Final Project

F.R.Castilla, O.Contrera

2022-12-11

Table of Contents

Abstract

Lupus is a chronic, complex autoimmune disease that can affect the joints, skin, brain, lungs, kidneys and blood vessels, causing widespread inflammation and tissue damage in the affected organs.

In order to study whether early detection of this disease is relevant to its survival, we used 87 people followed up for this study.

for this study we used 87 followed for 15+ years there have been 35 deaths, the time the disease has appeared and the duration which is the time it took to be detected.

Description data

the data has been extracted from http://lib.stat.cmu.edu/datasets/lupus

TIME(int) time during which the disease has occurred STATUS(int) 0 alive 1 dead DURATION(num) time during which the viropsy was taken LOG.1.DURATION.(num) log(1+duration)

```
data<-data.frame(read.table("~/Statistic/Project/lupus.txt",header = T))
dim(data)
```

```
## [1] 87  4
```

```
attach(data)
```

first we see that our table has a dimension of 87x4

our objective would be to see if the STATUS which is what tells us if the subject is alive (0) or dead (1).

as a first step we see how many of the subjects in the study have died or are alive.

```
table(STATUS)
```

```
## STATUS
##  0  1
## 52 35
```

we can see that at the end of the study there have been a total of 35 deaths out of 87 over the 15+ years of the study.

Technique

as our objective is to be able to see which predictors are the most important for predicting the STATUS and this is a binomial variable (0, 1) the model that can be most closely approximated is the logistic regression model.

Analysis

```
mean(data$DURATION[data$STATUS==0])
```

```
## [1] 9.228846
```

```
mean(data$DURATION[data$STATUS==1])
```

```
## [1] 12.34
```

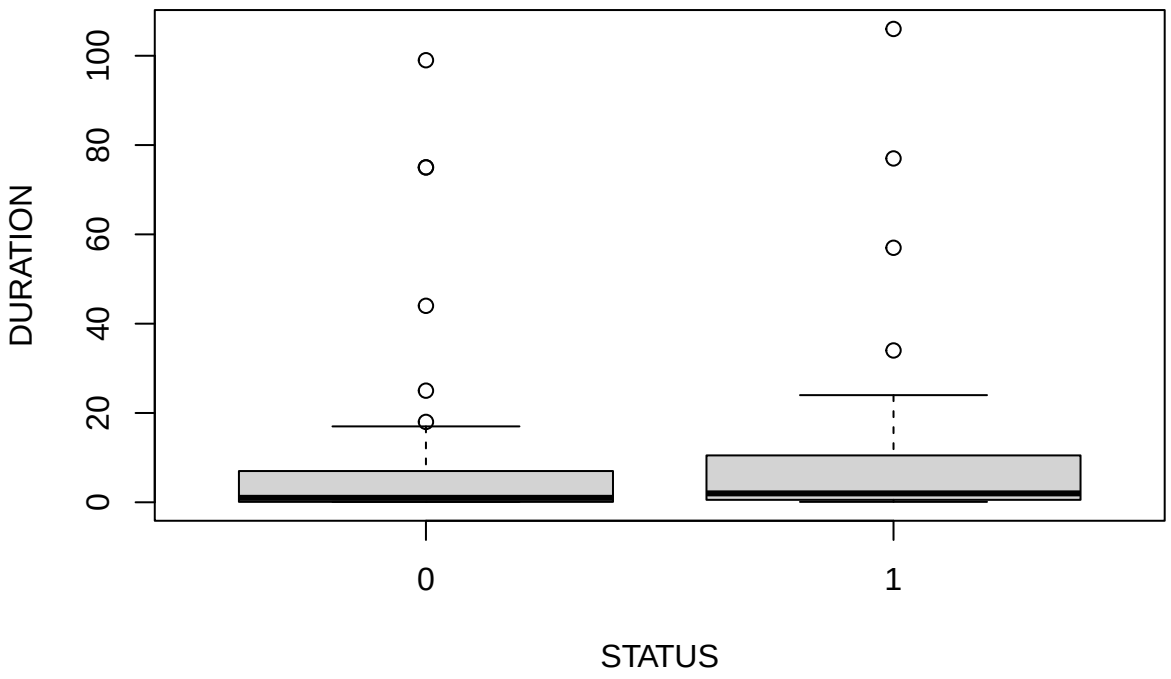
```
mean(data$DURATION)
```

```
## [1] 10.48046
```

with a mean of 10.48 being the time the subjects had the disease until they were biopsied and started treatment we can see that the mean for the live subjects is 9.23 and the mean for the deceased is 12.34 and

