(a) 18/20
(b) 14/15
(c) 20/20
(d) 13/15
(e) 10/10
(f) 10/10
total = 85/90

A Better Future for Infants: Risk Prediction Model for Infant Respiratory Disease

Lee Yen May

Introduction

Respiratory disease is a major health concern for infants worldwide. According to the World Health Organization (WHO), pneumonia alone accounts for 14% of all deaths of children under 5 years old, killing 740,180 children in 2019. Therefore, identifying the factors that increase the risk of respiratory disease in infants is critical for improving infant mortality rate. In this study, we aim to (1) explore the factors associated with the risk of respiratory disease in infants and (2) create a risk prediction model to estimate the probability of an infant developing respiratory disease within their first year of life.

To achieve these aims, we identify the significant predictors from variables of these three categories, namely infant characteristics, maternal characteristics and environmental factors. Data was collected from 18,422 infants across 40 countries. The independent variables investigated are shown below:

```
knitr::include_graphics("vartable.png")
```

Variable Name	Variable Description	Category	Variable Type	Data Type	Remarks
gender	gender	Infant Characteristics	categorical	string	F: Female M: Male
fed_method	feeding habits	Infant Characteristics	categorical	string	Supplement breastfeed only Bottle feed only Mix bottle and breastfeed
sleep_hour	average sleeping hours per day	Infant Characteristics	numerical	float	-
delivery_type	type of delivery	Infant Characteristics	categorical	integer	0: Natural birth 1: Ceasaran birth
weight	mother's weight	Maternal Characteristics	categorical	integer	0: Underweight 1: Normal weight
smoking	mother's smoking status	Maternal Characteristics	categorical	string	N: No Y: Yes
diabetes	mother's gestational diabetes	Maternal Characteristics	categorical	string	N: No Y: Yes
API	air pollution index of the country where the infant resides in	Environmental Factors	categorical	string	0-50: good 51-100: moderate 101-200: unhealthy 201-300: very unhealthy >300: hazardous

The response variable "status" follows a Bernoulli distribution $y_i \sim \text{Bernoulli}(P_i)$ with the following proba-

usually small letter to denote parameter

bility density function:

$$P(n) = \begin{cases} 1-p & \text{for } n = 0 \\ p & \text{for } n = 1 \end{cases}$$

suggest to use other letter. in stats, we keep "n" mostly to denote sample size

Each observation can only take on one of two possible outcomes, ie. presence or absence of the respiratory disease.

Exploratory Data Analysis

```
library(MASS)
library(knitr)
library(kableExtra)
```

```
## Warning in !is.null(rmarkdown::metadata$output) && rmarkdown::metadata$output
## %in% : 'length(x) = 3 > 1' in coercion to 'logical(1)'
```

```
train <- read.csv("MATH3046_train_data.csv")
str(train) # display the structure of the dataset
```

```
## 'data.frame':    18422 obs. of  9 variables:
##  $ status       : int  0 1 0 0 1 1 1 0 0 0 ...
##  $ gender       : chr  "M" "F" "F" "F" ...
##  $ fed_method   : chr  "Supplement" "Supplement" "Supplement" "Supplement" ...
##  $ sleep_hour   : num  11.6 12.5 12.5 12.5 12.5 ...
##  $ API          : chr  "101-200" "201-300" "201-300" "201-300" ...
##  $ diabetes     : chr  "Y" "Y" "Y" "Y" ...
##  $ delivery_type: int  0 1 1 1 1 0 0 0 0 0 ...
##  $ smoking      : chr  "N" "N" "N" "N" ...
##  $ weight       : int  1 1 1 0 0 1 1 1 1 1 ...
```

```
sum(is.na(train)) # check for NA values
```

```
## [1] 0
```

We determine the summary statistics to describe the relationship between the independent variables with the response variable status. By visual inspection, we have checked that all infants have non-smoker mothers hence variable "smoking" is not meaningful and will be dropped from the analysis.

