Logistic Regression

Benjamin Kedem

June 2020

Logistic Regression

Qustion: Suppose we have data on texting while driving. How could we use such data to quantify the effect of texting on the **chance** of an accident?

Answer: This can be done by **logistic regression**.

Example: Chance of an accident as a function of covariates.

Define:

y=Accident	1	Accident last year
	0	No accident lat year
x_2=Age		Measured in years
x_3=Vision	0	No problem
	1	Some problem
x_4=Drive_Ed	1	Yes
	0	No

If p is the probability of an accident, the objective is to get the log-odds $\log[p/(1-p)]$. Observe that $p = E(y)$, that is, the mean of y.

Logistic regression model:

$$\log\left(\frac{p}{1-p}\right) = \log\left(\frac{P(accident)}{1-P(accident)}\right) = \beta_0 + \beta_1 Age + \beta_2 Vision + \beta_3 Driv\_Ed$$

Since the Accident data are 0-1, we can get the likelihood of the parameters $L(\boldsymbol{\beta})$, and from it get the AIC and BIC.

Fact: $Odds = p/(1-p)$. Then

$$p = \frac{Odds}{1+Odds}$$

Hence:

$$Odds \leftrightarrow p$$

Why is this called logistic regression? Since we express p in terms of the logistic CDF.

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 Age + \beta_2 Vision + \beta_3 Drive\_Ed \equiv \boldsymbol{\beta}'\boldsymbol{x}$$

then, solving for p we have:

$$p = F_l(\boldsymbol{\beta}'\boldsymbol{x}) = \frac{1}{1+\exp(-\boldsymbol{\beta}'\boldsymbol{x})}$$

where $F_l(x)$ is the CDF of the logistic distribution. Observe that:

$$\boldsymbol{\beta}'\boldsymbol{x} = \log\left(\frac{p}{1-p}\right) = F_l^{-1}(p)$$

So:

$$g(p) = \boldsymbol{\beta}'\boldsymbol{x}$$

That is, a monotone function of the mean of y is modeled as a linear model!!! This is a special case of GLM.

The function $g(\cdot)$ is called **link function**.

Observe that $p = F_l(\boldsymbol{\beta}'\boldsymbol{x})$. Hence:
1. $0 \le p \le 1$.
2. $F_l^{-1}(p) = \boldsymbol{\beta}'\boldsymbol{x}$, that is F_l^{-1} is a **link function**.

```
DATA LOGISTIC;
INPUT ACCIDENT AGE VISION DRIVE_ED;
DATALINES;
1 17 1 1
1 44 0 0
1 48 1 0
1 55 0 0
1 75 1 1
0 35 0 1
0 42 1 1
0 57 0 0
0 28 0 1
0 20 0 1
0 38 1 0
0 45 0 1
0 47 1 1
0 52 0 0
0 55 0 1
```

```
1 68 1 0
1 18 1 0
1 68 0 0
1 48 1 1
1 17 0 0
1 70 1 1
1 72 1 0
1 35 0 1
1 19 1 0
1 62 1 0
0 39 1 1
0 40 1 1
0 55 0 0
0 68 0 1
0 25 1 0
0 17 0 0
0 45 0 1
0 44 0 1
0 67 0 0
0 55 0 1
1 61 1 0
1 19 1 0
1 69 0 0
1 23 1 1
1 19 0 0
1 72 1 1
1 74 1 0
1 31 0 1
1 16 1 0
1 61 1 0
;
PROC LOGISTIC DATA=LOGISTIC DESCENDING;
MODEL ACCIDENT=AGE VISION DRIVE_ED;
RUN;
QUIT;
```

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	63.827	58.158
SC	65.633	65.385
-2 Log L	61.827	50.15

Note LRT: $61.827 - 50.15 = 11.677$ is a value of $\chi^2_{(3)}$ with which we test the hypothesis the global hypothesis:

$$H_0 : \beta_1 = \beta_2 = \beta_3 = 0$$

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	$Pr > ChiSq$
Likelihood Ratio	11.6682	3	0.0086

