8.2 Nonparametric Logistic Regression

Definition. Let y be a binary response variable, and let $\pi = \mathbb{P}(y = 1)$. A **nonparametric binary logistic regression model** has the form

$$\ln\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m + loess(x_{m+1}) + \cdots + loess(x_k)$$

where $x_1, \ldots, x_m$ is a set of **regression variables** (all 0-1 and possibly numeric predictors), and $x_{m+1}, \ldots, x_k$ are **smoothing predictors** (must be numeric).

Example. Consider the respiratory infection data set introduced earlier. We regress infection (yes/no) on gender and visit (regression predictors), and age and BMI (smoothing predictors).

In SAS:

```
proc import out=resp_data datafile="./respiratory_infection.csv"
dbms=csv replace;

/*creating longform dataset*/
data longform;
set resp_data;
 array x[4] xerophthalmia1-xerophthalmia4;
 array i[4] infection1-infection4;
  do visit=1 to 4;
   xerophthalmia=x[visit];
     infection=i[visit];
   output;
   end;
keep gender age BMI xerophthalmia visit infection;
run;

/*specifying data for prediction*/
data point4pred;
input gender$ visit age BMI ;
cards;
M 2 10 15.6
;

/*fitting nonparametric logistic model*/
proc gam data=longform; *gam=generalized additive model;
```

```
class gender(ref="M");
 model infection(event='1')= param(gender visit)
loess(age) loess(BMI)/link=logist dist=binomial;
  score data=point4pred out=predicted;
 run;
```

Regression Model Analysis
Parameter Estimates

Parameter	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	-1.29323	1.14831	-1.13	0.2615
gender F	0.32729	0.34818	0.94	0.3484
gender M	0	.	.	.
visit	-0.60464	0.15889	-3.81	0.0002
Linear(age)	0.32107	0.07482	4.29	<.0001
Linear(BMI)	-0.03588	0.05324	-0.67	0.5011

Smoothing Model Analysis
Analysis of Deviance

Source	DF	Sum of Squares	Chi-Square	Pr > ChiSq
Loess(age)	3.27636	2.757153	2.7572	0.4794
Loess(BMI)	3.13320	7.962066	7.9621	0.0519

Linear trends of visit and age are significant at the 5% level, and the detrended loess curve for BMI is marginally significant at the 5% level.

```
proc print data=predicted noobs;
run;
```

gender	visit	age	BMI	P_infection	P_age	P_BMI	LINP_infection
M	2	10	15.6	0.52259	0.094163	-0.15210	0.090429

The predicted probability of infection is 0.52259.

In R:

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

#creating long-form data set
library(reshape2)
longform.data<- melt(infection.data[,c("gender","age","BMI",
"infection1","infection2","infection3","infection4")],
id.vars=c("gender","age","BMI"), variable.name="infection_visit",
value.name="infection")

#creating variable for visit
longform.data$visit<- ifelse(longform.data$infection_visit=="infection1", 1,
ifelse(longform.data$infection_visit=="infection2", 2,
ifelse(longform.data$infection_visit=="infection3", 3, 4)))

#specifying reference category
longform.data$gender.rel<- relevel(as.factor(longform.data$gender), ref="M")

#fitting nonparametric logistic regression
library(gam)
logistic.fit<- gam(infection ~ gender.rel + visit + lo(age) + lo(BMI),
data=longform.data, family=binomial)
coefficients(summary.glm(logistic.fit))
```

```
               Estimate Std. Error    z value     Pr(>|z|)
(Intercept) -1.35851259 1.14211831 -1.1894675 2.342558e-01
gender.relF  0.26470362 0.34558259  0.7659634 4.436981e-01
visit       -0.59528474 0.15771169 -3.7745124 1.603209e-04
lo(age)      0.31362575 0.07491736  4.1862895 2.835517e-05
lo(BMI)     -0.02858864 0.05298921 -0.5395181 5.895294e-01
```

```
#using fitted model for prediction
predict(logistic.fit, data.frame(gender.rel="M", visit=2, age=10, BMI=15.6),
type="response")
```

```
0.4762946
```

□