REVOLUTION ANALYTICS

Modeling in RRE Module 3: Logistic Regression

September 9, 2014

Logistic Regression

Let's extend our discussion about regression. There are actually several different types of regression models, each based on different characteristics in observed data.

Logistic regression applies to binary response variables - that is, terms that assume only two values. Some example binary variables are:

- Positive/Negative account balance (as in our Bank data)
- Under 18/Over 18 years of age (applicable to some concert venues)
- Good/Bad song review (applicable to ratings)
- Alive/Deceased subjects (applicable to Survival Analysis)

as well as many, many other applications…

Logistic Regression

Binary response variables can be modeled using Logistic Regression:

$$\frac{F(x)}{1 - F(x)} = e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m}$$

Where the above regression assumes multiple predictors and the function $F(t) = \frac{1}{1 + e^{-t}}$ is referred to as the logistic function, and t is a linear function to the explanatory variable x or a linear combination of explanatory variables.

Logistic Regression

Let's consider the previous equation in a little more detail, first by talking about the Logistic Function:

$$F(t) = \frac{1}{1 + e^{-t}}$$

The variable *t* represents a linear function, just like models we constructed in the last module:

$$t = \beta_0 + \beta_1 x$$

where *x* is the variable we are modeling.

Logistic Regression

If we are dealing with multiple predictor variables we would like to model, then our equation for *t* changes to a linear combination of those predictor variables. Mathematically, for *m*-predictor variables, the equation for *t* becomes:

$$t = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m$$

We are already familiar with both of these models from the previous module! Essentially, Logistic Regression wraps these fundamental models around some additional theory which we will consider briefly.

Logistic Regression

Euler's number is a common constant, $2.71828\ldots$ that turns up a lot in math.

For instance, it is very important in:

- calculating interest on a loan
- some forms of population growth such as for certain bacteria
- other examples involving increases on an exponential scale.

Logistic Regression

Alternately, the Natural Logarithm acts as the opposite to using Euler's number as the base of an exponent. In R, the standard log function actually is taking the natural log of a value, simply because of its vast use in both application and theory (much more so than the base 10 log) and because it is easy to convert logarithmic bases using laws of logarithms. Models involving the natural logarithm include:

- decaying population, perhaps for a bacteria that is exposed to pesticide
- carbon dating, a process referring to the decaying of half-life of Carbon atoms in an object
- other such problems involving a reducing sample

Logistic Regression

In Logistic Regression, the exponential term is used to model a binary response variable. Let's consider a supposed variable *x* as having a response equal to either 0 or 1. Statistically, our model should predict *x* as falling between one of these extreme values - for instance, perhaps a higher proportion of Carbon atoms in some object should have decayed over a certain period of time, so we would model the question of whether a given Carbon atom in that object would have decayed as closer to 1, but not a definite 1 (indicating that all of the Carbon atoms have decayed).

Modeling *t* with a linear response variable, we get the following plot for our Logistic Function:

```
x = -10:10
y = 1/(1 + exp(-x))
plot(x, y, type = "l", col = "red")
```

Logistic Regression

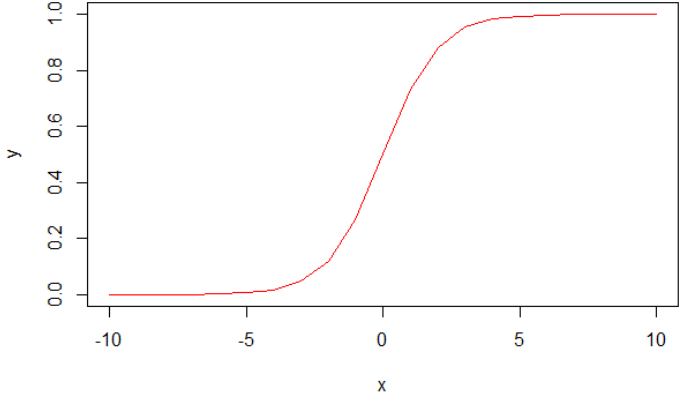

Figure:

Logistic Regression

Of course, we can see that the distributional proportion will depend on the linear modeling of *t*, but in the previous plot one may see the characteristic S-shape of the logistic curve. This shows values falling between minimum and maximum values 0 and 1, respectfully, while asymptotically approaching these values. Predictions based on the above model would strictly follow the line.

Logistic Regression

The rxLogit function fits a logistic regression model to data by specifying a formula and a root data set:

```
rxLogit(formula, data, ...)
```

Example: Logistic Regression

In our Bank data set, notice that the default variable has a binary response of either yes or no. Therefore, to model this variable it makes sense to use logistic regression.

