**Data Science for Medical Informatics**

**Lecture 2: Binary Response Variables + Logistic Regression + Fitting the model with R + goodness-of-fit**

**BINARY RESPONSE VARIABLE**

A binary variable is a variable with only two possible values. The focus of this lecture is on those binary variables which are response variables.

Example: A patient is admitted with abdominal sepsis (blood poisoning). The case is severe enough to warrant surgery. The patient is wheeled to the operation theatre. Let us speculate what will happen after the surgery. Let

Y = 1 if death follows surgery,

= 0 if the patient survives.

The outcome Y is random. Since it is random, we want to know its distribution.

Pr(Y = 1) = π, say, and

Pr(Y = 0) = 1 – π.

In simple terms, we want to know the chances (π) that a patient dies after surgery. Equivalently, what are the chances (1 – π) of survival after surgery? Are there any prognostic variables or factors which could influence the outcome Y. Surgeons list the following variables which could have some bearing on Y.

1. X1: Shock. Is the patient in a state of shock before the surgery?

X1 = 1 if yes,

= 0 if no.

1. X2: Nutrition. Is the patient undernourished?

X2 = 1 if yes,

= 0 if no.

1. X3: Alcoholism. Is the patient alcoholic?

X3 = 1 if yes,

= 0 if no.

1. X4: Age
2. X5: Bowel infarction. Has the patient bowel infarction?

X5 = 1 if yes,

= 0 if no.

The variables X1, X2, X3, and X5 are categorical covariates and X4 is a continuous covariate. The categorical variables are binary with only two possible values. It is felt that the outcome Y depends on these covariates.

Response variable Covariates, predictors, or independent variables

Y X1, X2, X3, X4, X5

Basic question: Which of the predictors have an impact on the response?

We do have data.

Structure of the data.

Patient Y X1 X2 X3 X4 X5

1 1 1 1 1 47 1

2 1 1 0 0 53 1

3 0 0 1 0 32 0

etc.

We have data on 106 patients. The entire data is posted on the blackboard under the name ‘Sepsis.’ The data is in EXCEL format.

A broad question of enquiry: How do the prognostic variables X1, X2, X3, X4, X5 influence the outcome of surgery?

Let us look at the data in R. Open an R console. The data is in my pen drive. Open the data and copy it to the clipboard (Windows). Download the data onto R.

> Sepsis <- read.delim("clipboard")

> dim(MB)

[1] 106 6

> head(Sepsis)

Shock Malnutrition Alcoholism Age Infarction Death

1 0 0 0 56 0 0

2 0 0 0 80 0 0

3 0 0 0 61 0 0

4 0 0 0 26 0 0

5 0 0 0 53 0 0

6 0 1 0 87 0 1

Let us look at the summary statistics of each variable.

> summary(Sepsis)

Shock Malnutrition Alcoholism Age

Min. :0.00000 Min. :0.0000 Min. :0.0000 Min. :17.00

1st Qu.:0.00000 1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.:33.00

Median :0.00000 Median :0.0000 Median :0.0000 Median :52.50

Mean :0.09434 Mean :0.3019 Mean :0.2075 Mean :51.28

3rd Qu.:0.00000 3rd Qu.:1.0000 3rd Qu.:0.0000 3rd Qu.:68.75

Max. :1.00000 Max. :1.0000 Max. :1.0000 Max. :94.00

Infarction Death

Min. :0.0000 Min. :0.0000

1st Qu.:0.0000 1st Qu.:0.0000

Median :0.0000 Median :0.0000

Mean :0.1321 Mean :0.1981

3rd Qu.:0.0000 3rd Qu.:0.0000

Max. :1.0000 Max. :1.0000

Each binary variable is treated as numeric. We need the counts of the levels of each binary variable. What is the class of each variable?

> apply(Sepsis, 2, class)

Shock Malnutrition Alcoholism Age Infarction Death

"integer" "integer" "integer" "integer" "integer" "integer"

Let us convert each binary variable into a factor (categorical variable with direction). Create another copy of Sepsis.

Sepsis1 <- Sepsis

> Sepsis1$Shock <- as.factor(Sepsis1$Shock)

> Sepsis1$Malnutrition <- as.factor(Sepsis1$Malnutrition)

> Sepsis1$Alcoholism <- as.factor(Sepsis1$Alcoholism)

> Sepsis1$Infarction <- as.factor(Sepsis1$Infarction)

> Sepsis1$Death <- as.factor(Sepsis1$Death)

Now do the summary.