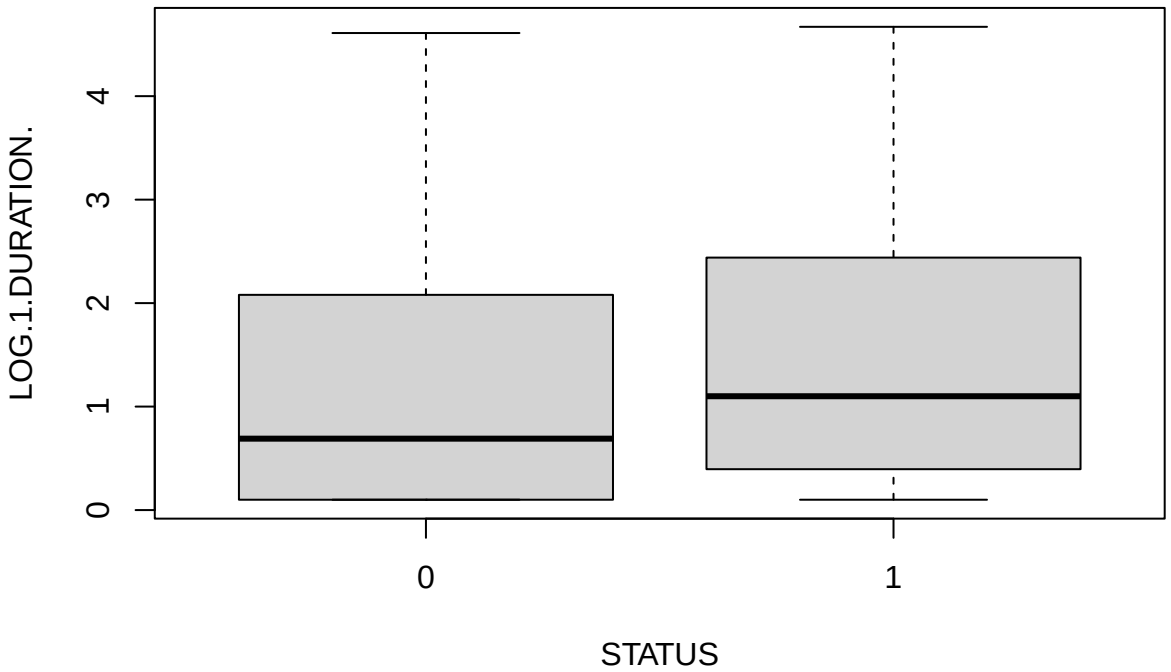
we can begin to see that there is a difference between the time the disease was started and the time it was treated.

```
boxplot(DURATION~STATUS)
```



in this boxplot it cannot be distinguished well so in order to work with a more comfortable scale we will work with the logarithm+1.

```
boxplot(LOG.1.DURATION.~STATUS)
```



where you can already see some difference but still not enough to be able to see if there is a relationship between the time that passed before the treatment started and whether it survives or not we create our model, also as we are dealing with a data that is binary live/dead I will use a STATUS~DURATION logistic model.

```
model1.1<-glm(STATUS~DURATION,family = binomial(link = 'logit'), trace=TRUE)
```

```
## Deviance = 116.9665 Iterations - 1
## Deviance = 116.8252 Iterations - 2
## Deviance = 116.8251 Iterations - 3
## Deviance = 116.8251 Iterations - 4
```

```
summary(model1.1)
```

```
##
## Call:
## glm(formula = STATUS ~ DURATION, family = binomial(link = "logit"),
##     trace = TRUE)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.2624  -0.9942  -0.9866   1.3589   1.3810
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.467618   0.244883  -1.910   0.0562 .
## DURATION     0.006719   0.010145   0.662   0.5078
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 117.26  on 86  degrees of freedom
## Residual deviance: 116.83  on 85  degrees of freedom
## AIC: 120.83
##
## Number of Fisher Scoring iterations: 4
```

```
anova(model1.1,test="Chisq")
```