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	SE	Wald Chi-Square	$Pr > ChiSq$
Intercept	1	-0.1883	0.9945	0.0359	0.8498
AGE	1	0.00656	0.0183	0.1290	0.7195
VISION	1	1.7096	0.7056	5.8708	0.0154
DRIVE_ED	1	-1.4937	0.7046	4.4949	0.0340

And we see AGE is not significant! Need to investigate.

```
PROC LOGISTIC DATA=LOGISTIC DESCENDING;
MODEL ACCIDENT=AGE VISION DRIVE_ED/selection=forward;
RUN;

Analysis of Maximum Likelihood Estimates

After two steps we get:
```

Parameter	DF	Estimate	SE	Wald Chi-Square	$Pr > ChiSq$
Intercept	1	0.1110	0.5457	0.0414	0.8389
VISION	1	1.7137	0.7049	5.9113	0.0150
DRIVE_ED	1	-1.5000	0.7037	4.5440	0.0330

So, the model for the probability of an accident p is:

$$\begin{aligned} logit &= \log\left(\frac{p}{1-p}\right) = \log(Odds\ of accident) \\ &= 0.1110 + 1.7137 * VISION - 1.5000 * DRIVE\_ED \end{aligned}$$

Or

$$\frac{p}{1-p} = \exp[0.1110 + 1.7137 * VISION - 1.5000 * DRIVE\_ED]$$

VISION=0, DRIVE_ED=1: ODDS=0.2493245
VISION=1, DRIVE_ED=0: ODDS=6.2009345

$$ODDS\ RATIO = \frac{6.2009345}{0.2493245} = 24.87094$$

Hence, if there is a vision problem, and no driver's ed then the odds for an accident increases almost 25 times.

VISION	DRIVE_ED	ODDS=p/(1-p)
0	0	1.1173950
1	0	6.2009345—Highest
0	1	0.2493245—Smallest
1	1	1.3836155

We saw that AGE was not included as its β was not significant. Let's look at the age distribution sorted by accident.

```
proc sort data=logistic;
   by accident;
run;
proc gchart data=logistic;
vbar age/Midpoints=10 to 90 by 5;
run;
```

From the plot, the "middle age" class tends to have less accidents. Therefore, it makes to replace AGE by AGEGROUP.

AGEGROUP =0 if AGE in [20,65] —"Middle age"
AGEGROUP= 1 otherwise. — "Young and old".

```
DATA LOGISTIC;
INPUT ACCIDENT AGE VISION DRIVE_ED;
IF AGE < 20 OR AGE > 65 THEN AGEGROUP=1;
ELSE AGEGROUP=0;
DATALINES;
1 17 1 1
1 44 0 0
1 48 1 0
1 55 0 0
1 75 1 1
0 35 0 1
0 42 1 1
0 57 0 0
0 28 0 1
0 20 0 1
0 38 1 0
0 45 0 1
0 47 1 1
0 52 0 0
0 55 0 1
1 68 1 0
1 18 1 0
1 68 0 0
```

```
1 48 1 1
1 17 0 0
1 70 1 1
1 72 1 0
1 35 0 1
1 19 1 0
1 62 1 0
0 39 1 1
0 40 1 1
0 55 0 0
0 68 0 1
0 25 1 0
0 17 0 0
0 45 0 1
0 44 0 1
0 67 0 0
0 55 0 1
1 61 1 0
1 19 1 0
1 69 0 0
1 23 1 1
1 19 0 0
1 72 1 1
1 74 1 0
1 31 0 1
1 16 1 0
1 61 1 0
;

PROC LOGISTIC DATA=LOGISTIC DESCENDING;
MODEL ACCIDENT=AGEGROUP VISION DRIVE_ED/SELECTION=FORWARD;
RUN;
QUIT;
```

`FORWARD selected AGEGROUP and VISION. The MLE's are:`

Parameter	DF	Estimate	SE	Wald χ^2	$Pr > ChiSq$
Intercept	1	-1.3334	0.5854	5.1886	0.0227
AGEGROUP	1	2.1611	0.8014	7.2711	0.0070
VISION	1	1.6258	0.7325	4.9265	0.0264