> summary(Sepsis1)

Shock Malnutrition Alcoholism Age Infarction Death

0:96 0:74 0:84 Min. :17.00 0:92 0:85

1:10 1:32 1:22 1st Qu.:33.00 1:14 1:21

Median :52.50

Mean :51.28

3rd Qu.:68.75

Max. :94.00

Questions

Statistical

1. Does shock have an impact on ‘Death?’
2. Do ‘Death’ and ‘Shock’ are significantly associated?

Prediction

1. How to predict the outcome of surgery (Death or Survival) for a patient given information on the prognostic variables for the patient? Prediction can be model based. If a model is built using the data, how to assess how good the model is?

Data mining

1. Search for pockets of data in the predictors’ space such that in each pocket only one outcome (Death or Survival) is predominant. How to mount a search for such pockets?

Statistical question can be answered naively. Let us look at the 2x2 contingency table.

> Shock <- table(Sepsis$Death, Sepsis$Shock)

> Shock

0 1

0 82 3

1 14 7

> rownames(Shock) <- c("Survived", "Died")

> colnames(Shock) <- c("No Shock", "Shock")

> Shock

No Shock Shock

Survived 82 3

Died 14 7

Pr(Death|Shock) = Pr(Death = 1|Shock = 1) = 7/10 = 70%

Pr(Death|No Shock) = Pr(Death = 1|Shock = 0) = 14/96 = 14.6%

If the patient is in shock, he/she is much more likely to die than when not in shock (70% vs 14.5%). Is the difference significant? Can we say that being in shock or not has a great impact whether or not the patient survives? How strong we feel about it? Statistics allows you to formalize evidence. In this naïve approach, we have ignored the impact of the other predictors. We will answer this question through a model building endeavor.

Model Building: Logistic Regression

Recall Pr(Y = 1) = π

We want to model π as a function of X1 through X5. In order to indicate the dependence, we write

π = π(X1, X2, X3, X4, X5).

Why one wants to build a model? If a model is in place, one could use the model to assess the chances of survival after surgery for a patient before he is wheeled into the operation theatre. How? The surgeon could get information on X1, X2, X3, X4, X5 for the patient, calculate π = π(X1, X2, X3, X4, X5) from the postulated model, and then the chances of survival (1 – π) after surgery.

A possible model?

π(X1, X2, X3, X4, X5) = β0 + β1X1 + β2X2 + β3X3 + β4X4 + β5X5

This is like a multiple regression model. This model is not acceptable. The left hand side of the model π is a probability. The right hand side of the model could be any real number.

Why not model Y directly as a function of the covariates? For example,

Y = β0 + β1X1 + β2X2 + β3X3 + β4X4 + β5X5?

This is not acceptable. The left hand side Y takes only two values 0 and 1, but the right hand side could be any real number.

Logistic regression model

π(X1, X2, X3, X4, X5) = .

This model looks very formidable. The left hand side of the model is a probability and hence its value should always between zero and one. The right hand is always a number between zero and one. Why?

This model has 6 unknown parameters β0, β1, β2, β3, β4, and β5. We need to know the values of these parameters before it can be used. We can estimate the parameters of the model if we have data on a sample of patients. We do have data. We have data on 106 patients.

A digression: The approach I presented is purely statistical. Identify the variables of interest, designate the response variable or dependent variable, designate the covariates or independent variables, postulate a model, fit the model, and check it’s goodness-of-fit.

Engineers, physicists, and computer scientists will look at the problem from a different angle. They work directly with the dependent or response variable. I will talk about their approach later.

Problems (Back to our problem)

1. How does one estimate the parameters of the model using the data? There are two standard methods available. 1. Method of maximum likelihood. Write the likelihood of the data. Maximize the likelihood with respect to the parameters. 2. Method of weighted least squares. The least squares principle is used to minimize certain sum of squares. This method is much simpler than the method of maximum likelihood. Asymptotically, both methods are equivalent. If the sample is large, the estimates will be more or less the same.
2. Once the model is estimated, we need to check whether or not the model adequately summarizes the data. We need to assess how well the model fits the data. We may have to use some goodness-of-fit tests to make the assessment.
3. If the model fits the data well, we need to examine the impact of each and every covariate on the response variable. It is tantamount to identifying risk factors. We need to test the significance of each and every covariate in the model. If a covariate is not significant, we could remove the covariate from the model and then fit a leaner and tighter model to the data.
4. If a particular model does not fit the data well, explore other models, which can do a better job.
5. If an adequate model is fitted, explain how the model can be used in practice. Spend time on interpreting the model.

