7.1 Binary Logistic Regression

Suppose the response variable y is **binary** (or **dichotomous**) variable, that is, it assumes only two possible values, which we will denote by 0 and 1. The relation between y and predictors $x_1, \dots, x_k$ can't be modeled by a general linear regression because the error terms would not be normally distributed. Instead of modeling y, we model the probability that y is equal to 1. We write $\pi = \mathbb{P}(y = 1)$. Note that π is also the mean of y. Indeed, $\mathbb{E}(y) = (1)(\pi) + (0)(1 - \pi) = \pi$.

Definition. The **binary** (or **dichotomous**) **logistic regression model** with the predictors $x_1, \dots, x_k$ has the form:

$$\pi = \mathbb{E}(y) = \frac{\exp\{\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k\}}{1 + \exp\{\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k\}}.$$

The name "logistic" comes from the fact that the distribution with the cdf $F(x) = \dfrac{e^x}{1 + e^x}$, $-\infty < x < \infty$, is called the **logistic distribution**.

Definition. A **generalized linear regression model** has the form $g(\mathbb{E}(y)) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$, where g is called the **link function**.

The binary logistic regression can be written as

$$\ln \frac{\pi}{1 - \pi} = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k.$$

Thus, a binary logistic regression is a generalized linear model with the link function $\ln \dfrac{\pi}{1 - \pi}$ which is called the **logit function** of π (from the words "logistic" and "unit").

Further, note that the ratio $\dfrac{\pi}{1 - \pi} = \dfrac{\mathbb{P}(y = 1)}{\mathbb{P}(y = 0)}$ represents the odds in favor of the event $y = 1$. The binary logistic regression is a linear model for the natural logarithm of the odds and hence is sometimes called **log-odds model**.

Definition. The **fitted binary logistic model** has the form

$$\widehat{\pi} = \widehat{\mathbb{E}}(y) = \frac{\exp\{\widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \cdots + \widehat{\beta}_k x_k\}}{1 + \exp\{\widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \cdots + \widehat{\beta}_k x_k\}},$$

or, written in terms of the estimated odds,

$$\frac{\widehat{\pi}}{1 - \widehat{\pi}} = \exp\{\widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \cdots + \widehat{\beta}_k x_k\}.$$

Definition. In the binary logistic regression model, the estimated regression coefficients yield the following **interpretation**:

• If a predictor variable x_1 is numeric, then the quantity $(\exp\{\widehat{\beta}_1\}-1)\cdot 100\%$ represents the **estimated percent change in odds** when x_1 is increased by one unit, and the other predictors are held fixed. This can be seen by writing:

$$\frac{\frac{\widehat{\pi}|_{x_1+1}}{1-\widehat{\pi}|_{x_1+1}}-\frac{\widehat{\pi}|_{x_1}}{1-\widehat{\pi}|_{x_1}}}{\frac{\widehat{\pi}|_{x_1}}{1-\widehat{\pi}|_{x_1}}}\cdot 100\% = \Big(\frac{\frac{\widehat{\pi}|_{x_1+1}}{1-\widehat{\pi}|_{x_1+1}}}{\frac{\widehat{\pi}|_{x_1}}{1-\widehat{\pi}|_{x_1}}}-1\Big)\cdot 100\%$$

$$= \Big(\frac{\exp\{\widehat{\beta}_0+\widehat{\beta}_1(x_1+1)+\widehat{\beta}_2\,x_2+\cdots+\widehat{\beta}_k\,x_k\}}{\exp\{\widehat{\beta}_0+\widehat{\beta}_1\,x_1+\cdots+\widehat{\beta}_k\,x_k\}}-1\Big)\cdot 100\% = \big(\exp\{\widehat{\beta}_1\}-1\big)\cdot 100\%\,.$$

