Data Science for Biomedical Research

(Adjusted) odds ratios; Prediction; Misclassification Rate; categorical predictors; Grouped data

We have a binary response variable Y. We have several covariates or predictors. How can we use the predictors to predict the response? One approach is to use a model for prediction. Logistic Regression is a way to do that. Let us work with the Sepsis data. Download the data.

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

> dim(MB)

[1] 106 6

> head(MB)

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

Fit a logistic regression model to the data.

Pr(Death) = Pr(Y = 1) =

and Pr(Y = 0) =

or in the form

ln() = log odds of death versus life

= β0 + β1Shock + β2Mal + β3Alcoholism + β4Age + β5Infarction

Let us estimate the parameters.

> MB1 <- glm(Death ~ ., data = MB, family = binomial)

> summary(MB1)

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

Goodness-of-fit

Is the model a good fit?

H0: The response probabilities follow the logistic regression model pattern.

Let us calculate the p-value under the null hypothesis. The residual deviance calculates discrepancy between what was observed under Y and what was predicted under the model. The idea is similar to a chi-squared goodness-of-fit procedure. In our context, residual deviance for a data point, say the first observation, is calculated as follows:

=

Residual deviance reported is the sum of squares of all individual residual deviances. If the model is a good fit, Residual Deviance has a chi-squared distribution with degrees of freedom = number of observations – number of parameters estimated.

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

pvalue

[1] 0.999968

The pvalue is much higher than 0.05. There is no reason to reject the null hypothesis.

Find confidence 95% intervals for each regression parameter.

> confint(MB1)

Waiting for profiling to be done...

2.5 % 97.5 %

(Intercept) -15.64470031 -5.5904775

Shock 1.63400028 6.3126637

Malnutrition -0.16922472 2.7410100

Alcoholism 1.64969091 5.5892967

Age 0.03979177 0.1600556

Infarction 0.71036445 5.3867158

Shock: 1.63 ≤ β1 ≤ 6.31 (Zero is outside.)

Undernourishment: - 0.17 ≤ β2 ≤ 2.74 (Zero is inside the interval. What does it mean?)

Alcoholism: 1.65 ≤ β3 ≤ 5.59

Age: 0.04 ≤ β4 ≤ 0.16

Bowel Infarction: 0.71 ≤ β5 ≤ 5.39

Model Based Odds Ratio

Odds Ratio Death vs Shock = = exp(β1)

We are calculating the odds of Death vs Life under two different scenarios. These two scenarios differ only with respect to Shock.

Odds Ratio Death vs Shock = = exp(β1)

Check this.

This is the (adjusted) odds ratio between Death and Shock adjusted for the presence of other predictors. Other adjusted odds ratios are similarly defined.

Let us estimate the odds ratios and obtain their confidence intervals.

Basic idea: 95% Confidence Interval of the Shock coefficient: 1.63 ≤ β1 ≤ 6.31

95% Confidence Interval between Death and Shock:

exp(1.63) ≤ exp(β1) ≤ exp(6.31)

> oddsratio <- exp(MB1$coefficients)

> oddsratio

(Intercept) Shock Malnutrition Alcoholism Age Infarction

5.806743e-05 3.940394e+01 3.375627e+00 2.864230e+01 1.096532e+00 1.640500e+01

Interpretation: We know that alcoholism is the most significant predictor of outcome surgery. Other conditions remaining the same,

Odds ratio of Outcome of surgery versus alcoholism

= = 28.64

We can stress the role of alcoholism on death after surgery. Suppose if one is not alcoholic, and Pr(Death) = Pr(Survival) = 50%. Then if one is alcoholic,

Pr(Death)/Pr(Survival) = 28.64.

This means that Pr(Death) = 28.64/(1 + 28.64) = 96.6%. The risk of death after surgery goes up by 96.6%.

Medical Interpretation:

Even keel: Chances of death after surgery are 50% if one is not alcoholic.

If one is alcoholic, chances of death after surgery are 97%.

If out of every 100 people who are not alcoholic 50 die, then 47 more die if they are alcoholic.

One can get confidence intervals for the odds ratios.

> exp(confint(MB1))

Waiting for profiling to be done...

2.5 % 97.5 %

(Intercept) 1.605436e-07 3.733245e-03

Shock 5.124333e+00 5.515120e+02

Malnutrition 8.443192e-01 1.550263e+01

Alcoholism 5.205371e+00 2.675474e+02

Age 1.040594e+00 1.173576e+00

Infarction 2.034733e+00 2.184847e+02

The intervals are very wide. Sample is small.

Prediction

Let us calculate the prediction probability of Y = 1 from the fitted model for each patient in the study.