Before we pursue all these objectives, let us look at the model from another angle.

π(X1, X2, X3, X4, X5) = Probability of death after surgery for a patient with covariate values X1, X2, X3, X4, X5

= 

1 - π(X1, X2, X3, X4, X5) = Probability of survival after surgery for a patient with covariate values X1, X2, X3, X4, X5

= 

 = Odds of Death versus Survival after surgery

= exp{β0 + β1X1 + β2X2 + β3X3 + β4X4 + β5X5}

ln() = log odds = logit = β0 + β1X1 + β2X2 + β3X3 + β4X4 + β5X5

This is like a multiple regression model. The log odds are a linear function of the covariates! This form of the model is very useful for interpretation. The parameter β0 is called the intercept of the model. The parameter β1 is called the regression coefficient associated with the variable ‘Shock.’ The parameter β2 is called the regression coefficient associated with the variable ‘Malnutrition,’ etc. These regression coefficients indicate how much impact the corresponding covariates have on the response variable.

The logistic regression model can be spelled out either in the form

π(X1, X2, X3, X4, X5) =

,

or in the form

ln() = β0 + β1X1 + β2X2 + β3X3 + β4X4 + β5X5.

Both are equivalent.

Let us fit a logistic regression model to the ‘Sepsis’ data.

Look at the trick I have used. I did not name all the covariates!

Fit the logistic regression model to the data in R.

> Sepsis2 <- glm(Death ~ ., data = Sepsis, family = binomial)

> summary(Sepsis2)

Call:

glm(formula = Death ~ ., family = binomial, data = MB)

Deviance Residuals:

Min 1Q Median 3Q Max

-1.3277 -0.4204 -0.0781 -0.0274 3.2946

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -9.75391 2.54170 -3.838 0.000124 \*\*\*

Shock 3.67387 1.16481 3.154 0.001610 \*\*

Malnutrition 1.21658 0.72822 1.671 0.094798 .

Alcoholism 3.35488 0.98210 3.416 0.000635 \*\*\*

Age 0.09215 0.03032 3.039 0.002374 \*\*

Infarction 2.79759 1.16397 2.403 0.016240 \*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 105.528 on 105 degrees of freedom

Residual deviance: 53.122 on 100 degrees of freedom

AIC: 65.122

Number of Fisher Scoring iterations: 7

Comments: Write the prediction model. Discuss the implications of the model. Malnutrition is not significant.

A tight model: Remove ‘Malnutrition.’

Sepsis3 <- glm(Death ~ Shock + Alcoholism + Age + Infarction, data = MB, family = binomial)

> summary(Sepsis3)

Call:

glm(formula = Death ~ Shock + Alcoholism + Age + Infarction,

family = binomial, data = MB)

Deviance Residuals:

Min 1Q Median 3Q Max

-1.26192 -0.50391 -0.10690 -0.04112 3.06000

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -8.89459 2.31689 -3.839 0.000124 \*\*\*

Shock 3.70119 1.10353 3.354 0.000797 \*\*\*

Alcoholism 3.18590 0.91725 3.473 0.000514 \*\*\*

Age 0.08983 0.02922 3.075 0.002106 \*\*

Infarction 2.38647 1.07227 2.226 0.026039 \*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 105.528 on 105 degrees of freedom

Residual deviance: 56.073 on 101 degrees of freedom

AIC: 66.073

Number of Fisher Scoring iterations: 7

All covariates are significant.

What is the range of the patients with respect to Age. This information is useful later.

> summary(Sepsis$Age)

Min. 1st Qu. Median Mean 3rd Qu. Max.

17.00 33.00 52.50 51.28 68.75 94.00

> sd(MB$Age)

[1] 20.7877

Graphics

How do we present a logistic regression model graphically?

Logistic regression:

Fitted model

Pr(Death) = exp(-8.89458992 + 3.70119321\*Shock + 3.18590397\*Alcoholism + 0.08983175\*Age + 2.38646847\*Infarction)/

(1 + exp(-8.89458992 + 3.70119321\*Shock + 3.18590397\*Alcoholism + 0.08983175\*Age + 2.38646847\*Infarction))

An idea:

Plot Pr(Death) against ‘Age’ for each choice of Shock, Alcoholism, and Infarction.

What does the output Sepsis3 contain?