For the categorical variables, we compare the frequency and the percentage within their respective levels for infant with and without disease.

```
cat_vars <- c("gender", "fed_method", "delivery_type", "weight", "diabetes", "API")
cat_table <- data.frame()

for (var in cat_vars) {
  for (lvl in unique(train[[var]])) {
    # subset the train data by the variable and its level
    sub_train <- subset(train, train[[var]] == lvl)

    for (i in c(0, 1)) {
```

```
      freq <- sum(sub_train$status == i) # calculate frequency
      perc <- freq/nrow(sub_train) # calculate percentage

      # add the results to the output dataframe cat_table
      cat_table <- rbind(cat_table, data.frame(
        variable = var,
        level = lvl,
        status = i,
        frequency = freq,
        percentage_within_lvl = perc
      ))
    }
  }
}

cat_table
```

```
##         variable                     level status frequency
## 1         gender                         M      0      4647
## 2         gender                         M      1      2301
## 3         gender                         F      0      7664
## 4         gender                         F      1      3810
## 5     fed_method                Supplement      0      8260
## 6     fed_method                Supplement      1      4156
## 7     fed_method           breastfeed only      0      3459
## 8     fed_method           breastfeed only      1      1662
## 9     fed_method          Bottle feed only      0       488
## 10    fed_method          Bottle feed only      1       251
## 11    fed_method Mix bottle and breastfeed      0       104
## 12    fed_method Mix bottle and breastfeed      1        42
## 13 delivery_type                           0      0     11084
## 14 delivery_type                           0      1      5417
## 15 delivery_type                           1      0      1227
## 16 delivery_type                           1      1       694
## 17        weight                           1      0      9928
## 18        weight                           1      1      4846
## 19        weight                           0      0      2383
## 20        weight                           0      1      1265
## 21      diabetes                           Y      0      7966
## 22      diabetes                           Y      1      4102
## 23      diabetes                           N      0      4345
## 24      diabetes                           N      1      2009
## 25           API                     101-200      0      8566
## 26           API                     101-200      1      4273
## 27           API                     201-300      0      1653
## 28           API                     201-300      1       879
## 29           API                        0-50      0      1092
## 30           API                        0-50      1       469
## 31           API                      51-100      0       715
## 32           API                      51-100      1       337
## 33           API                        >300      0       285
## 34           API                        >300      1       153
##    percentage_within_lvl
```

```
## 1              0.6688256
## 2              0.3311744
## 3              0.6679449
## 4              0.3320551
## 5              0.6652706
## 6              0.3347294
## 7              0.6754540
## 8              0.3245460
## 9              0.6603518
## 10             0.3396482
## 11             0.7123288
## 12             0.2876712
## 13             0.6717169
## 14             0.3282831
## 15             0.6387298
## 16             0.3612702
## 17             0.6719913
## 18             0.3280087
## 19             0.6532346
## 20             0.3467654
## 21             0.6600928
## 22             0.3399072
## 23             0.6838212
## 24             0.3161788
## 25             0.6671859
## 26             0.3328141
## 27             0.6528436
## 28             0.3471564
## 29             0.6995516
## 30             0.3004484
## 31             0.6796578
## 32             0.3203422
## 33             0.6506849
## 34             0.3493151
```

We should check the distribution of continuous variable, if it is not normally distributed, median and IQR would be a better summary stats

For the continuous variable "sleep_hour", we compare its means and standard deviations for infant with and without disease.

```
cont_table <- data.frame()

for (val in c(0, 1)) {
  # subset the train data for the current value of "status"
  sub_train <- train[train$status == val, ]
  sleep_hour_mean <- mean(sub_train$sleep_hour) # calculate mean
  sleep_hour_sd <- sd(sub_train$sleep_hour) # calculate sd

  # add the results to the output dataframe cont_table
  cont_table <- rbind(cont_table, data.frame(status = val,
                                             sleep_hour_mean = sleep_hour_mean,
                                             sleep_hour_sd = sleep_hour_sd))
}

cont_table
```

```
##   status sleep_hour_mean sleep_hour_sd
## 1      0        12.05771     0.4613962
## 2      1        12.07157     0.4721355
```

To test if there is a difference between the status of respiratory disease and each independent variable, we use chi-squared test for the categorical variables and t-test for the continuous variable.

For categorical variables, the hypothesis test is expressed as

$$H_0 : \beta_i = 0 H_1 : \beta_i \neq 0$$

where β_i represents the coefficient of each categorical variable within the dataset.

```r
# chi-squared test
sub_train <- subset(train, select = -c(sleep_hour, smoking))

for (col in colnames(sub_train)) {
  if (col != "status") {
    tbl <- table(sub_train[[col]], sub_train$status)
    chi_test <- chisq.test(tbl)
    print(paste("Chi-squared test for", col))
    print(chi_test)
  }
}
```

```
## [1] "Chi-squared test for gender"
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  tbl
## X-squared = 0.011427, df = 1, p-value = 0.9149
##
## [1] "Chi-squared test for fed_method"
##
##  Pearson's Chi-squared test
##
## data:  tbl
## X-squared = 3.1835, df = 3, p-value = 0.3642
##
## [1] "Chi-squared test for API"
##
##  Pearson's Chi-squared test
##
## data:  tbl
## X-squared = 10.903, df = 4, p-value = 0.02768
##
## [1] "Chi-squared test for diabetes"
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  tbl
## X-squared = 10.465, df = 1, p-value = 0.001217
##
## [1] "Chi-squared test for delivery_type"
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  tbl
## X-squared = 8.2979, df = 1, p-value = 0.003969
##
## [1] "Chi-squared test for weight"
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  tbl
## X-squared = 4.5588, df = 1, p-value = 0.03275
```

For the only continuous variable "sleep_hour", the hypothesis test is expressed as

$$H_0 : \bar{y} = \bar{x} H_1 : \bar{y} \neq \bar{x}$$

we are testing the population parameters, hence we should use "mu" here not bar(x) or bar(y)

where y represents the sample mean for "status" and x represents the sample mean for "sleep_hour".

```
# t-test
sub_train <- train[, c("status", "sleep_hour")]
t_test_result <- t.test(sleep_hour ~ status, data = sub_train)
t_test_result
```

```
##
##  Welch Two Sample t-test
##
## data:  sleep_hour by status
## t = -1.89, df = 11944, p-value = 0.05878
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -0.028232536  0.000514313
## sample estimates:
## mean in group 0 mean in group 1
##        12.05771        12.07157
```

The results for categorical variables are summarised below:

```
knitr::include_graphics("catgorical table.png")
```