Check:
$P(\chi^2_{(1)} > 5.1886) = 0.02273552$
$P(\chi^2_{(1)} > 7.2711) = 0.007007288$
$P(\chi^2_{(1)} > 4.9265) = 0.02644783$

This time β of AGEGROUP is significant and the model becomes:

$$\log\left(\frac{p}{1-p}\right) = -1.3334 + 2.1611 * AGEGROUP + 1.6258 * VISION$$

Young or Old, with a vision problem:
AGEGROUP=1, VISION=1, ODDS=11.62898, p=0.920817
Middle age with good vision:
AGEGROUP=0, VISION=0, ODDS=0.2635796, p=0.2085975

ODDS RATIO: 11.62898/0.2635796=44.11942

Model	AIC
Accident=Vision + DR_ED	56.2874
Accident=Age+Vision+DR_ED	58.1583
Accident=AGEGROUP+Vision	52.4340—Better model

PROC GENMOD

The GENMOD procedure fits generalized linear models, as defined by Nelder and Wedderburn (1972).

Instead of PROC LOGISTIC we could use PROC GENMOD which is more general.

```
DATA LOGISTIC;
INPUT ACCIDENT AGE VISION DRIVE_ED;
DATALINES;
1 17 1 1
1 44 0 0
1 48 1 0
1 55 0 0
1 75 1 1
0 35 0 1
0 42 1 1
0 57 0 0
0 28 0 1
0 20 0 1
0 38 1 0
0 45 0 1
0 47 1 1
0 52 0 0
0 55 0 1
1 68 1 0
1 18 1 0
```

```
1 68 0 0
1 48 1 1
1 17 0 0
1 70 1 1
1 72 1 0
1 35 0 1
1 19 1 0
1 62 1 0
0 39 1 1
0 40 1 1
0 55 0 0
0 68 0 1
0 25 1 0
0 17 0 0
0 45 0 1
0 44 0 1
0 67 0 0
0 55 0 1
1 61 1 0
1 19 1 0
1 69 0 0
1 23 1 1
1 19 0 0
1 72 1 1
1 74 1 0
1 31 0 1
1 16 1 0
1 61 1 0
;
PROC GENMOD DATA=LOGISTIC DESCENDING;
MODEL ACCIDENT=AGE VISION DRIVE_ED/dist = bin
                        link = logit;
RUN;
QUIT;
```

Example from SAS Web Page: Five drugs: A,B,C,D,E. Each drug is tested on a number of different subjects. The outcome of each experiment is the presence or absence of a positive response in a subject. The following artificial data represent the number of positive responses r in the n subjects for each of the five different drugs, labeled A through E. The response is measured for different levels of a continuous covariate x for each drug.

The drug type and the continuous covariate x are explanatory variables in this experiment. The number of positive responses r is modeled as a binomial random variable for each combination of the explanatory variable values, with the binomial number of trials parameter equal to the number of subjects n and the binomial probability equal to the probability of a response.

A logistic regression for these data is a generalized linear model with response equal to the binomial proportion r/n. The probability distribution is binomial, and the link function is logit. For these data, drug and x are explanatory variables. The probit and the complementary log-log link functions are also appropriate for binomial data.

PROC GENMOD performs a logistic regression on the data in the following SAS statements.

```
data drug;
   input drug$ x r n @@;
   datalines;
A  .1   1  10   A  .23  2  12   A  .67  1   9
B  .2   3  13   B  .3   4  15   B  .45  5  16   B  .78  5  13
C  .04  0  10   C  .15  0  11   C  .56  1  12   C  .7   2  12
D  .34  5  10   D  .6   5   9   D  .7   8  10
E  .2  12  20   E  .34 15  20   E  .56 13  15   E  .8  17  20
;

proc genmod data=drug;
   class drug;
   model r/n = x drug / dist = bin
                        link = logit;
run;
```