logistic回归与glm（）函数

一、logistic回归

许多情况下假设因变量是正态分布是不合理的，比方说因变量是类别型的，如二值变量（是否）。对于这类问题计量的简单线性概率模型（LPM）会有很多缺陷比如

一、将自变量的特定组合数值带入回归后的模型中可能出现概率大于1或小于0的情况；

二、多数情况下预测的概率不可能与自变量是线性关系的。

为了克服LPM的缺陷计量计量经济学家提出了logistic模型。

logistic模型如下：

$$P(y = 1|\mathbf{x}) = G（\beta_0 +\beta_1x_1+\beta_2x_2+…+\beta_kx_k）（*）$$

其中G（x）= $\frac{e^x}{1+ex}$ 。为方便观察可将（*）式变形为：

$$Ln(\frac{p(y=1|\mathbf{x})}{1-p(y=1|\mathbf{x})}) =\beta_0 +\beta_1x_1+\beta_2x_2+…+\beta_kx_k$$

二、glm函数

在R语言中通常利用glm（）函数来实现logistic回归，glm()类似于lm（）函数，但是多了一些参数，他的基本函数类型为：

glm(formula, family = family(link = function), data=)

而lm（）的基本函数类型为：

lm（formula, data=）

因为lm（）是在残差服从正太分布的假设下所做的回归，而glm（）通过family参数提供非经典假设下的回归方法。

glm（）中概率分布family以及连接函数function的可以有如下设置：

分布族	默认连接函数
binomial	（link = “logit”）
gaussian	（link = “identity”）
gamma	（link =”inverse”）
poisson	（link =“1/mu^2”）
quasi	（link = “log”）
quasibinomial	（link = “logit”）
quasipoisson	（link = “log”）

三、代码演示

1、读取数据

```
data(Affairs, package="AER")
summary(Affairs)
Affairs$ynaffair[Affairs$affairs > 0 ] <- 1
Affairs$ynaffair[Affairs$affairs == 0 ] <- 0
Affairs$ynaffair <-
factor(Affairs$ynaffair,levels=c(0,1),labels=c("NO","YES"))
table(Affairs$ynaffair)
```

2、利用glm（）做logistic回归

```
fit.full<-glm(ynaffair ~ gender + age + yearsmarried + children + religiousness + education + occupation + rating, data=Affairs, family=binomial())
#性别、年龄、婚龄、是否有小孩、宗教信仰程度（5分制，1分表示反对，5分非常信仰）、学历、职业、婚姻的自我评分（5分制，1分表示非常不幸福，5分表示非常幸福）
summary(fit.full)
fit.reduced <- glm(ynaffair ~ age + yearsmarried +religiousness + rating, data= Affairs, family=binomial())
summary(fit.reduced)
anova(fit.reduced, fit.full ,test="Chisq")
coef(fit.reduced)
exp(coef(fit.reduced))
```

3、评价预测变量对结果概率的影响

```
testdata<-
data.frame(rating=c(1,2,3,4,5),age=mean(Affairs$age),yearsmarried=mean(Affairs$
yearsmarried),religiousness=mean(Affairs$religiousness))
testdata
testdata$prob <- predict(fit.reduced,newdata=testdata,type="response")
testdata
```

4、检验过度离势

过度离势指观测的响应变量的方差大于期望的二项分布的方差。它会导致奇异的标准误检验和不精确的显著性检验。

```
deviance(fit.reduced)/df.residual(fit.reduced)
#结果远大于1则认为存在标准误。
fit<-glm(ynaffair~age+yearsmarried + religiousness + rating, family = binomial(), data = Affairs)
fit.od <- glm(ynaffair ~ age + yearsmarried + religiousness + rating, family = quasibinomial(), data = Affairs)
pchisq(summary(fit.od)$dispersion*fit$df.residual,fit$df.residual,lower = F)
```