	Categorical Variables	Levels	Frequency		% within level		chi-squared statistic	p-value	Reject H0?
			w/o disease: 0	w/ disease: 1	w/o disease: 0	w/ disease: 1			
Infant specifications	**gender**	F	7664	3810	66.79%	33.21%	0.011427	0.9149	NO
		M	4647	2301	66.88%	33.12%			
	fed_method	Supplement	8260	4156	66.53%	33.47%	3.1835	0.3642	NO
		breastfeed only	3459	1662	67.55%	32.45%			
		Bottle feed only	488	251	66.04%	33.96%			
		Mix bottle and breastfeed	104	42	71.23%	28.77%			
	delivery_type	0: Natural birth	11084	5417	67.17%	32.83%	8.2979	0.003969	YES
		1: Ceasaran birth	1227	694	63.87%	36.13%			
Maternal specifications	**weight**	0: Underweight	2383	1265	65.32%	34.68%	4.5588	0.03275	YES
		1: Normal weight	9928	4846	67.20%	32.80%			
	diabetes	N: No	4345	2009	68.38%	31.62%	10.465	0.001217	YES
		Y: Yes	7966	4102	66.01%	33.99%			
Environmental Factors	**API**	0-50: good	1092	469	69.96%	30.04%	10.903	0.02768	YES
		51-100: moderate	715	337	67.97%	32.03%			
		101-200: unhealthy	8566	4273	66.72%	33.28%			
		201-300: very unhealthy	1653	879	65.28%	34.72%			
		>300: hazardous	285	153	65.07%	34.93%			
		Total	12311	6111	66.83%	33.17%			

Based on summary statistics, we identify that we have 6200 more diseased infants than healthy infants. Therefore we analyse the relationship between "status" and each categorical variable using percentage split within each level rather than percentage split within the whole sample for a more meaningful interpretation. We observe the percentage split for each categorical variable as follows:

gender: The percentage of diseased male infants (33.12%) is approximately the same as the percentage of diseased female infants (33.21%), therefore gender most likely does not contribute to disease status.

comment on p-value

fed_method: The percentage of diseased infants is approximately the same among those who were given supplements (33.47%), only breast-fed (32.45%) and only bottle-fed (33.96%). However, infants who were both bottle-fed and breast-fed (28.77%) seem to have a lower percentage of developing disease compared to the three other feeding methods. This may imply a relationship with disease status but further investigation is needed.

delivery_type: The percentage of infants with disease is higher among those who were born via C-section (36.13%) compared to those who were born naturally (32.83%). This suggests that delivery type is potentially related to disease status.

weight: There is a higher percentage of diseased infants with underweight mothers (34.68%) compared to normal weight mothers (32.68%). This suggests that weight may have an effect on disease status.

diabetes: The percentage of infants with disease is slightly higher among those who have diabetic mothers (33.99%) compared to those who do not have diabetic mothers (31.62%). This suggests that the mother's diabetic status may have an effect on infant's disease status.

API: The percentage of infants with disease increases as air quality in their residing region worsens, with the highest percentage in the "hazardous" level (34.93%) and the lowest percentage in the "good" level (30.04%). This indicates that air quality is highly likely to contribute to infant's disease status.

In summary, the findings imply that fed_method, delivery_type, weight, diabetes and API may be related to the disease status, while gender may not be related. However, they should be analysed with statistical tests to verify our interpretations.

you have done chi-square test here?

Based on the p-values from the chi-squared tests above, we can reject the null hypothesis for delivery_type, weight, diabetes and API at a 5% significance level, indicating that these variables have a significant effect on disease status. On the other hand, we cannot reject the null hypothesis for gender and fed_method, indicating the otherwise. This result is consistent with our findings.

consistent with what findings?

the purpose of chi-square test is to test whether the observed differences in the percentages are significant. hence suggest to comment on the significance of variable in each corresponding paragraph above

The results for the continuous variable sleep_hour is summarised below:

```
knitr::include_graphics("cont table.png")
```

sleep_hour	mean	s.d.	t statistic	p-value	95% C.I	95% C.I.	Reject H0?
w/o disease: 0	12.05771	0.4613962	-1.89	0.05878	-0.02823	0.00051	NO
w/ disease: 1	12.07157	0.4721355					

We observe that its mean and variance are similar. This implies that sleep_hour does not affect disease status. This is also consistent with the result from t-test where we reject the null hypothesis at a 5% significance level.

nice!

Model Assumption

Based on the relationships identified above, we expect to use logistic regression for model building. However, the data needs to fulfill 5 assumptions of a logistic regression model.

1. Logistic regression requires the dependent variable to be binary or count. The response variable status is binary.
2. Logistic regression requires all independent variables to have independent observations. Since we have excluded smoking which has all values of "1", this assumption is fulfilled.

