

## Chapter 5: Categorical Data Analysis and Object-Oriented Programming (OOP)

### I. Statistics

#### 5.1. Categorical data analysis

##### 1. Logistic regression models:

$$Y_i \sim \text{Binomial}(n_i, p_i), 0 \leq Y_i \leq n_i, i = 1, \dots, m,$$

$$\text{logit} \left[ \frac{E(Y_i)}{n_i} \right] = \text{logit}(p_i) = \log \left( \frac{p_i}{1 - p_i} \right) = \sum_{j=1}^p \beta_j X_{ij}.$$

##### Example:

The following table refers to 661 children with birth weights 650 g and 1749 g all of whom survived for at least one year. The variables of interest are:

**Cardiac**: mild heart problems of the mother during pregnancy

**Comps**: gynaecological problems during pregnancy

**Smoking**: mother smoked at least one cigarette per day during the first months of pregnancy.

**BW**: was the birth weight less than 1250

<b>Cardiac</b>		Yes				No			
<b>Comps</b>		Yes		No		Yes		No	
<b>Smoking</b>		Yes	No	Yes	No	Yes	No	Yes	No
<b>BW</b>	Yes	10	25	12	15	18	12	42	45
	No	7	5	22	19	10	12	202	205

Analyze the data and interpret the relationship of the children weights and mother's habits and health conditions.

##### Example (Splus):

```
BW.yes=c(10,25,12,15,18,12,42,45)
```

```
BW.no=c(7,5,22,19,10,12,202,205)
```

```
BW=cbind(BW.yes,BW.no)
```

```
cardiac=factor(rep(c("0","1"),each=4))
```

```
comps=factor(rep(rep(c("0","1"),each=2),2))
```

```
smoking=factor(rep(c("0","1"),4))
```

```
survived.glm=glm(BW~cardiac+comps+smoking,family=binomial(link=logit))
```

```
anova(survived.glm,test="Chisq")
summary(survived.glm)
survived.glm2=glm(BW~cardiac+comps,family=binomial(link=logit))
anova(survived.glm2,test="Chisq")
summary(survived.glm2)
```

## 2. Log-linear models:

$$Y_i \sim \text{Poisson}(\lambda_i), i = 1, \dots, n,$$

$$\log[E(Y_i)] = \log(\lambda_i) = \sum_{j=1}^p \beta_j X_{ij}.$$

### Example:

The data given in Splus data frame *Insurance* (in the library MASS) consist of the numbers of policy-holders, *Holders*, the numbers of car insurance claims made by those policyholders, *Claims*. There are three explanatory variables, *District* (four levels), *Group* (of car, four levels), and *Age* (four ordered levels). Please analyze the data using log-linear models with offset  $\log(\text{Holders})$ .

### Example (Splus):

```
library(MASS)
data=Insurance
age=factor(data$Age)
group=factor(data$Group)
dis=factor(data$District)
hol=log(data$Holders)
cla=data$Claims
dataf=data.frame(hol,dis,group,age,cla)
glm1=glm(cla~offset(hol)+dis+group+age,family=poisson,data=dataf)
anova(glm1,test="Chisq")
summary(glm1,cor=F)
```