> names(Sepsis3)

[1] "coefficients" "residuals" "fitted.values"

[4] "effects" "R" "rank"

[7] "qr" "family" "linear.predictors"

[10] "deviance" "aic" "null.deviance"

[13] "iter" "weights" "prior.weights"

[16] "df.residual" "df.null" "y"

[19] "converged" "boundary" "model"

[22] "call" "formula" "terms"

[25] "data" "offset" "control"

[28] "method" "contrasts" "xlevels"

One can recall any of these using the $ symbol.

> Sepsis3$coefficients

(Intercept) Shock Alcoholism Age Infarction

-8.89458992 3.70119321 3.18590397 0.08983175 2.38646847

Operation plotting

Graph 1

Shock = 1; Alcoholism = 1; Infarction = 1

> curve(exp(-8.89458992 + 3.70119321 + 3.18590397 + 0.08983175\*x + 2.38646847)/

+ (1 + exp(-8.89458992 + 3.70119321 + 3.18590397 + 0.08983175\*x + 2.38646847)),

+ from = 15, to = 95, lwd = 2, col = "red", ylim = c(0, 1), xlab = "Age",

+ ylab = "Probability", main = "Logistic Regression", sub = "Sepsis Data")

Graph 2

Shock = 1; Alcoholism = 1; Infarction = 0

> curve(exp(-8.89458992 + 3.70119321 + 3.18590397 + 0.08983175\*x)/

+ (1 + exp(-8.89458992 + 3.70119321 + 3.18590397 + 0.08983175\*x)), col = "blue", lwd = 2, add = T)

Graph 3

Shock = 1; Alcoholism = 0; Infarction = 1

> curve(exp(-8.89458992 + 3.70119321 + 0.08983175\*x + 2.38646847)/

+ (1 + exp(-8.89458992 + 3.70119321 + 0.08983175\*x + 2.38646847)), col = "green", lwd = 2, add = T)

Graph 4

Shock = 0; Alcoholism = 1; Infarction = 1

> curve(exp(-8.89458992 + 3.18590397 + 0.08983175\*x + 2.38646847)/

+ (1 + exp(-8.89458992 + 3.18590397 + 0.08983175\*x + 2.38646847)), col = "black", lwd = 2, add = T)

Graph 5

Shock = 1; Alcoholism = 0; Infarction = 0

> curve(exp(-8.89458992 + 3.70119321 +0.08983175\*x)/

+ (1 + exp(-8.89458992 + 3.70119321 + 0.08983175\*x)), col = "gray", lwd = 2, add = T)

Graph 6

Shock = 0; Alcoholism = 1; Infarction = 0

> curve(exp(-8.89458992 + 3.18590397 + 0.08983175\*x)/

+ (1 + exp(-8.89458992 + 3.18590397 + 0.08983175\*x)), col = "violet", lwd = 2, add = T)

Graph 7

Shock = 0; Alcoholism = 0; Infarction = 1

> curve(exp(-8.89458992 + 0.08983175\*x + 2.38646847)/

+ (1 + exp(-8.89458992 + 0.08983175\*x + 2.38646847)), col = "orange", lwd = 2, add = T)

Graph 8

Shock = 0; Alcoholism = 0; Infarction = 0

> curve(exp(-8.89458992 + 0.08983175\*x)/(1 + exp(-8.89458992 + 0.08983175\*x)), col = "magenta", lwd = 2, add = T)

Add ‘legend’ to the graph.

> legend("right", pch = rep(16, 8), col = c("red", "blue", "green", "black",

+ "gray", "violet", "orange", "magenta"), legend = c("S=Y, A=Y, I=Y",

+ "S=Y, A=Y, I=N", "S=Y, A=N, I=Y", "S=N, A=Y, I = Y", "S=Y, A=N, I=N", "S=N, A=Y, I=N", "S=N, A=N, I=Y", "S=N, A=N, I=N"))



Interpretation:

Goodness-of-fit

Is the model fitted an adequate summary of the data?

H0: The response probability follows the logistic regression model, i.e.,

Pr(Death) =  for some β s.

H1: H0 not true

In simple terms, the null hypothesis states that the model is an adequate fit of the data.

Test Statistic: Residual Deviance with the stated degrees of freedom

Theory: If the null hypothesis is true, Residual Deviance has a chi-squared distribution with the stated degrees of freedom. A large value of Residual Deviance casts doubt on the null hypothesis. Calculate the p-value under the null hypothesis.

pvalue <- pchisq(53.122, 100, lower.tail = F)

The p-value is much larger than 0.05. There is no reason to reject the null hypothesis. The model adequately summarizes the data.