not sure i understand this

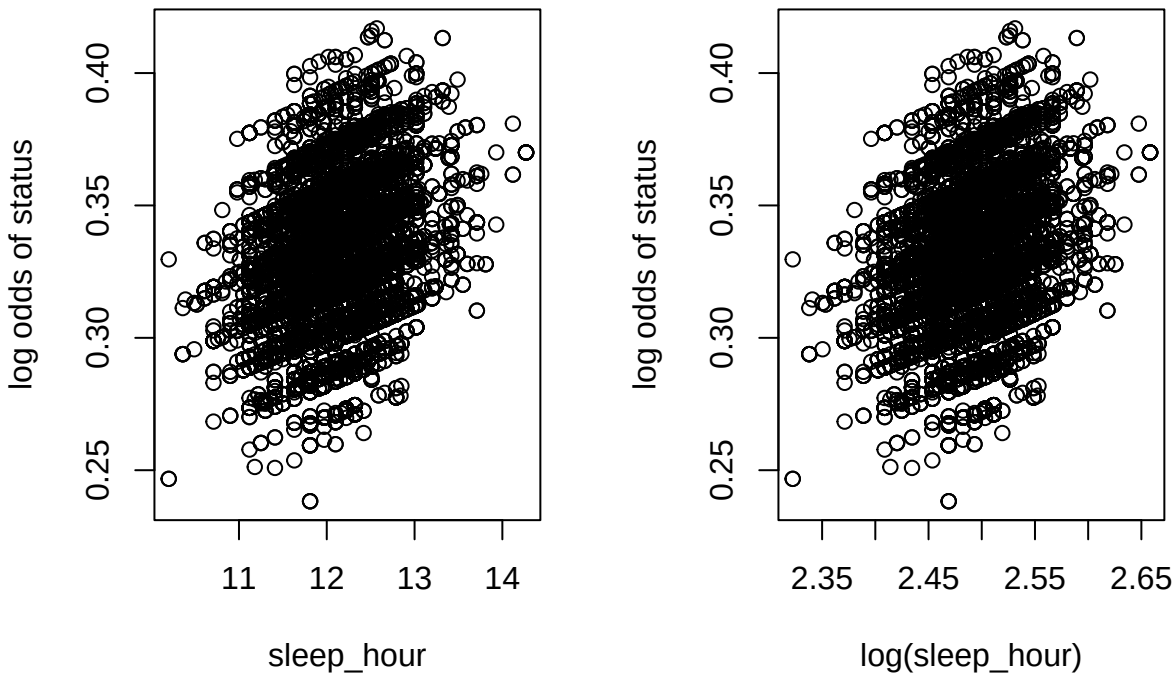
3. Logistic regression requires there to be little or no multicollinearity among the independent variables, ie. the independent variables should not be too highly correlated with each other. Since we only have one continuous variable, multicollinearity is not a concern.

not really, sometime continuous and categorical variable can be correlated too, but not within the scope of this coursework

4. Logistic regression assumes linearity of continuous independent variables and log odds (ie. logit) of the dependent variable. Therefore, we assess the linearity assumption between log(p/(1-p)) and average sleeping hour, where p is the logit derived from full model. It is not meaningful to check linearity assumption for categorical variables.

```
fullModel0 = glm(status ~ gender + fed_method + sleep_hour + delivery_type + weight + diabetes + API, fa
eta <- predict(fullModel0, newdata = train, type = "response")
p <- exp(eta) / (1 + exp(eta))

par(mfrow=c(1,2))
plot(train$sleep_hour,log(p/(1-p)),xlab="sleep_hour",ylab="log odds of status")
plot(log(train$sleep_hour),log(p/(1-p)),xlab="log(sleep_hour)",ylab="log odds of status")
```



On the left plot, we observe that sleep_hour has a weak positive linear relationship to the log odds of status. As seen on the right plot, taking the log of sleep_hour also produced similar outcome and did not improve the linear relationship further. Therefore, it does not satisfy the linearity assumption very well. To guarantee that we fulfill the linearity assumption, we categorise sleep_hour into three groups: - <11.6 hours - 11.6-12.9 hours - >12.9 hours

i think the plot doesn't suggest departure from linearity assumption, but do suggest the lack of relationship

```
# define the cut points for each group
cut_points <- c(-Inf, 11.6, 12.9, Inf)

# use the cut function to create a factor variable with three levels
sleep_hour_groups <- cut(train$sleep_hour, cut_points,
                         labels = c("<11.6 hours", "11.6-12.9 hours", ">12.9 hours"))
```

5. Logistic regression typically requires a large sample size. Our train dataset consists of 18,422 observations and our test dataset consists of 500 observations hence data size is large enough.

not relly an assumption

Model Building

By fulfilling all 5 model assumptions, we conclude that we can build our prediction model using logistic regression. We format the independent variables according to the requirement in glm:

```
status = train$status
gender = factor(train$gender)
fed_method = factor(train$fed_method)
API = factor(train$API, levels = c("0-50", "51-100", "101-200", "201-300", ">300"))
```

```
diabetes = factor(train$diabetes)
delivery_type = factor(train$delivery_type)
weight = factor(train$weight)
```

We choose logistic regression using the backwards selection method in the following steps: 1. Create full model.

```
fullModel = glm(status ~ gender + fed_method + sleep_hour_groups +
                delivery_type + weight + diabetes + API, family = "binomial"(link=logit))
summary(fullModel)
```

```
##
## Call:
## glm(formula = status ~ gender + fed_method + sleep_hour_groups +
##     delivery_type + weight + diabetes + API, family = binomial(link = logit))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0771  -0.9068  -0.8713   1.4533   1.6856
##
## Coefficients:
##                                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)                       -0.91032    0.11058  -8.232  < 2e-16 ***
## genderM                           -0.01898    0.03328  -0.570  0.56849
## fed_methodbreastfeed only         -0.09176    0.08370  -1.096  0.27297
## fed_methodMix bottle and breastfeed -0.23156    0.19923  -1.162  0.24513
## fed_methodSupplement              -0.01085    0.08038  -0.135  0.89265
## sleep_hour_groups11.6-12.9 hours   0.09907    0.04914   2.016  0.04380 *
## sleep_hour_groups>12.9 hours       0.29335    0.09442   3.107  0.00189 **
## delivery_type1                     0.14230    0.05096   2.793  0.00523 **
## weight1                           -0.08138    0.03901  -2.086  0.03697 *
## diabetesY                          0.09535    0.03331   2.862  0.00420 **
## API51-100                          0.09821    0.08630   1.138  0.25509
## API101-200                         0.14966    0.05853   2.557  0.01056 *
## API201-300                         0.21028    0.06943   3.028  0.00246 **
## API>300                            0.22110    0.11510   1.921  0.05474 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 23410  on 18421  degrees of freedom
## Residual deviance: 23363  on 18408  degrees of freedom
## AIC: 23391
##
## Number of Fisher Scoring iterations: 4
```

