6.2.2 - Fitting the Model in R

6.2.2 - Fitting the Model in R

There are different ways to run logistic regression depending on the format of the data. Here are some general guidelines to keep in mind with a simple example outlined in dataformats.R where we created two binary random variables with n number of trials, e.g., $n = 100$. [1]

```
##create two binary vectors of length 100
x=sample(c(0,1),100, replace=T)
y=sample(c(0,1),100, replace=T)
```

If the data come in a tabular form, i.e., response pattern is with counts (as seen in the previous example), the data are said to be "grouped".

```
> ##create a 2x2 table with counts
> xytab=table(x,y)
> xytab
y
x 0 1
0 24 29
1 26 21
```

glm(): Let Y be the response variable capturing the number of events with the number of success ($Y = 1$) and failures ($Y = 0$). We need to create a response table that has a count for both the "success" and "failure" out of n trials in its columns. Notice that the count table below could be also the number of success $Y = 1$, and then a column computed as $n - Y$

```
> count=cbind(xytab[,2],xytab[,1])
> count
[,1] [,2]
0 29 24
1 21 26
```

We also need a categorical predictor variable.

```
> xfactor=factor(c("0","1"))
> xfactor
[1] 0 1
Levels: 0 1
```

We need to specify the response distribution and a link, which in this case is done by specifying `family=binomial("logit")`.

```
> tmp3=glm(count~xfactor, family=binomial("logit"))
> tmp3
Call: glm(formula = count ~ xfactor, family = binomial("logit"))
Coefficients:
(Intercept) xfactor1
0.1892 -0.4028
Degrees of Freedom: 1 Total (i.e. Null); 0 Residual
Null Deviance: 1.005
Residual Deviance: -4.441e-15 AIC: 12.72
```

If data come in a matrix form, i.e., subject × variables matrix with one line for each subject, like a database, where data are "ungrouped".

```
> xydata=cbind(x,y)

> xydata ## 100 rows, we are showing first 7
x y
[1,] 0 1
[2,] 0 1
[3,] 0 0
[4,] 0 1
[5,] 1 0
[6,] 0 1
[7,] 0 0.....
```

glm(): We need a binary response variable Y and a predictor variable x, which in this case was also binary. We need to specify the response distribution and a link, which in this case is done by specifying family=binomial("logit")

```
> tmp1=glm(y~x, family=binomial("logit"))
> tmp1
Call: glm(formula = y ~ x, family = binomial("logit"))
Coefficients:
(Intercept) x
0.1892 -0.4028
Degrees of Freedom: 99 Total (i.e. Null); 98 Residual
Null Deviance: 138.6
Residual Deviance: 137.6 AIC: 141.6
```

We will follow the R output through to explain the different parts of model fitting. The output from SAS (or from many other software) will be essentially the same.

Example 6-2: Student Smoking

Let's begin with the collapsed 2×2 table:

	Student smokes	**Student does not smoke**
1–2 parents smoke	816	3203

Neither parent smokes	188	1168

In R, we can use the `glm()`function and specify the `family = binomial(link = logit)`. See the files smoke.R and the output generated in smoke.out. That R code corresponds to the SAS code discussed in the previous section:

```
#### define the explanatory variable with two levels:
#### 1=one or more parents smoke, 0=no parents smoke

parentsmoke=as.factor(c(1,0))

#### NOTE: if we do parentsmoke=c(1,0) R will treat this as
#### a numeric and not categorical variable
#### need to create a response vector so that it has counts for both "success" and
"failure"

response<-cbind(yes=c(816,188),no=c(3203,1168))
response

#### fit the logistic regression model
smoke.logistic<-glm(response~parentsmoke, family=binomial(link=logit))

#### OUTPUT

smoke.logistic
summary(smoke.logistic)
anova(smoke.logistic)
```

Here is the R output for the 2 × 2 table that we will use in R for logistics regression:

```
> response
yes no
[1,] 816 3203
[2,] 188 1168
```

Please Note: the table above is different from the one given from the SAS program. We input y and n as data columns in SAS; here we just input data columns as yes and no.