> MB2 <- predict(MB1, newdata = MB, type = "response")

> head(MB2)

1 2 3 4 5 6

0.0100174835 0.0845795486 0.0157879878 0.0006371069 0.0076163433 0.3728426160

> round(MB2, 3)

1 2 3 4 5 6 7 8 9 10 11 12 13

0.010 0.085 0.016 0.001 0.008 0.373 0.000 0.490 0.011 0.177 0.923 0.005 0.000

14 15 16 17 18 19 20 21 22 23 24 25 26

0.004 0.001 0.004 0.002 0.001 0.970 0.001 0.180 0.001 0.366 0.926 0.085 0.000

27 28 29 30 31 32 33 34 35 36 37 38 39

0.039 0.150 0.830 0.000 0.000 0.132 0.187 0.006 0.145 0.060 0.295 0.980 0.586

40 41 42 43 44 45 46 47 48 49 50 51 52

0.596 0.012 0.251 0.003 0.035 0.035 0.001 0.000 0.085 0.079 0.030 0.002 0.001

53 54 55 56 57 58 59 60 61 62 63 64 65

0.180 0.244 0.000 0.466 0.008 0.094 0.001 0.001 0.065 0.028 0.003 0.002 0.976

66 67 68 69 70 71 72 73 74 75 76 77 78

0.299 0.048 0.081 0.895 0.001 0.002 0.439 0.095 0.055 0.945 0.002 0.241 0.001

79 80 81 82 83 84 85 86 87 88 89 90 91

0.001 0.164 0.000 0.238 0.331 0.439 0.536 0.155 0.729 0.191 0.000 0.103 0.943

92 93 94 95 96 97 98 99 100 101 102 103 104

0.001 0.015 0.000 0.002 0.449 0.023 0.004 0.000 0.490 0.747 0.339 0.001 0.308

105 106

0.225 0.212

What are we getting here? For each patient in our data, MB2 is calculating Pr(Death) using the fitted model.

Give a name to these predicted probabilities.

> MB2 <- round(MB2, 3)

Misclassification rate

Let us actually predict ‘Death’ or ‘Survival’ using the model.

Predict ‘Death’ if Pr(Death) ≥ 0.5.

Predict ‘Survival’ if Pr(Death) < 0.5.

Let us do it.

> MB3 <- ifelse(MB2 >= 0.5, 1, 0)

> MB3

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20

0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 1 0

21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40

0 0 0 1 0 0 0 0 1 0 0 0 0 0 0 0 0 1 1 1

41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60

0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0

61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80

0 0 0 0 1 0 0 0 1 0 0 0 0 0 1 0 0 0 0 0

81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99 100

0 0 0 0 1 0 1 0 0 0 1 0 0 0 0 0 0 0 0 0

101 102 103 104 105 106

1 0 0 0 0 0

Interpretation: For Patient 1, the model predicted ‘Survival’ after surgery. In reality the patient survived. The model did a good job. And so on. Let us find out how many mistakes the model committed.

> ConfusionMatrix <- table(MB$Death, MB3)

> ConfusionMatrix

MB3

0 1

0 84 1

1 8 13

> rownames(ConfusionMatrix) <- c("ObservedSurvival", "ObservedDeath")

> colnames(ConfusionMatrix) <- c("PredictedSurvival", "PredictedDeath")

> ConfusionMatrix

MB3

PredictedSurvival PredictedDeath

ObservedSurvival 84 1

ObservedDeath 8 13

> MisclassificationRate <- 9/106

> MisclassificationRate

[1] 0.08490566

The misclassification rate is 8%. Not bad.

How to handle predictors which are categorical (factor)? An example from multiple regression lecture.

Multi-Level Predictors in Multiple Regression

Predictors are of several kinds: integer; numeric; factor. A factor could have more than two levels. A factor might have a sense of direction. Or, might not. Let us look at an example. Install and activate the package ‘faraway.’ Download the data ‘happy.’

Genesis of data: A survey was conducted how happy one is. Happiness is measured on a scale from 1 to 10. Query: What factors influence happiness?

> data(package = "faraway")

> data(happy)

> head(happy)

happy money sex love work

1 10 36 0 3 4

2 8 47 1 3 1

3 8 53 0 3 5

4 8 35 1 3 3

5 4 88 1 1 2

6 9 175 1 3 4

Get documentation on the data.