Let's predict default based on age and balance:

```
infile <- file.path("data", "BankXDF.xdf")
BankDS <- RxXdfData(file = infile)
logitMod <- rxLogit(default ~ age + balance, data = BankDS)
```

```
## Rows Read: 10000, Total Rows Processed: 10000, Total Chunk Time: 0.002 seconds
## Rows Read: 10000, Total Rows Processed: 20000, Total Chunk Time: 0.003 seconds
## Rows Read: 10000, Total Rows Processed: 30000, Total Chunk Time: 0.002 seconds
## Rows Read: 10000, Total Rows Processed: 40000, Total Chunk Time: 0.002 seconds
## Rows Read: 5211, Total Rows Processed: 45211, Total Chunk Time: 0.002 seconds
##
...
```

Example: Logistic Regression

The RRE output for Logistic Regression is a little different from the linear regression output of the prior module.

```
summary(logitMod)

## Call:
## rxLogit(formula = default ~ age + balance, data = BankDS)
##
## Logistic Regression Results for: default ~ age + balance
## Data: BankDS (RxXdfData Data Source)
## File name: data/BankXDF.xdf
...
```

Apart from the usual terms is the log-likelihood ratio statistic, testing the fit between the null model and the alternative model. Essentially, this value is used to compute a p-value concerning the relevancy of the model.

From the above coefficients, we can reconstruct the logistic model

Example: Logistic Regression

Mathematically, the model reconstructed to two decimal places is:

$$\frac{F(x)}{1-F(x)} = e^{-3.08-0.008391 \cdot age-0.00227 \cdot balance}$$

Perhaps the interpretation of a logistic model is a little less apparent than it was with linear regression. In this case, exponentiating the coefficients and interpreting them as odds-ratios should make much more sense.

Example: Odds-Ratio

Computing the Odds-Ratios:

```
## Age
exp(-0.008391)

## Balance
exp(-0.002266)
```

Interpreting the results for individuals in our Bank data,

- Given a one unit increase in age (equivalent to an additional year), the odds of having credit in default (versus not having credit in default) decreases by a factor of approximately 0.9916.
- Given a one unit increase in balance (equivalent to having one additional Euro), the odds of having credit in default (versus not having credit in default) decreases by a factor of approximately 0.9977.

Exercise: Logistic Regression

For this exercise, we will pull on all of the information we have learned so far in this module. Let's construct a model that indicates the success of the Portuguese banking institution's success with its marketing campaign.

Exercise: Logistic Regression

- Use logistic regression to model whether or not a client subscribed to a term deposit (i.e. the variable y in the Bank data set), given information on the client's age, balance, housing, and marital status.
- After constructing the model, interpret your results using odds-ratios for each variable coefficient.

Exercise: Solution

First, using ScaleR to construct the logistic regression using the Bank data source:

```
logitMod <- rxLogit(y ~ age + balance + housing + marital, data = BankDS)
```

```
## Rows Read: 10000, Total Rows Processed: 10000, Total Chunk Time: 0.002 seconds
## Rows Read: 10000, Total Rows Processed: 20000, Total Chunk Time: 0.006 seconds
## Rows Read: 10000, Total Rows Processed: 30000, Total Chunk Time: 0.004 seconds
## Rows Read: 10000, Total Rows Processed: 40000, Total Chunk Time: 0.006 seconds
## Rows Read: 5211, Total Rows Processed: 45211, Total Chunk Time: 0.005 seconds
##
...
```

Analyze the summary using the summary function:

```
summary(logitMod)
```

```
## Call:
## rxLogit(formula = y ~ age + balance + housing + marital, data = BankDS)
##
## Logistic Regression Results for: y ~ age + balance + housing +
```

Exercise: Solution

```
df <- as.data.frame(logitMod$coefficients)
df[2] <- exp(df[1])
names(df) <- c("coeff", "oddsRatio")
df
```

```
##                       coeff oddsRatio
## (Intercept)      -2.052e+00    0.1284
## age               8.701e-03    1.0087
## balance           3.153e-05    1.0000
## housing=yes      -8.250e-01    0.4382
## housing=no               NA        NA
...
```

Exercise: Solution

All of the variables listed above were determined to be highly significant. Accordingly, our variable interpretations are listed as follows:

- For age, given a one year increase, the odds of the client subscribing to a term deposit (versus not subscribing) increases by a factor of approximately 1.008739.
- For balance, given a one euro increase, the odds of the client subscribing to a term deposit (versus not subscribing) increases by a factor of approximately 1.000032.
- Given yes on housing, the odds of the client subscribing to a term deposit (versus not subscribing) decreases by a factor of approximately 0.4382246.
- Given being married, the odds of the client subscribing to a

Recap

Let's review some of the concepts covered in this module:

- How does logistic regression differ from linear regression?
  - In what ways is it similar?
- What is a binary response variable?
  - Can you come up with an example?
- How do you interpret an odds-ratio?

REVOLUTION ANALYTICS

Thank you

Revolution Analytics is the leading commercial provider of software and support for the popular open source R statistics language.

www.revolutionanalytics.com, 1.855.GET.REVO, Twitter: @Revo-lutionR