2. Perform backwards stepwise regression. The independent variable "gender" is dropped from the model.

```
stepwiseModel <- step(fullModel, direction = "backward")
```

```
## Start:  AIC=23390.78
```

```
## status ~ gender + fed_method + sleep_hour_groups + delivery_type +
##     weight + diabetes + API
##
##                      Df Deviance   AIC
## - gender              1    23363 23389
## <none>                     23363 23391
## - fed_method          3    23369 23391
## - weight              1    23367 23393
## - API                 4    23373 23393
## - delivery_type       1    23371 23397
## - sleep_hour_groups   2    23373 23397
## - diabetes            1    23371 23397
##
## Step:  AIC=23389.1
## status ~ fed_method + sleep_hour_groups + delivery_type + weight +
##     diabetes + API
##
##                      Df Deviance   AIC
## <none>                     23363 23389
## - fed_method          3    23369 23389
## - weight              1    23368 23392
## - API                 4    23374 23392
## - sleep_hour_groups   2    23373 23395
## - delivery_type       1    23371 23395
## - diabetes            1    23372 23396
```

```
summary(stepwiseModel)
```

```
##
## Call:
## glm(formula = status ~ fed_method + sleep_hour_groups + delivery_type +
##     weight + diabetes + API, family = binomial(link = logit))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0727  -0.9103  -0.8749   1.4495   1.6806
##
## Coefficients:
##                                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)                       -0.91297    0.11048  -8.264  < 2e-16 ***
## fed_methodbreastfeed only         -0.09047    0.08367  -1.081  0.27958
## fed_methodMix bottle and breastfeed -0.23281    0.19921  -1.169  0.24255
## fed_methodSupplement              -0.01100    0.08038  -0.137  0.89110
## sleep_hour_groups11.6-12.9 hours   0.09440    0.04846   1.948  0.05139 .
## sleep_hour_groups>12.9 hours       0.28552    0.09341   3.057  0.00224 **
## delivery_type1                     0.14286    0.05095   2.804  0.00504 **
## weight1                           -0.08156    0.03901  -2.091  0.03656 *
## diabetesY                          0.09622    0.03328   2.892  0.00383 **
## API51-100                          0.09929    0.08628   1.151  0.24980
## API101-200                         0.14807    0.05846   2.533  0.01132 *
## API201-300                         0.20934    0.06941   3.016  0.00256 **
## API>300                            0.22689    0.11466   1.979  0.04783 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 23410  on 18421  degrees of freedom
## Residual deviance: 23363  on 18409  degrees of freedom
## AIC: 23389
##
## Number of Fisher Scoring iterations: 4
```

3. Based on stepwiseModel, we observe that the p-value for all levels of "fed_method" is greater than 0.05. Therefore, we decide to drop "fed_method".

```
fittedModel = glm(status ~ sleep_hour_groups + delivery_type + weight +
                  diabetes + API, family = "binomial"(link=logit))
summary(fittedModel)
```

```
##
## Call:
## glm(formula = status ~ sleep_hour_groups + delivery_type + weight +
##     diabetes + API, family = binomial(link = logit))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0752  -0.9019  -0.8667   1.4524   1.6276
##
## Coefficients:
##                                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)                     -0.93412    0.07890 -11.840  < 2e-16 ***
## sleep_hour_groups11.6-12.9 hours  0.08399    0.04816   1.744  0.08117 .
## sleep_hour_groups>12.9 hours      0.25328    0.09211   2.750  0.00596 **
## delivery_type1                    0.13083    0.05066   2.582  0.00981 **
## weight1                          -0.08117    0.03900  -2.081  0.03743 *
## diabetesY                         0.09601    0.03325   2.887  0.00389 **
## API51-100                         0.09440    0.08623   1.095  0.27362
## API101-200                        0.14579    0.05843   2.495  0.01259 *
## API201-300                        0.20874    0.06934   3.010  0.00261 **
## API>300                           0.22412    0.11459   1.956  0.05048 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 23410  on 18421  degrees of freedom
## Residual deviance: 23369  on 18412  degrees of freedom
## AIC: 23389
##
## Number of Fisher Scoring iterations: 4
```

We decide to select fittedModel as our final prediction model due to the following reasons: 1. The AIC for fittedModel and stepwiseModel are both 23389, indicating that fittedModel fits the data as well as stepwiseModel. 2. Since the residual deviance for fittedModel is 23369 and for stepwiseModel is 23363, the residual deviance is only slightly higher which is negligible. With one less predictor variable, we conclude that fittedModel is a better choice than stepwiseModel as it is easier to interpret and apply in practice.