> ?happy

Discussion of the variables: ‘happy’ is numeric. I want to take this variable as the response variable. What makes one happy? Predictors are: money; sex; love; work. Clearly, ‘money’ is numeric. The others seem to be factors. ‘sex’ is binary (1 = sex life is good; 0 = No sex). Don’t disturb it. ‘love’ seems to be ternary (1 = lonely; 2 = secure relationship; 3 = deep sense of belonging). It has a sense of direction. One can keep it as numeric. It will have only one regression coefficient. Or, keep it as categorical. One of the levels has to be base line. It will have two regression coefficients. ‘work’ seems to have five levels. It has a sense of direction. One could keep it as numeric. It will have only one regression coefficient. Or, make it categorical. One of the levels has to be base line. It will have four regression coefficients. Let us try both.

Let us find the status of each variable.

> apply(happy, 2, class)

happy money sex love work

"numeric" "numeric" "numeric" "numeric" "numeric"

Create a duplicate copy of ‘happy.’

> happy1 <- happy

Change ‘love’ and ‘work’ into factors.

> happy1$love <- as.factor(happy1$love)

> happy1$work <- as.factor(happy1$work)

Model 1: ‘love’ and ‘work’ are numeric.

Happy = β0 + β1\*money + β2\*sex + β3\*love + β4\*work + ε

Fit this model to the data.

> happy2 <- lm(happy ~ ., data = happy)

> summary(happy2)

Call:

lm(formula = happy ~ ., data = happy)

Residuals:

Min 1Q Median 3Q Max

-2.7186 -0.5779 -0.1172 0.6340 2.0651

Coefficients:

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

(Intercept) -0.072081 0.852543 -0.085 0.9331

money 0.009578 0.005213 1.837 0.0749 .

sex -0.149008 0.418525 -0.356 0.7240

love 1.919279 0.295451 6.496 1.97e-07 \*\*\*

work 0.476079 0.199389 2.388 0.0227 \*

---

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

Residual standard error: 1.058 on 34 degrees of freedom

Multiple R-squared: 0.7102, Adjusted R-squared: 0.6761

F-statistic: 20.83 on 4 and 34 DF, p-value: 9.364e-09

Interpretation … ‘love’ and ‘work’ are significant.

Model 2: ‘love’ and ‘work’ are categorical (factor)

R takes love = 1 as base line; work = 1 as base line.

Happy = β0 + β1\*money + β2\*sex + β3\*love2 + β4\*love3 + β5\*work2 + β6\*work3 + β7\*work4 + β8\*work5 + ε

> happy3 <- lm(happy ~ ., data = happy1)

> summary(happy3)

Call:

lm(formula = happy ~ ., data = happy1)

Residuals:

Min 1Q Median 3Q Max

-2.42802 -0.50047 -0.05559 0.63203 2.16380

Coefficients:

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

(Intercept) 3.370241 0.993288 3.393 0.00196 \*\*

money 0.008374 0.005464 1.533 0.13587

sex -0.345443 0.470171 -0.735 0.46822

love2 1.850241 0.766964 2.412 0.02217 \*

love3 3.845091 0.722507 5.322 9.39e-06 \*\*\*

work2 -0.792463 0.920846 -0.861 0.39629

work3 0.113597 0.899973 0.126 0.90040

work4 0.808892 0.857931 0.943 0.35329

work5 0.382735 1.128814 0.339 0.73693

---

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

Residual standard error: 1.059 on 30 degrees of freedom

Multiple R-squared: 0.744, Adjusted R-squared: 0.6757

F-statistic: 10.9 on 8 and 30 DF, p-value: 4.769e-07

Interpretation:

1. ‘money’ does not have a significant impact on happiness.
2. ‘sex’ does not have a significant impact on happiness.
3. We can only compare the baseline with every other category of ‘love.’ Love = 1 and Love = 2 can be compared. The p-value is 0.02. The coefficient of ‘love2’ is positive. Being Love = 2 has significantly more impact on happiness than being Love = 1. Compare Love = 1 and Love = 3. The p-value is 9.39\*10-6. Being Love = 3 has significantly much more impact on happiness than being Love = 1.

One can also examine overall impact of ‘love’ and ‘work’ on happiness.

> anova(happy3)

Analysis of Variance Table

Response: happy

Df Sum Sq Mean Sq F value Pr(>F)

money 1 9.649 9.649 8.6025 0.006375 \*\*

sex 1 1.940 1.940 1.7294 0.198449

love 2 75.542 37.771 33.6757 2.146e-08 \*\*\*

work 4 10.657 2.664 2.3753 0.074334 .

Residuals 30 33.648 1.122

---

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

Interpretation …

If ‘money’ is the only predictor in the model, then it is significant. Suppose ‘money’ is in the model. Does adding ‘sex’ improve the model? No. Suppose ‘money’ and ‘sex’ are in the model. Does adding ‘love’ improve the model? Yes

Logistic Regression for Grouped data

Activate the package ‘MASS.’