• If a predictor variable x_1 is an **indicator** (or a **0-1**) **variable**, then the quantity $\exp\{\widehat{\beta}_1\}\cdot 100\%$ represents the **estimated percent ratio in odds** when $x_1=1$ and when $x_1=0$, while the other predictors are assumed constant. This can be seen by writing

$$\frac{\frac{\widehat{\pi}|_{x_1=1}}{1-\widehat{\pi}|_{x_1=1}}}{\frac{\widehat{\pi}|_{x_1=0}}{1-\widehat{\pi}|_{x_1=0}}}\cdot 100\% = \frac{\exp\{\widehat{\beta}_0+\widehat{\beta}_1\,\cdot 1+\widehat{\beta}_2\,x_2+\cdots+\widehat{\beta}_k\,x_k\}}{\exp\{\widehat{\beta}_0+\widehat{\beta}_1\cdot 0+\widehat{\beta}_2\,x_2+\cdots+\widehat{\beta}_k\,x_k\}}\cdot 100\%$$

$$= \exp\{\widehat{\beta}_1\}\cdot 100\%.$$

Definition. In a binary logistic regression, for a specified set of predictors $x_1^0,\dots,x_k^0$, the **predicted probability** π^0 can be found as

$$\pi^0 = \frac{\exp\{\widehat{\beta}_0+\widehat{\beta}_1\,x_1^0+\cdots+\widehat{\beta}_k\,x_k^0\}}{1+\exp\{\widehat{\beta}_0+\widehat{\beta}_1\,x_1^0+\cdots+\widehat{\beta}_k\,x_k^0\}}.$$

Example. A cohort study was conducted to investigate what factors are co-morbidities of pneumonia. The data file "pneumonia_data.csv" contains data on individuals' age, gender, and indicators (1=yes/0=no) of diabetes, asthma, hypertension, cardiovascular disease, obesity, and pneumonia; intensity of tobacco use (0=no/1=light/2=heavy); and PM2.5 measurement for the place of residence (in micro grams per cubic meter). We run SAS and R codes to regress pneumonia on the other variables.

In SAS:

```
proc import out=pneumonia datafile="./pneumonia_data.csv"
dbms=csv replace;

/*fitting logistic model*/
proc genmod;
class gender(ref="M") tobacco_use(ref="0");
   model pneumonia(event="1")= age gender diabetes asthma
hypertension cardiovascular obesity tobacco_use PM2_5/
             dist=binomial link=logit;
run;
```

Analysis Of Maximum Likelihood Parameter Estimates								
Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	-5.4991	1.1541	-7.7611	-3.2370	22.70	<.0001
age		1	0.0469	0.0178	0.0119	0.0818	6.91	0.0086
gender	F	1	0.4683	0.5531	-0.6157	1.5523	0.72	0.3972
gender	M	0	0.0000	0.0000	0.0000	0.0000	.	.
diabetes		1	0.9070	0.6459	-0.3590	2.1730	1.97	0.1603
asthma		1	1.6495	0.6730	0.3304	2.9686	6.01	0.0142
hypertension		1	0.1431	0.6955	-1.2201	1.5064	0.04	0.8370
cardiovascular		1	-0.8383	1.2590	-3.3059	1.6293	0.44	0.5055
obesity		1	-0.2583	0.6771	-1.5853	1.0687	0.15	0.7028
tobacco_use	1	1	1.6605	0.6016	0.4814	2.8397	7.62	0.0058
tobacco_use	2	1	4.8892	1.2059	2.5258	7.2527	16.44	<.0001
tobacco_use	0	0	0.0000	0.0000	0.0000	0.0000	.	.
PM2_5		1	-0.0013	0.0112	-0.0231	0.0206	0.01	0.9104
Scale		0	1.0000	0.0000	1.0000	1.0000		

From this output, significant predictors are age, asthma, and light and heavy tobacco use (the p-values are less than 0.05). The fitted model is

$$\frac{\widehat{\pi}}{1-\widehat{\pi}} = \exp\{-5.4991 + 0.0469 \cdot age + 0.4683 \cdot female + 0.9070 \cdot diabetes$$