A combined table of odds ratio, 95% confidence interval and p-value is shown as below:

```
OR <- exp(coef(fittedModel)[-1]) # odds ratio
CI <- exp(confint(fittedModel, level = 0.95))[-1,] # 95% confidence intervals of odds ratio
```

```
## Waiting for profiling to be done...
```

```
pval <- coef(summary(fittedModel))[-1, 'Pr(>|z|)'] # p-value

table <- cbind(OR, CI, pval)
table
```

```
##                                       OR     2.5 %    97.5 %        pval
## sleep_hour_groups11.6-12.9 hours 1.087623 0.9900453 1.1958035 0.081169365
## sleep_hour_groups>12.9 hours     1.288238 1.0746862 1.5421985 0.005964196
## delivery_type1                   1.139771 1.0316632 1.2583283 0.009809316
## weight1                          0.922041 0.8543295 0.9954697 0.037430794
## diabetesY                        1.100767 1.0313987 1.1750007 0.003886938
## API51-100                        1.098999 0.9278245 1.3010480 0.273617799
## API101-200                       1.156948 1.0324457 1.2982523 0.012593026
## API201-300                       1.232129 1.0759315 1.4120493 0.002609194
## API>300                          1.251218 0.9981044 1.5644254 0.050477251
```

Though the p-values of API51-100 and API>300 are greater than 0.05 when evaluated at a 5% significance level, we still retain them in our model because the other 2 API levels are statistically significant and API>300 is only slightly greater than 0.05. Similarly, we retain sleep_hour_groups11.6-12.9 hours because it is only slightly larger than 0.05 and sleep_hour_groups>12.9 hours is statistically significant.

The odds ratio of sleep_hour_groups>12.9 hours, sleep_hour_groups11.6-12.9 hours, API>300, API201-300, API101-200, API51-100, delivery_type1, diabetesY are greater than 1 and the odds ratio of weight1 is lower than 1.

An odds ratio larger than 1 indicates a positive association to the response variable status. After controlling for the other variables, we interpret the following effects:

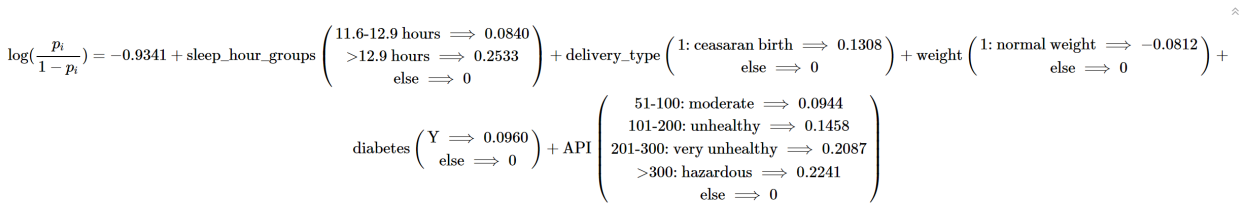
- An infant who sleeps more than 12.9 hours per day increases their odds of developing respiratory disease by 28.82% whereas an infant who sleeps for 11.6 to 12.9 hours per day increases their odds by 8.76%.
- An infant who is exposed to hazardous, very unhealthy, unhealthy and moderate air pollution has an increased odds of developing respiratory disease by 25.12%, 23.21%, 15.69% and 9.90% respectively. It also implies that the odds of developing disease increases as air quality deteriorates in the region where the infant resides.
- An infant who is delivered via C-section has an increased odds of developing disease by 13.98% compared to an infant delivered naturally.
- An infant who has a diabetic mother has an increased odds of developing disease by 10.08% compared to a non-diabetic mother.

On the other hand, an odds ratio smaller than 1 implies a negative association to the response variable. This means that a mother possessing normal weight decreases her infant's odds of developing disease by 7.80%.

By the 95% confidence interval of the odds ratio of all variables included in our model, we can conclude that interpretations of odds above are statistically significant because all confidence intervals do not contain 0, hence we reject the null hypothesis.

We use these variables to form our fitted model as follows:

```
knitr::include_graphics("model.png")
```



Hence,

```
knitr::include_graphics("eqeta.png")
```

$$\eta_i = -0.9341 + 0.0840 \times \text{sleep\_hour\_groups11.6-12.9 hours} + 0.2533 \times \text{sleep\_hour\_groups>12.9 hours} + 0.1308 \times \text{delivery\_type1} - 0.0812 \times \text{weight1} + 0.0960 \times \text{diabetesY} + 0.0944 \times \text{API51-100} + 0.1458 \times \text{API101-200} + 0.2087 \times \text{API201-300} + 0.2241 \times \text{API>300}$$

and the probability of an infant developing respiratory disease within their first year of life is expressed as $p_i = \frac{\exp(\eta_i)}{1-\exp(\eta_i)}$.