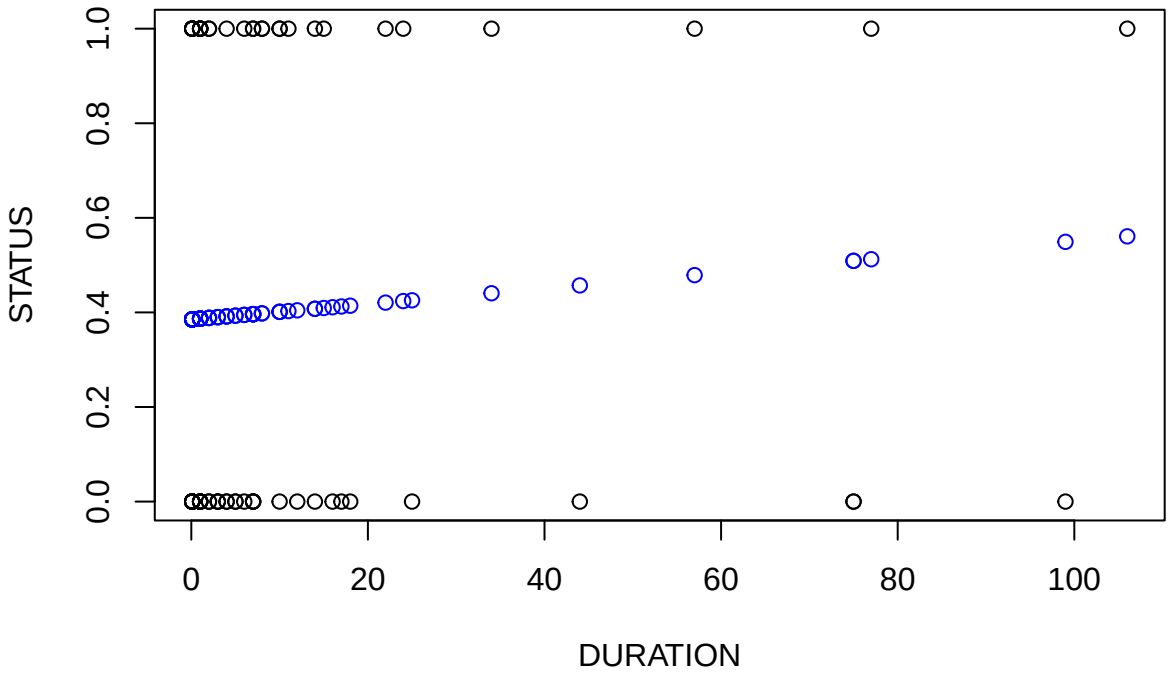
```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: STATUS
##
## Terms added sequentially (first to last)
##
##
##          Df Deviance Resid. Df Resid. Dev Pr(>Chi)
## NULL                        86     117.26
## DURATION  1  0.43916        85     116.83   0.5075
```

```
dchisq(0.43916 ,1)
```

```
## [1] 0.4833217
```

with the summary of the model we can already start to see that the DURATION variable does not have an influence.

```
plot(STATUS~DURATION)+
points(DURATION,fitted(model1.1),pch=1,col="blue")
```




```
## integer(0)
```