$+1.6495 \cdot asthma + 0.1431 \cdot hypertension - 0.8383 \cdot cardiovascular - 0.2583 \cdot obesity$
$+1.6605 \cdot tobacco\ use\ light + 4.8892 \cdot tobacco\ use\ heavy - 0.0013 \cdot PM2.5\}.$

As age increases by one year, the estimated odds in favor of pneumonia increase by $(\exp(0.0469) - 1) \cdot 100\% = 4.80172\%$. For individuals with asthma, the estimated odds of pneumonia are $\exp(1.6495) \cdot 100\% = 520.4377\%$ of those without asthma. For light tobacco users, the estimated odds of pneumonia are $\exp(1.6605) \cdot 100\% = 526.1941\%$ of those who don't use tobacco. For heavy tobacco users, the estimated odds of pneumonia are $\exp(4.8892) \cdot 100\% = 13,284.73\%$ of those who don't use tobacco.

In R:

```
pneumonia.data<- read.csv(file="./pneumonia_data.csv",
header=TRUE, sep=",")

#specifying reference categories
gender.rel<- relevel(as.factor(pneumonia.data$gender), ref="M")
tobacco.use.rel<- relevel(as.factor(pneumonia.data$tobacco_use), ref="0")

#fitting logistic model
summary(fitted.model<- glm(pneumonia ~ age + gender.rel + diabetes
+ asthma + hypertension + cardiovascular + obesity + tobacco.use.rel
+ PM2_5, data=pneumonia.data, family=binomial(link=logit)))
```

```
Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)      -5.499074   1.154121  -4.765 1.89e-06
age               0.046856   0.017826   2.629  0.00858
gender.relF       0.468256   0.553071   0.847  0.39719
diabetes          0.906976   0.645933   1.404  0.16028
asthma            1.649520   0.673008   2.451  0.01425
hypertension      0.143117   0.695544   0.206  0.83698
cardiovascular   -0.838314   1.259000  -0.666  0.50550
obesity          -0.258283   0.677056  -0.381  0.70285
tobacco.use.rel1  1.660522   0.601617   2.760  0.00578
tobacco.use.rel2  4.889229   1.205856   4.055 5.02e-05
PM2_5            -0.001255   0.011151  -0.113  0.91037
```

Next, we use the fitted model to predict the probability of pneumonia for a 55-year old female who is obese, has asthma, and lives in an area with PM2.5

of 13.3. We write

$$\mathbb{P}^0(pneumonia) =$$

$$= \frac{\exp\{-5.4991 + 0.0469 \cdot 55 + 0.4683 + 1.6495 - 0.2583 - 0.0013 \cdot 13.3\}}{1 + \exp\{-5.4991 + 0.0469 \cdot 55 + 0.4683 + 1.6495 - 0.2583 - 0.0013 \cdot 13.3\}} = 0.254.$$

In SAS:

```
/*using fitted model for prediction*/
data prediction;
input age gender$ diabetes asthma hypertension cardiovascular obesity
tobacco_use PM2_5;
cards;
55 F 0 1 0 0 1 0 13.3
;

data pneumonia;
 set pneumonia prediction;
run;

proc genmod;
class gender tobacco_use;
  model pneumonia(event="1")= age gender diabetes asthma hypertension
    cardiovascular obesity tobaccouse PM2_5/dist=binomial link=logit;
 output out=outdata p=pprob_pneumonia;
run;

proc print data=outdata (firstobs=201) noobs;
var pprob_pneumonia;
run;
```

pprob_pneumonia
0.25366

In R:

```
#using fitted model for prediction
print(predict(fitted.model, type="response", data.frame(age=55,
gender.rel="F", diabetes=0, asthma=1, hypertension=0, cardiovascular=0,
obesity=1, tobacco.use.rel="0", PM2_5=13.3)))
```

```
0.2536582
```

□