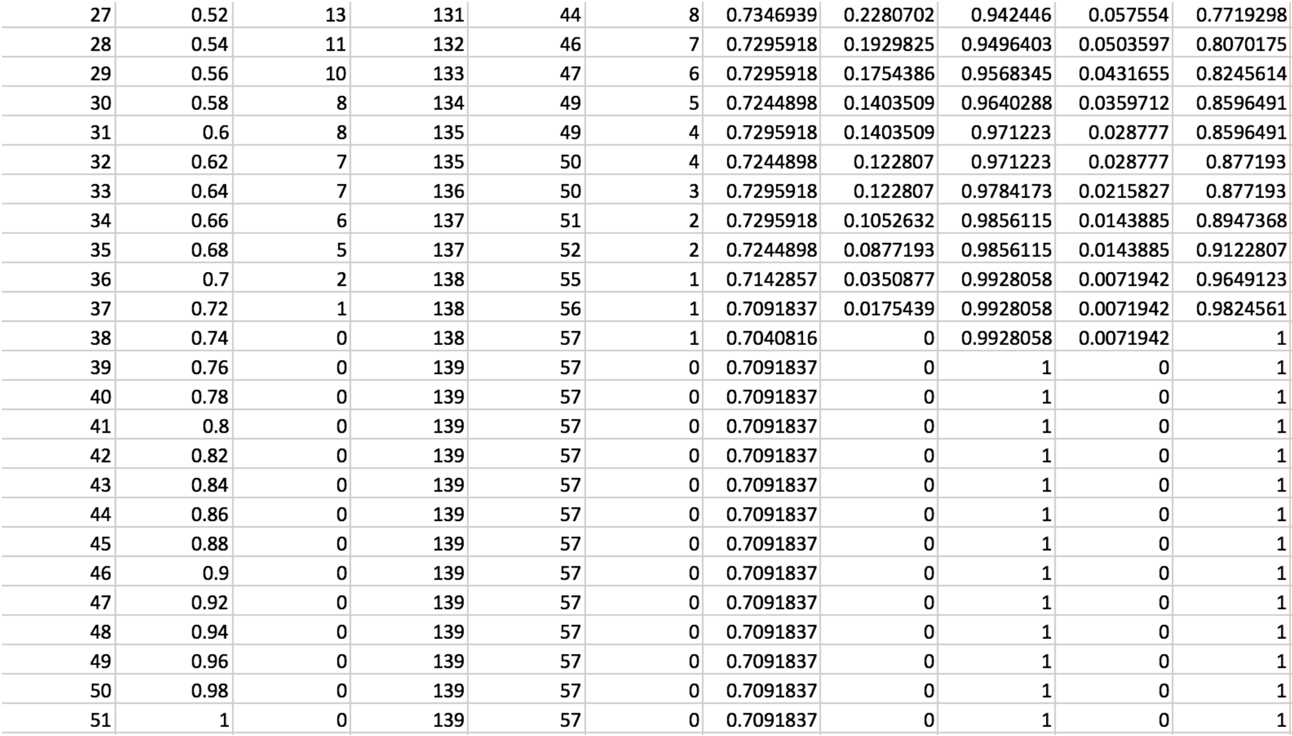
20.



21.



	prob.level	correct.event	correct.non.e	incorrect.eve	incorrect.nor	correct.perce	sensitivity	specificity	false.pos	false.neg
1	0	57	0	0	139	0.2908163	1	0	1	0
2	0.02	57	0	0	139	0.2908163	1	0	1	0
3	0.04	57	0	0	139	0.2908163	1	0	1	0
4	0.06	57	0	0	139	0.2908163	1	0	1	0
5	0.08	57	0	0	139	0.2908163	1	0	1	0
6	0.1	57	0	0	139	0.2908163	1	0	1	0
7	0.12	57	22	0	117	0.4030612	1	0.1582734	0.8417266	0
8	0.14	55	37	2	102	0.4693878	0.9649123	0.2661871	0.7338129	0.0350877
9	0.16	53	49	4	90	0.5204082	0.9298246	0.352518	0.647482	0.0701754
10	0.18	50	56	7	83	0.5408163	0.877193	0.4028777	0.5971223	0.122807
11	0.2	46	69	11	70	0.5867347	0.8070175	0.4964029	0.5035971	0.1929825
12	0.22	40	74	17	65	0.5816327	0.7017544	0.5323741	0.4676259	0.2982456
13	0.24	38	80	19	59	0.6020408	0.6666667	0.5755396	0.4244604	0.3333333
14	0.26	38	82	19	57	0.6122449	0.6666667	0.5899281	0.4100719	0.3333333
15	0.28	38	85	19	54	0.627551	0.6666667	0.6115108	0.3884892	0.3333333
16	0.3	38	92	19	47	0.6632653	0.6666667	0.6618705	0.3381295	0.3333333
17	0.32	35	100	22	39	0.6887755	0.6140351	0.7194245	0.2805755	0.3859649
18	0.34	34	104	23	35	0.7040816	0.5964912	0.7482014	0.2517986	0.4035088
19	0.36	27	111	30	28	0.7040816	0.4736842	0.7985612	0.2014388	0.5263158
20	0.38	24	117	33	22	0.7193878	0.4210526	0.8417266	0.1582734	0.5789474
21	0.4	21	117	36	22	0.7040816	0.3684211	0.8417266	0.1582734	0.6315789
22	0.42	20	119	37	20	0.7091837	0.3508772	0.8561151	0.1438849	0.6491228
23	0.44	19	123	38	16	0.7244898	0.3333333	0.8848921	0.1151079	0.6666667
24	0.46	19	126	38	13	0.7397959	0.3333333	0.9064748	0.0935252	0.6666667
25	0.48	18	130	39	9	0.755102	0.3157895	0.9352518	0.0647482	0.6842105
26	0.5	17	130	40	9	0.75	0.2982456	0.9352518	0.0647482	0.7017544