`summary(smoke.logistic)` gives the following model information:

```
call:
glm(formula = response ~ parentsmoke, family = binomial(link = logit))

Deviance Residuals:
[1]  0  0

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.82661    0.07858 -23.244  < 2e-16 ***
parentsmoke1  0.45918    0.08782   5.228 1.71e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance:  2.9121e+01  on 1  degrees of freedom
Residual deviance: -3.7170e-13  on 0  degrees of freedom
AIC: 19.242

Number of Fisher Scoring iterations: 2
```

Model Information & Model Convergence Status

```
Call:
glm(formula = response ~ parentsmoke, family = binomial(link = logit))
```

```
(Dispersion parameter for binomial family taken to be 1)
```

```
Number of Fisher Scoring iterations: 2
```

These sections tell us which dataset we are manipulating, the labels of the response and explanatory variables and what type of model we are fitting (e.g., binary logit), and the type of scoring algorithm for parameter estimation. Fisher scoring is a variant of Newton-Raphson method for ML estimation. In logistic regression they are equivalent. Since we are using an iterative procedure to fit the model, that is, to find the ML estimates, we need some indication if the algorithm converged.

Overdispersion is an important concept with discrete data. In the context of logistic regression, overdispersion occurs when the observed variance in the data tends to be larger than what the binomial model would predict. If $Y_i \sim \text{Binomial}(n_i, \pi_i)$, the mean is $\mu_i = n_i\pi_i$, and the variance is $n_i\pi_i(1-\pi_i$. Both of these rely on the parameter π_i, which can be too restrictive. If overdispersion is present in a dataset, the estimated standard errors and test statistics for individual parameters and the overall goodness-of-fit will be distorted and adjustments should be made. We will look at this briefly later when we look into continuous predictors.

Table of Coefficients

```
Coefficients:
Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.82661 0.07858 -23.244 < 2e-16 ***
parentsmoke1 0.45918 0.08782 5.228 1.71e-07 ***
--
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

This information gives us the fitted model:

$$\log \frac{\hat{\pi}_i}{1-\hat{\pi}_i} = \hat{\beta}_0 + \hat{\beta}_1 X_i = -1.837 + 0.459\text{parentsmoke1}$$

where parentsmoke1 is a **dummy variable** ((e.g. **design variable**) that takes value 1 if at least one parents is smoking and 0 if neither is smoking.

$x_1 = 1$ ("parentsmoke1") if parent smoking = at least one,
$x_1 = 0$ ("parentsmoke0") if parent smoking = neither

The group for parent smoking = 0 is the baseline. We are modeling the probability of "children smoking".

Estimated $\beta_0 = -1.827$ with a standard error of 0.078 is significant and it says that log-odds of a child smoking versus not smoking if neither parents is smoking (the baseline level) is -1.827 and it's statistically significant.

Estimated $\beta_1 = 0.459$ with a standard error of 0.088 is significant and it says that log-odds-ratio of a child smoking versus not smoking if at least one parent is smoking versus neither parents is smoking (the baseline level) is 0.459 and it's statistically significant. $\exp(0.459) = 1.58$ are the estimated odds-ratios.

Testing Individual Parameters

Testing the hypothesis that the probability of the characteristic depends on the value of the jth variable.

Testing $H_0: \beta_j = 0$ versus $H_A: \beta_j \neq 0$

The Wald chi-squared statistics $z^2 = (\hat{\beta}_j / \text{SE}(\hat{\beta}_k))^2$ for these tests are displayed along with the estimated coefficients in the "Coefficients" section.

The standard error for $\hat{\beta}_1$, 0.0878, agrees exactly with the standard error that we can calculate from the corresponding 2×2 table. The values indicate the significant relationship between the logit of the odds of student smoking in parents' smoking behavior. This information indicates that parents' smoking behavior is a significant factor in the model. We could compare z^2 to a chi-square with one degree of freedom; the p-value would then be the area to the right of z^2 under the χ^2_1 density curve.

A value of z^2 (Wald statistic) bigger than 3.84 indicates that we can reject the null hypothesis $\beta_j = 0$ at the .05-level.

$$\beta_1 : \left(\frac{0.4592}{0.0878}\right)^2 = 27.354$$

Confidence Intervals of Individual Parameters:

An approximate $(1-\alpha)100\%$ confidence interval for β_j is given by

$$\hat{\beta}_j \pm z_{(1-\alpha/2)} \times SE(\hat{\beta}_j)$$

For example, a 95% confidence interval for β_1

$$0.4592 \pm 1.96(0.0878) = (0.287112, 0.63128)$$

where 0.0878 is the "Standard Error"' from the "Coefficients" section. Then, the 95% confidence interval for the odds-ratio of a child smoking if one parent is smoking in comparison to neither smoking is

$$(\exp(0.287112), \exp(0.63128)) = (1.3325, 1.880)$$

Thus, there is a strong association between parent's and children smoking behavior. But does this model fit?

Overall goodness-of-fit testing

Test: H_0: current model vs. H_A: saturated model