so we are going to look if there is another model that can explain the deaths so we are going to try STATUS~DURATION+TIME and STATUS~TIME,

```
model1.2<-glm(STATUS~DURATION+TIME,family = binomial(link = 'logit'), trace=TRUE)
```

```
## Deviance = 80.91537 Iterations - 1
## Deviance = 80.21662 Iterations - 2
## Deviance = 80.20605 Iterations - 3
## Deviance = 80.20604 Iterations - 4
## Deviance = 80.20604 Iterations - 5
```

```
summary(model1.2)
```

```
##
## Call:
## glm(formula = STATUS ~ DURATION + TIME, family = binomial(link = "logit"),
##     trace = TRUE)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0159  -0.7175  -0.3359   0.6207   2.1395
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.753482   0.734113   3.751 0.000176 ***
## DURATION     0.001742   0.017815   0.098 0.922116
## TIME        -0.023631   0.005051  -4.678  2.9e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
```

```
##     Null deviance: 117.264  on 86  degrees of freedom
## Residual deviance:  80.206  on 84  degrees of freedom
## AIC: 86.206
##
## Number of Fisher Scoring iterations: 5
```

```
anova(model1.2,test="Chisq")
```

```
## Deviance = 116.9665 Iterations - 1
## Deviance = 116.8252 Iterations - 2
## Deviance = 116.8251 Iterations - 3
## Deviance = 116.8251 Iterations - 4
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: STATUS
##
## Terms added sequentially (first to last)
##
##
##          Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                        86    117.264
## DURATION  1    0.439        85    116.825    0.5075
## TIME      1   36.619        84     80.206 1.436e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
dchisq(36.619 ,1)
```

```
## [1] 7.367871e-10
```

```
dchisq(0.439,1)
```

```
## [1] 0.4834484
```

```
model1.3<-glm(STATUS~TIME,family = binomial(link = 'logit'), trace=TRUE)
```

```
## Deviance = 80.9429 Iterations - 1
## Deviance = 80.22585 Iterations - 2
## Deviance = 80.21562 Iterations - 3
## Deviance = 80.21562 Iterations - 4
## Deviance = 80.21562 Iterations - 5
```

```
summary(model1.3)
```

```
##
## Call:
## glm(formula = STATUS ~ TIME, family = binomial(link = "logit"),
##     trace = TRUE)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0184  -0.7201  -0.3294   0.6163   2.1373
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
```

```
## (Intercept)  2.773602   0.706277   3.927 8.60e-05 ***
## TIME        -0.023685   0.005027  -4.711 2.46e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 117.264  on 86  degrees of freedom
## Residual deviance:  80.216  on 85  degrees of freedom
## AIC: 84.216
##
## Number of Fisher Scoring iterations: 5
```

```
anova(model1.3,test="Chisq")
```

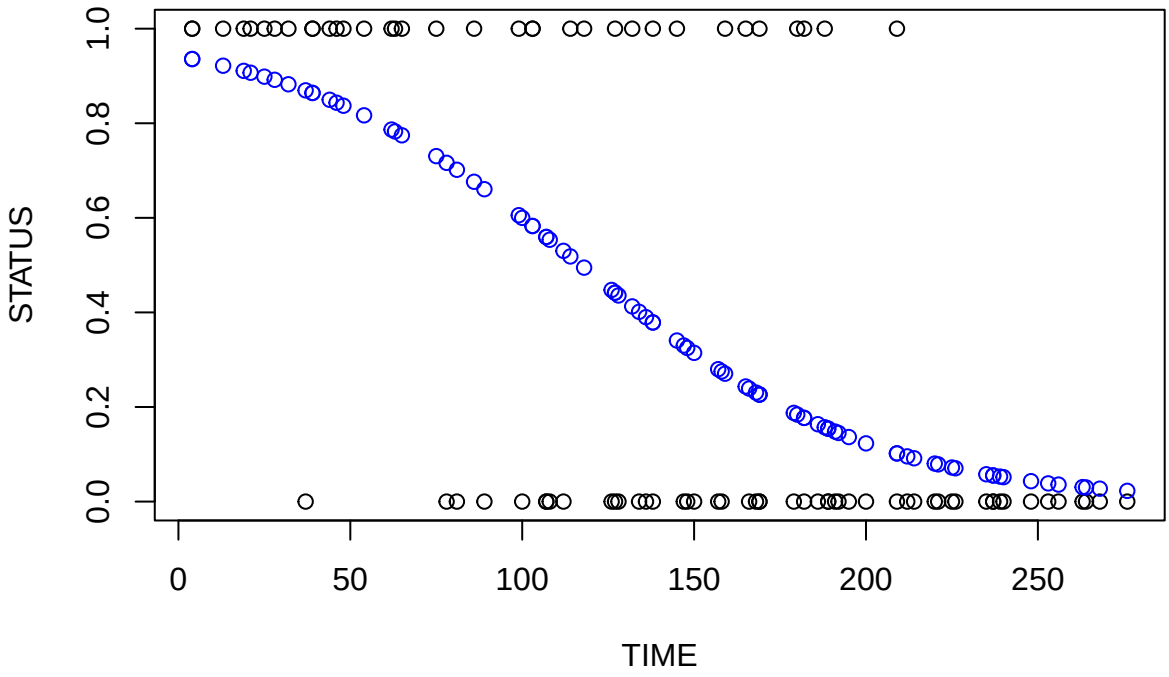
```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: STATUS
##
## Terms added sequentially (first to last)
##
##
##      Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                    86    117.264
## TIME  1   37.049        85     80.216 1.152e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
dchisq(36.619 ,1)
```

```
## [1] 7.367871e-10
```