27	0.52	13	131	44	8	0.7346939	0.2280702	0.942446	0.057554	0.7719298
28	0.54	11	132	46	7	0.7295918	0.1929825	0.9496403	0.0503597	0.8070175
29	0.56	10	133	47	6	0.7295918	0.1754386	0.9568345	0.0431655	0.8245614
30	0.58	8	134	49	5	0.7244898	0.1403509	0.9640288	0.0359712	0.8596491
31	0.6	8	135	49	4	0.7295918	0.1403509	0.971223	0.028777	0.8596491
32	0.62	7	135	50	4	0.7244898	0.122807	0.971223	0.028777	0.877193
33	0.64	7	136	50	3	0.7295918	0.122807	0.9784173	0.0215827	0.877193
34	0.66	6	137	51	2	0.7295918	0.1052632	0.9856115	0.0143885	0.8947368
35	0.68	5	137	52	2	0.7244898	0.0877193	0.9856115	0.0143885	0.9122807
36	0.7	2	138	55	1	0.7142857	0.0350877	0.9928058	0.0071942	0.9649123
37	0.72	1	138	56	1	0.7091837	0.0175439	0.9928058	0.0071942	0.9824561
38	0.74	0	138	57	1	0.7040816	0	0.9928058	0.0071942	1
39	0.76	0	139	57	0	0.7091837	0	1	0	1
40	0.78	0	139	57	0	0.7091837	0	1	0	1
41	0.8	0	139	57	0	0.7091837	0	1	0	1
42	0.82	0	139	57	0	0.7091837	0	1	0	1
43	0.84	0	139	57	0	0.7091837	0	1	0	1
44	0.86	0	139	57	0	0.7091837	0	1	0	1
45	0.88	0	139	57	0	0.7091837	0	1	0	1
46	0.9	0	139	57	0	0.7091837	0	1	0	1
47	0.92	0	139	57	0	0.7091837	0	1	0	1
48	0.94	0	139	57	0	0.7091837	0	1	0	1
49	0.96	0	139	57	0	0.7091837	0	1	0	1
50	0.98	0	139	57	0	0.7091837	0	1	0	1
51	1	0	139	57	0	0.7091837	0	1	0	1

The maximum of the percent correct is 0.755102. And when the probability level is at 0.48, the percent correct is at the highest level. So I would use the probability level of 0.48 as the cut-off. The percent of correct observations is 0.755102 in this model whereas the percent of correct observations is 0.7143 in the model with only age. So it is different. And using this model we can improve the correctness of prediction.

Discussion about the model

22.

Advantage:

- Predict the individual’s disease risk infected by the virus in a relatively high correctness using his/her age and living area sector.
- The prediction cost is low and the explanatory variables can be easily collected to predict their risk.
- The model is simple for use to predict since there is no interaction between the categorical variables and the quantitative variable and no high level categorical variable which would require more dummy variable.

Limitation:

- The correctness of prediction is not high enough to identify the exact disease risk person, which is only 75.5% percent of correct observations.
- Only consider two factors influencing the dengue fever. There may be some other factors that determine the disease status.
- Underdispersion occurs in the model.

Recommendation:

- Consider adding the saving account variable into the model and test its significance. It may account for the difference in disease status that we are seeing.
- Collect more information about the characteristics of the individuals such as their family member's status and their favorite foods. These factors may account for the disease.