```
Residual deviance: -3.7170e-13 on 0 degrees of freedom
AIC: 19.242
```

The goodness-of-fit statistics, deviance, G^2 from this model is zero, because the model is saturated. If we want to know the fit of the intercept only model that is provided by

Test: H_0: intercept only model vs. H_A: saturated model

```
Null deviance: 2.9121e+01 on 1 degrees of freedom
```

Suppose that we fit the intercept-only model. This is accomplished by removing the predictor from the model statement, like this:

```
glm(response~1, family=binomial(link=logit)
```

The goodness-of-fit statistics are shown below.

```
ull deviance: 29.121 on 1 degrees of freedom

Residual deviance: 29.121 on 1 degrees of freedom
AIC: 46.362
N
```

The deviance $G^2 = 29.1207$ is precisely equal to the G^2 for testing independence in this 2×2 table. Thus by the assumption, the intercept-only model or the null logistic regression model states that student's smoking is unrelated to parents' smoking (e.g., assumes independence, or odds-ratio=1). But clearly, based on the values of the calculated statistics, this model (i.e., independence) does NOT fit well.

Analysis of deviance table

In R, we can test factors' effects with the anova function to give an analysis of deviance table. We only include one factor in this model. So R dropped this factor (parentsmoke) and fit the intercept-only model to get the same statistics as above, i.e., the deviance $G^2 = 29.121$.

```
> anova(smoke.logistic)
Analysis of Deviance Table
Model: binomial, link: logit
Response: response
Terms added sequentially (first to last)
           Df Deviance Resid. Df Resid. Dev
NULL                           1     29.121
parentsmoke  1   29.121         0      0.000
```

This example shows that analyzing a 2×2 table for association is equivalent to logistic regression with a single dummy variable. We can further compare these tests to the loglinear model of independence in Lesson 10.

The goodness-of-fit statistics, X^2 and G^2, are defined as before in the tests of independence and loglinear models (e.g. compare observed and fitted values). For the chi^2 approximation to work well, we need the n_is to be sufficiently large so that the expected values, $\hat{\mu}_i \geq 5$, and $n_i - \hat{\mu}_i \geq 5$ for most of the rows. We can afford to have about 20% of these values less than 5, but none of them should fall below 1.

With real data, we often find that the n_is are not big enough for an accurate test, and there is no single best solution to handle this but a possible solution may rely strongly on the data and context of the problem.

Legend

[1]	Link
↥	Has Tooltip/Popover
⬚	Toggleable Visibility

Source: https://online.stat.psu.edu/stat504/lesson/6/6.2/6.2.2

Links:

1. https://online.stat.psu.edu/onlinecourses/sites/stat504/files/lesson06/dataformats.R