I think the model that can be 3 is the one that fits best STATUS~TIME

```
plot(STATUS~TIME)+
points(TIME,fitted(model1.3),pch=1,col="blue")
```




```
## integer(0)
```

Another model to look at is the logarithm of DURATION.

```
model1.4<-glm(STATUS~LOG.1.DURATION.,family = binomial(link = 'logit'), trace=TRUE)
```

```
## Deviance = 116.3639 Iterations - 1
## Deviance = 116.2203 Iterations - 2
## Deviance = 116.2203 Iterations - 3
## Deviance = 116.2203 Iterations - 4
```

```
summary(model1.4)
```

```
##
## Call:
## glm(formula = STATUS ~ LOG.1.DURATION., family = binomial(link = "logit"),
##     trace = TRUE)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.2364  -1.0111  -0.9271   1.2977   1.4504
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -0.6389     0.3269  -1.954   0.0507 .
## LOG.1.DURATION.   0.1684     0.1652   1.019   0.3080
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 117.26  on 86  degrees of freedom
## Residual deviance: 116.22  on 85  degrees of freedom
## AIC: 120.22
```

```
##
## Number of Fisher Scoring iterations: 4
```

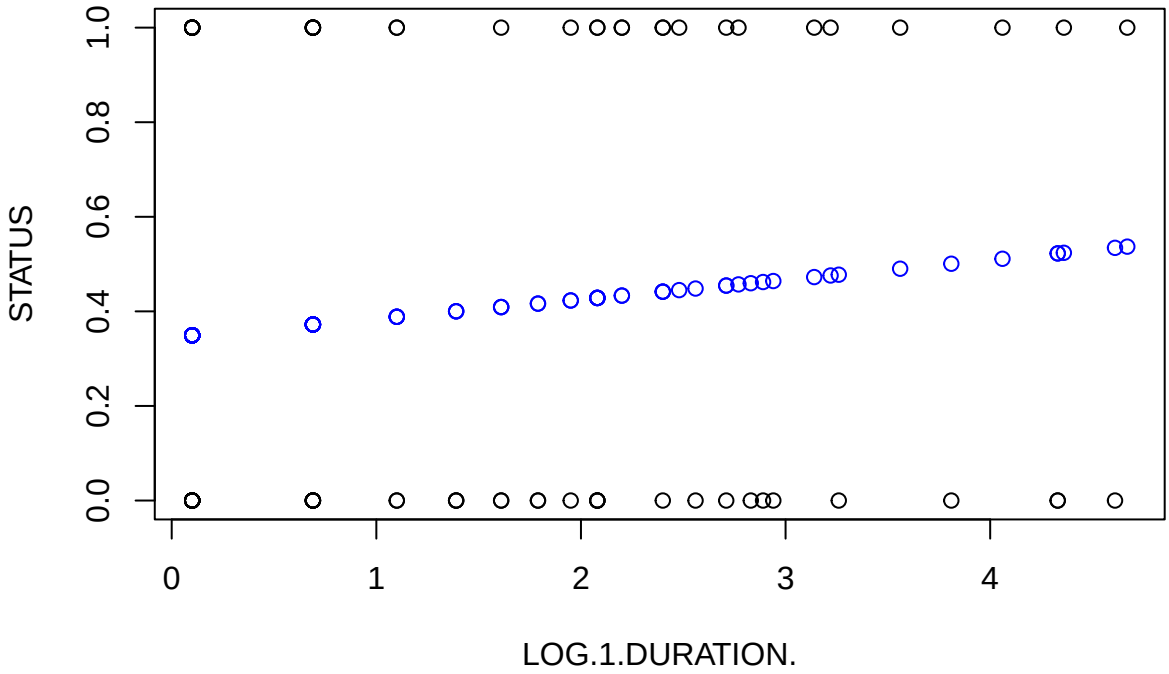
```
anova(model1.4,test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: STATUS
##
## Terms added sequentially (first to last)
##
##
##                 Df Deviance Resid. Df Resid. Dev Pr(>Chi)
## NULL                                86     117.26
## LOG.1.DURATION.  1    1.044        85     116.22   0.3069
```

```
dchisq(1.044 ,1)
```

```
## [1] 0.2316638
```

```
plot(STATUS~LOG.1.DURATION.)+
points(LOG.1.DURATION.,fitted(model1.4),pch=1,col="blue")
```




```
## integer(0)
```