> data(package = "MASS")

Download the data ‘snails.’

> data(snails)

> dim(snails)

[1] 96 6

> head(snails)

Species Exposure Rel.Hum Temp Deaths N

1 A 1 60.0 10 0 20

2 A 1 60.0 15 0 20

3 A 1 60.0 20 0 20

4 A 1 65.8 10 0 20

5 A 1 65.8 15 0 20

6 A 1 65.8 20 0 20

Look at its documentation.

> ?snails

starting httpd help server ... done

Technically, we are looking at 20\*96 snails. The data are grouped.

> summary(snails)

Species Exposure Rel.Hum Temp Deaths

A:48 Min. :1.00 Min. :60.00 Min. :10 Min. : 0.000

B:48 1st Qu.:1.75 1st Qu.:64.35 1st Qu.:10 1st Qu.: 0.000

Median :2.50 Median :68.15 Median :15 Median : 1.000

Mean :2.50 Mean :68.03 Mean :15 Mean : 2.865

3rd Qu.:3.25 3rd Qu.:71.83 3rd Qu.:20 3rd Qu.: 4.250

Max. :4.00 Max. :75.80 Max. :20 Max. :16.000

N

Min. :20

1st Qu.:20

Median :20

Mean :20

3rd Qu.:20

Max. :20

I am unhappy: Exposure is numeric; Temp is numeric.

Change their status. Create a duplicate copy of the data.

> snails1 <- snails

> snails1$Exposure <- as.factor(snails1$Exposure)

> class(snails1$Exposure)

[1] "factor"

> snails1$Temp <- as.factor(snails1$Temp)

> class(snails1$Temp)

[1] "factor"

I need the following to work with grouped data.

> snails1$Survival <- 20 - snails1$Deaths

> head(snails1)

Species Exposure Rel.Hum Temp Deaths N Survival

1 A 1 60.0 10 0 20 20

2 A 1 60.0 15 0 20 20

3 A 1 60.0 20 0 20 20

4 A 1 65.8 10 0 20 20

5 A 1 65.8 15 0 20 20

6 A 1 65.8 20 0 20 20

> summary(snails1)

Species Exposure Rel.Hum Temp Deaths N

A:48 1:24 Min. :60.00 10:32 Min. : 0.000 Min. :20

B:48 2:24 1st Qu.:64.35 15:32 1st Qu.: 0.000 1st Qu.:20

3:24 Median :68.15 20:32 Median : 1.000 Median :20

4:24 Mean :68.03 Mean : 2.865 Mean :20

3rd Qu.:71.83 3rd Qu.: 4.250 3rd Qu.:20

Max. :75.80 Max. :16.000 Max. :20

Survival

Min. : 4.00

1st Qu.:15.75

Median :19.00

Mean :17.14

3rd Qu.:20.00

Max. :20.00

R code for grouped data. Write the model. A surprise! Alpha-numeric principle is flouted.

> Model <- glm(cbind(Deaths, Survival) ~ Species + Exposure + Rel.Hum +

+ Temp, data = snails1, family = binomial)

> summary(Model)

Call:

glm(formula = cbind(Deaths, Survival) ~ Species + Exposure +

Rel.Hum + Temp, family = binomial, data = snails1)

Deviance Residuals:

Min 1Q Median 3Q Max

-1.68917 -0.43933 -0.00007 0.22205 1.29245

Coefficients:

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

(Intercept) -16.84750 2045.08744 -0.008 0.99343

SpeciesB 1.28736 0.16161 7.966 1.64e-15 \*\*\*

Exposure2 18.94452 2045.08726 0.009 0.99261

Exposure3 21.17994 2045.08725 0.010 0.99174

Exposure4 22.12363 2045.08725 0.011 0.99137

Rel.Hum -0.10543 0.01377 -7.654 1.95e-14 \*\*\*

Temp15 0.57467 0.19791 2.904 0.00369 \*\*

Temp20 0.94304 0.19403 4.860 1.17e-06 \*\*\*

---

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

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 539.721 on 95 degrees of freedom

Residual deviance: 31.748 on 88 degrees of freedom

AIC: 206.61

Number of Fisher Scoring iterations: 19

Adequacy of the model.

> pvalue <- pchisq(31.748, 88, lower.tail = FALSE)

> pvalue

[1] 1

Fantastic!

Discussion

Species A and B are significantly different. The coefficient is positive. Species B has a higher chance of death.

Relative humidity is very, very significant. The higher the humidity is, the lower the chances of death.

Temperature 10 and 15 are significantly different; Temperature 10 and 20 are significantly different.