Model Testing

We now test our prediction model with an additional 500 observations provided by the scientist.

```
test <- read.csv("MATH3046_test_data.csv")
str(test) # display the structure of the dataset
```

```
## 'data.frame':    500 obs. of  9 variables:
##  $ status       : int  0 0 1 1 0 0 0 0 1 0 ...
##  $ gender       : chr  "F" "M" "M" "F" ...
##  $ fed_method   : chr  "breastfeed only" "Supplement" "Bottle feed only" "Supplement" ...
##  $ sleep_hour   : num  11.3 11.7 12 11.6 11.3 ...
##  $ API          : chr  ">300" "201-300" "201-300" "101-200" ...
##  $ diabetes     : chr  "Y" "N" "N" "N" ...
##  $ delivery_type: int  0 0 0 0 0 0 0 1 0 0 ...
##  $ smoking      : chr  "N" "N" "N" "N" ...
##  $ weight       : int  1 1 1 1 1 0 1 1 1 1 ...
```

```
sum(is.na(test)) # check for NA values
```

```
## [1] 0
```

```
status = test$status
gender = factor(test$gender)
fed_method = factor(test$fed_method)
API = factor(test$API, levels = c("0-50", "51-100", "101-200", "201-300", ">300"))
diabetes = factor(test$diabetes)
delivery_type = factor(test$delivery_type)
weight = factor(test$weight)

cut_points <- c(-Inf, 11.6, 12.9, Inf)
sleep_hour_groups <- cut(test$sleep_hour, cut_points,
                         labels = c("<11.6 hours", "11.6-12.9 hours", ">12.9 hours"))
```

We calculate the predicted probabilities of the 500 infants developing respiratory disease within their first year of life.

```
test <- data.frame(status,  gender, fed_method, sleep_hour_groups,
                   API, diabetes,   delivery_type,  weight)
prediction <- predict(fittedModel, newdata = test)
probabilities <- exp(prediction) / (1 + exp(prediction))
```

How was this done? base on what criteria you decided 0.36 is an appropriate cutoff to evaluate accuracy?

To test the model accuracy, we selected a decision threshold of 0.36 by trial and error and generated its confusion matrix. The confusion matrix allows us to see how many of the actual positives were correctly identified as positives (true positives) and how many were incorrectly identified as negatives (false negatives). It also allows us to see how many of the actual negatives were correctly identified as negatives (true negatives) and how many were incorrectly identified as positives (false positives).

```
# Convert probabilities to binary class labels
probabilities_labels <- ifelse(probabilities > 0.36, 1, 0)

# Create a confusion matrix
conf_matrix <- table(test$status, probabilities_labels)

conf_matrix
```

```
##    probabilities_labels
##       0   1
##   0 303  30
##   1 132  35
```

```
# Calculate evaluation metrics
accuracy <- sum(diag(conf_matrix)) / sum(conf_matrix)
precision <- conf_matrix[2,2] / sum(conf_matrix[,2])
recall <- conf_matrix[2,2] / sum(conf_matrix[2,])
f1_score <- 2 * precision * recall / (precision + recall)

# Print the confusion matrix and evaluation metrics
cat("Confusion Matrix:\n", conf_matrix, "\n")
```

```
## Confusion Matrix:
##  303 132 30 35
```

```
cat("Accuracy:", accuracy, "\n")
```

```
## Accuracy: 0.676
```

```
cat("Precision:", precision, "\n")
```

```
## Precision: 0.5384615
```

```
cat("Recall:", recall, "\n")
```

```
## Recall: 0.2095808
```

```
cat("F1-Score:", f1_score, "\n")
```

```
## F1-Score: 0.3017241
```

Based on our confusion matrix output, we observe that: - There are 303 healthy infants correctly predicted as no disease. - There are 35 unhealthy infants correctly predicted as developing disease. - There are 30 healthy infants incorrectly predicted as developing disease. - There are 132 unhealthy infants incorrectly predicted as no disease.

Based on the performance metrics, we interpret that: - Accuracy: Our model predicted 67.6% of the cases correctly. - Precision: When our model predicted a positive case, it was correct 53.8% of the time. - Recall: Our model correctly identified 21.0% of the actual positive cases - F1-Score: An F1-Score of 0.302 implies that our model has a moderately performing precision and recall.

Summary

This report investigated the relationship between infant disease status and eight predictors from three categories: infant characteristics, maternal characteristics and environmental factors. For infant characteristics, the variables include infant gender, feeding habits, average daily sleeping hour and delivery type. For maternal characteristics, we have data on the mother's weight, smoking status and diabetic status. The air pollution index is the only environmental factor in our analysis.

Based on our data collected from 18,422 infants across 40 countries, we have explored the factors associated with the disease status using three summary statistics: percentage frequency within each value of the categorical variable, mean and standard deviation. We used percentage frequency to summarise the characteristics of our categorical variables. Our findings revealed that the split between diseased female infants and diseased male infants is approximately the same. We also observed this outcome in feeding methods. There is a higher frequency percentage of diseased infants who were delivered through c-sections and had diabetic and underweight mothers. The percentage frequency for diseased infants also showed an increasing trend with the increased severity in the air pollution index. On the other hand, we use mean and standard deviation to inform us of average daily sleeping hours' characteristics, which is a continuous variable. The means and standard deviations appeared similar for diseased and non-diseased infants. With our expectation that gender, feeding methods, and average daily sleeping hour have no relationship compared to the diseased status, we tested the significance of the difference between disease status and our predictors with chi-squared and t-tests. Our findings implied that infant gender, feeding method and average daily sleeping hours did not show significance to disease status, hence are consistent with our expectation formed from the summary statistics.

