Complete/quasi-complete separation in Logistic Regression

When running a binary logistic regression we estimate parameters for a specified model based on the sample data that has been collected. Most of the time, we use what is called Maximum Likelihood Estimation. However, based on specifics within our data, sometimes these estimation methods fail.

In this example we are going to talk about a situation that can often occur when running logistic regression called complete separation or quasi-complete separation.

We are going to use the dataset "sepsis data.xlsx". Firstly, we need to load the dataset.

```
library(readxl)
sepsis <- read_excel("sepsis data.xlsx")
View(sepsis)
```

Suppose we have a sample of 52 newborns tested for sepsis. There are four variables in our dataset:

- infected(0=no/1=yes) = infected with sepsis or not
- GA = gestational age
- IL6 = Levels of Interleukin 6
- blood_culture(0=negative/1=positive) = whether the result of the blood
- culture was positive/negative

We are going to run a simple logistic regression model using variable "infected" as the binary dependent variable and variable "blood_culture" as the independent variable.

Let's start

```
# Fit the model
model.q <- glm(infected ~ blood_culture, data = sepsis, family = binomial())
#Summarise the model
summary(model.q)

##
## Call:
## glm(formula = infected ~ blood_culture, family = binomial(),
##     data = sepsis)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.84460  -0.84460   0.00008   0.00008   1.55176
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -0.8473     0.3984  -2.127   0.0334 *
## blood_culture   20.4134  2292.7633   0.009   0.9929
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 70.152  on 51  degrees of freedom
## Residual deviance: 36.652  on 50  degrees of freedom
## AIC: 40.652
##
## Number of Fisher Scoring iterations: 18
```

```
#To obtain the OR and the 95% CI we need to exponentiate the estimate
cbind(exp(coef(model.q)), exp(confint(model.q)))
```

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

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: algorithm did not converge

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

##                                   2.5 %        97.5 %
## (Intercept)   4.285714e-01 1.863789e-01  9.073728e-01
## blood_culture 7.335207e+08 3.818474e-41 1.027733e+300
```

So, we run the model, model summary looks somewhat strange (we notice that the Std.Error for `blood_culture` is quite big) and then for the OR and 95%CI we notice a large number of warnings and we end up with a very large OR for `blood_culture` (7.335 x 10^8) and an even larger 95%CI (3.81 x 10^{-41}, 1.02 x 10^{300}). All the above warnings indicate that the algorithm has not converged.

When we have such a result, the first thing we are going to check is whether there is complete/quasi-complete separation in the dataset.

Complete separation or quasi-complete separation occurs when a linear combination of the predictors yields a perfect prediction of the response variable for all or the most values of the predictors. So. when we have an unexpectedly large OR with an infinite 95%CI we suspect that there might be a complete or quasi-complete separation in the data.

Please note: it is wrong to report such OR and 95%CI!.

The simplest way to check whether such a separation exists in the data is by using a two-way table between the dependent and the independent variable. In this example, we have

```
table(sepsis$infected, sepsis$blood_culture,dnn = c("Infected", "Blood Cultur
e"))
```

```
##          Blood Culture
## Infected  0  1
##        0 21  0
##        1  9 22
```

Here we notice that non-infected patients (Infected=0) have no positive blood cultures for sepsis. This "perfect prediction" of the outcome variable is what causes the estimates, and thus our model, to fail.

Another way to examine whether there is a separation in our data is by using the `detectseparation` package.

```
library(detectseparation)
model.sep <- glm(infected ~ blood_culture, data = sepsis, family = binomial()
, method = "detect_separation")
model.sep
```

```
## Implementation: ROI | Solver: lpsolve
## Separation: TRUE
## Existence of maximum likelihood estimates
##   (Intercept) blood_culture
##             0           Inf
## 0: finite value, Inf: infinity, -Inf: -infinity
```

We use this package only to identify whether a separation exists and in which variable is detected. It does not need to estimate the model to detect separation. So, here we notice that `Separation: TRUE` so there is separation in our data and that there is an `Inf` under the variable `blood culture` meaning that infinite estimates were identified for this variable.

Often, separation occurs when the data set is too small to observe events with low probabilities. In addition, the more predictors are in the model, the more likely separation is to occur because the individual groups in the data have smaller sample sizes. Separation occurs when there is a category or range of a predictor with only one value of the response. We need diversity, or variation among the response to estimate the model.

We usually cannot do much to fix the problem. Some remedies (depending on the independent variable) are:

- Increasing the amount of observations.
- Consider what the separation means. Complete separation and quasi-complete separation can indicate important relationships.
- Consider an alternative model. – this applies when a complete separation or a quasi-complete separation is detected in a multivariate model.
- Check to see whether you can combine categories in problematic variables.

For this particular example we can not do anything apart from indicating the type of separation and the relationship between the dependent and independent variable. (all the patients with no sepsis had negative blood culture).