In the end the best model looking at the anova test is model 1.3 STATUS~TIME.

To see how much the model explains the variance we can use the pseudoR, to calculate the pseudoR we first have to make the model null and then use the formula $R^2_{McFadden} = 1 - \frac{likelihoodmodel}{likelihoodnullmodel}$

```
nullmodel<-glm(STATUS~1,family = binomial(link = 'logit'), trace=TRUE)
```

```
## Deviance = 117.3952 Iterations - 1
## Deviance = 117.2643 Iterations - 2
```

```
## Deviance = 117.2643 Iterations - 3
## Deviance = 117.2643 Iterations - 4
```

```
PseudoR1 <- 1-logLik(model1.3)/logLik(nullmodel)
PseudoR1
```

```
## 'log Lik.' 0.3159417 (df=2)
```

```
PseudoR <- 1-logLik(model1.1)/logLik(nullmodel)
PseudoR
```

```
## 'log Lik.' 0.003745009 (df=2)
```

```
PseudoR2<- 1-logLik(model1.4)/logLik(nullmodel)
PseudoR2
```

```
## 'log Lik.' 0.008903181 (df=2)
```

With this we can see that while TIME can explain 32% of the variance, only 0.3% of the variance can be explained by DURATION and with its logarithm of 0.8% an improvement can be seen but it still does not explain the variance.

At the end we calculate the coefidence interval of the models using the formula $CI = \bar{x} \pm z. * SE$

```
M <- summary(model1.1)$coefficients
se <- M[2,2]#standard error
m<-mean(DURATION)
llb <-  m- qnorm(0.975)*se
ulb <- m + qnorm(0.975)*se
cib <- c(llb, ulb)
print(paste0("the CI of this model, of DURATION is ",  cib[1]," ",cib[2]))
```

```
## [1] "the CI of this model, of DURATION is 10.4605752244692 10.5003443157607"
```

```
M <- summary(model1.3)$coefficients
se <- M[2,2]#standard error
m<-mean(TIME)
llb <-  m- qnorm(0.975)*se
ulb <- m + qnorm(0.975)*se
cib <- c(llb, ulb)
print(paste0("the CI of this model, of TIME is ",  cib[1]," ",cib[2]))
```

```
## [1] "the CI of this model, of TIME is 142.33497401833 142.354681154084"
```

```
M <- summary(model1.4)$coefficients
se <- M[2,2]#standard error
m<-mean(LOG.1.DURATION.)
llb <-  m- qnorm(0.975)*se
ulb <- m + qnorm(0.975)*se
cib <- c(llb, ulb)
print(paste0("the CI of this model, of LOG.1.DURATION. is ", cib[1]," ",cib[2]))
```

```
## [1] "the CI of this model, of LOG.1.DURATION. is 1.09393586510313 1.74169631880491"
```

Conclusion

If we look at the analysis we can see that early detection is not relevant as there is no clear relationship between morbidity and early detection as seen in the variable DURATION and LOG.1.DURATION, what is closely related is the time in which you have suffered from the disease (TIME).

Bibliografy

Slide logistic regression. Jan Graffelman

medlineplus. https://medlineplus.gov/spanish/lupus.html

Apendix

you can find rmarkdawn on this github: https://github.com/rafaelcastilla/Project_Statistics