Subsequently, we built a prediction model to estimate the probability of an infant developing respiratory disease within their first year of life using logistic regression. The assumptions of the logistic regression model

led to our decision to eliminate the mother's smoking status and categorise the infant's average sleeping hours into three groups. We executed these data processing steps because all infants' mothers are non-smokers, and the categorising enabled the linearity assumption to be fulfilled. We removed infant gender through backward stepwise regression. From analysing the p-values of predictor estimates evaluated at a 5% significance level, we found variable feeding methods insignificant and decided to remove it from our analysis. The prediction model has an AIC score of 23,389, as low as the backward stepwise regression model, and residual deviance of 23369, slightly higher than the backward stepwise regression model. AIC represents how well the model fits the data, and residual deviance compares the observed data with the predicted data; therefore, a low AIC score and residual deviance indicate a better fit. Since the increase is negligible, we proceed with the simplified model with one variable less. In conclusion, we identified the best subset of predictors with two variables from infant characteristics - average sleeping hours by group, and delivery type, two variables from maternal characteristics - weight and diabetic status, and the only environmental factor - air pollution index.

The analysis of p-value on the final model evaluated at a 5% significance level verified that our variable selection method is appropriate. The analysis of odds ratio showed that the average sleeping hours by group and air pollution index are the most dominant predictors. Our results showed that infants sleeping more than 12.9 hours on average per day increases their odds of developing respiratory diseases by 29%. In contrast, an average of 11.6 to 12.9 hours leads to a 9% increase in odds. Furthermore, the higher the air pollution index the infant is exposed to, the higher their odds of developing the disease. This relationship is demonstrated by a 25%, 23%, 16% and 10% increase in odds for a hazardous, very unhealthy, unhealthy and moderate air pollution index, respectively. The odds of developing disease for an infant delivered via c-section is 14% higher than an infant delivered naturally. The odds of developing disease for an infant with a diabetic mother is 9% higher than an infant with a healthy mother. Finally, an infant with a mother possessing normal weight can offset the odds of developing disease by 8%. In terms of model performance, we set a decision threshold of 0.36 on the predicted probabilities, which means that an infant with a predicted probability greater than 0.36 is classified as diseased, and a probability of 0.36 or lower indicates non-diseased. We tested the model on an additional 500 observations and analysed its confusion matrix to visualise the distribution of true positives, false negatives, false positives and true negatives cases. As a result, our prediction model predicted 67.6% of the cases accurately, classified 53.8% of the positive cases precisely, identified 21.0% of actual positive cases correctly and has an F1-Score of 0.302, indicating a moderately performing model.

The moderate performance can be explained by the lack of variables in our analysis, which is a limitation to our model's predictive power. While we investigated eight predictors, only five were significant to our disease status. Other factors could affect disease status but were not accounted for. For instance, the scientist can extend the variables categorised as infant characteristics with variables such as genetics, Family history of respiratory disease, immunisation status, allergies and birth weight. This is because if the family has history of respiratory diseases, the infant is likely to be inherited with the relevant genomes. We can also suspect that infant who are not vaccinated, have allergies and low birthweight are more vulnerable to developing respiratory diseases. For maternal characteristics, the scientist should also collect data of infants with smoker mothers since our dataset only includes infants with non-smoker mothers, which hinders us from utilising the variable. The scientist can also expand the category to parental characteristics to consider both parents and evaluate factors such as household income, father's education level and mother's education level. This is because we can suspect that household with lower income, and parents with lower awareness on health issues are likely to have lesser access to proper healthcare facilities and infants from these backgrounds are less likely to receive proper diagnosis or check up. For environmental factors, having only API as a factor is limiting the model, the variables could be expanded to country, climate and housing conditions. This is because the country of residence can affect the level of air pollution, access to healthcare, and other environmental factors that can contribute to respiratory disease. Climate conditions such as temperature, humidity, and seasonal changes can also affect respiratory health. Additionally, housing conditions such as mold, dust, and poor ventilation can increase the risk of respiratory problems.

By considering these factors, we can develop a more comprehensive understanding of the potential risk factors for infant respiratory disease and improve our ability to predict and prevent it. However, since these variables were not explored, we were unable to investigate them and improve the model accuracy further.

Advice for the Scientist

Regarding scientist's interest on knowing the probability of an infant developing respiratory disease with their first month of live, the provided model is not appropriate for answering the question because the model was built to predict the probability of an infant developing respiratory disease within their first year of life only.

Though some of the variables here could also be adopted, such as all the maternal characteristics and the environmental factors, the scientist would need to collect infant characteristics data specifically related to the first month of an infant's life, such as the infant's birth weight, length of gestation and gestational age. The scientist could also extend on the variables for maternal characteristics such as any relevant medical history of the mother or the infant. Based on this data, the scientist can build a new logistic regression model to predict the probability of an infant developing respiratory disease within their first month of life. However, the accuracy and reliability of such a model would need to be retested again, using similar techniques above such confusion matrix.

The potential model could look like:

```
knitr::include_graphics("eqpo.png")
```

$$\eta_i = \beta_0 + \beta_1 \times \text{birth\_weight} + \beta_2 \times \text{gestation\_length} + \beta_3 \times \text{gestational\_age} + \beta_4 \times \text{weight} + \beta_5 \times \text{diabetes} + 0.0944 \times \text{API} + 0.1458 \times \text{delivery\_type} + 0.2087 \times \text{smoking} + 0.2241 \times \text{diabetes}$$