有序logit/probit线性模型

宋歌 2015080086 数52

5/30/2018

1 研究目的

探讨有序logit/probit线性模型的原理，并在给定数据集上进行试验。

2 研究原理

2.1 有序响应变量线性模型的建立过程

对于有序多分类的响应变量Y，我们假设存在一个潜在的连续型变量Y^*。
若将Y的取值按照顺序记为$1, 2, \ldots, J$，则有

$$
\begin{aligned}
Y &= 1 & -\infty < &Y^* \leqslant \theta_1 \\
Y &= 2 & \theta_1 < &Y^* \leqslant \theta_2 \\
&\vdots & &\vdots \\
Y &= J & \theta_{J-1} < &Y^* < \infty
\end{aligned}
$$

其中，

$$\theta_0 = -\infty < \theta_1 < \theta_2 < \cdots < \theta_{J-1} < \theta_J = \infty$$

是假定的分界点，将连续变量Y^* 的取值分为了$\mathbb{R}$上的J个区间。

我们假设Y^*与X之间的线性模型为

$$Y^* = X\beta + e, \qquad e \sim \mathbb{F}$$

从而我们可以建立如下模型

$$P(Y^* \leqslant j | X = x) = F(\theta_j - x^T\beta), \qquad j = 0, 1, 2, \ldots, J$$

$$P(Y = j | X = x) = F(\theta_j - x^T\beta) - F(\theta_{j-1} - x^T\beta)$$

其中常见的F选取为logistic函数或者标准正态概率分布函数，分别对应了有序logit/probit线性模型：

$$\text{ordered logit model:}\quad F(\theta_j - x^T\beta) = \frac{1}{1+exp(-(\theta_j - x^T\beta))} \quad \Rightarrow \quad \theta_j - x^T\beta = \log\frac{P(Y^* \leqslant j|x)}{1-P(Y^* \leqslant j|x)}$$
$$\text{ordered probit model:}\quad F(\theta_j - x^T\beta) = \Phi(\theta_j - x^T\beta) \quad \Rightarrow \quad \theta_j - x^T\beta = \Phi^{-1}(P(Y^* \leqslant j|x))$$

2.2 有序响应变量logit线性模型参数的解释

对于以上建立的有序响应变量logit线性模型，我们采用极大似然的方式进行参数估计。
对于给定的观测值(x_i, y_i)，$i = 1, 2, \ldots, n$，似然函数写为

$$\begin{aligned} L(\beta,\theta|Y,X) &= \prod_{i=1}^{n} P(Y_i = y_i|X = x_i, \beta, \theta) \\ &= \prod_{i=1}^{n}\prod_{j=1}^{J}\left[\text{logit}(\theta_j - x^T\beta) - \text{logit}(\theta_{j-1} - x^T\beta)\right]^{I\{y_i=j\}} \end{aligned}$$

对数似然写为

$$\begin{aligned} l(\beta,\theta|Y,X) &= \sum_{i=1}^{n}\sum_{j=1}^{J} I\{y_i = j\}\log\left[\text{logit}(\theta_j - x^T\beta) - \text{logit}(\theta_{j-1} - x^T\beta)\right] \\ &= \sum_{i=1}^{n}\sum_{j=1}^{J} I\{y_i = j\}\log\left[\frac{1}{1+\exp(-(\theta_j - x^T\beta))} - \frac{1}{1+\exp(-(\theta_{j-1} - x^T\beta))}\right] \end{aligned}$$

通过最大化似然函数得到参数的估计$\hat{\beta}, \hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_{J-1})$，即回归得到的模型为

$$log\frac{P(\hat{Y}_i \leqslant j|x_i)}{1-P(\hat{Y}_i \leqslant j|x_i)} = \hat{\theta}_j - x_i^T\hat{\beta}, \quad j = 1, 2, \ldots, J-1$$

即通过有序logistic回归，对于每一个样本点(x_i, y_i)，都给出了y_i属于各个有序类别的分数/概率。

3 实验过程与结果讨论

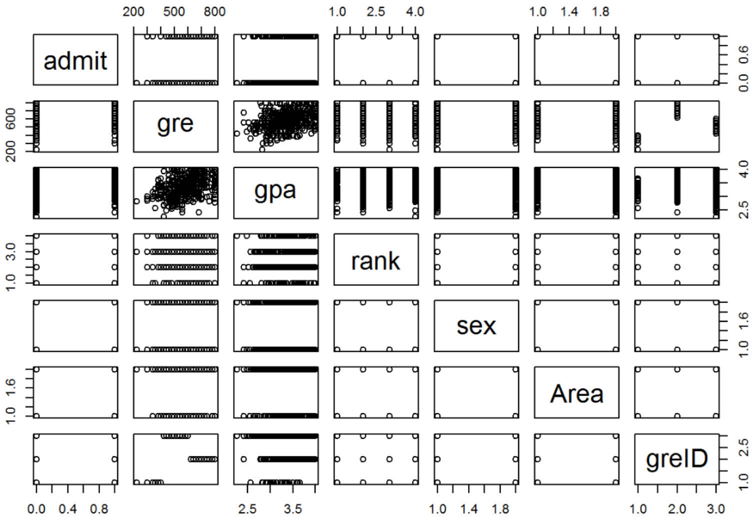
3.1 数据分析与处理

- 数据分析：

```
#读取并分析原始数据
dat <- read.table('pgBinary.txt')
summary(dat)
```

```
##      admit            gre             gpa             rank
##  Min.   :0.0000   Min.   :220.0   Min.   :2.260   Min.   :1.000
##  1st Qu.:0.0000   1st Qu.:520.0   1st Qu.:3.130   1st Qu.:2.000
##  Median :0.0000   Median :580.0   Median :3.395   Median :2.000
##  Mean   :0.3175   Mean   :587.7   Mean   :3.390   Mean   :2.485
##  3rd Qu.:1.0000   3rd Qu.:660.0   3rd Qu.:3.670   3rd Qu.:3.000
##  Max.   :1.0000   Max.   :800.0   Max.   :4.000   Max.   :4.000
##      sex        Area     greID
##  female:215   A:108   低: 31
##  male  :185   C:292   高:174
##                       中:195
##
##
##
```

```
plot(dat)
```



连续型变量：gre，gpa；
分类型变量：admit为二元分类变量，rank为四元分类变量，sex为二元分类变量，Area为二元分类变量，greID为三元分类变量；
从散点图中也可观察出，除了gre,gpa其余变量均为分类变量；
其中gre与gpa之间有一定的线性关系；gre与greID之间呈现出分段函数的形式；

- 创建哑变量

```
#admit,rank,sex,Area,greID化为哑变量
dat$admit <- factor(dat$admit)
dat$rank <- factor(dat$rank)
dat$sex <- factor(dat$sex)
dat$Area <- factor(dat$Area)
```

- 定义响应变量顺序

```
#定义低中高顺序以便回归
y <- c()
for(i in 1:400){
  if(dat$greID[i] == "低"){
    y[i] <- "a"
  }
  if(dat$greID[i] == "中"){
    y[i] <- "b"
  }
  if(dat$greID[i] == "高"){
    y[i] <- "c"
  }
}

y <- factor(y)
```

将“低，中，高”换为"a,b,c"，从而在回归时，R能够按照我们想要的顺序识别哑变量greID；

3.2 有序probit线性模型

- 方差分析
R中ANOVA用于检验两个模型之间相对的显著性。在这里我们设置了没有任何解释变量的空模型作为0假设；

```
probit0 = polr(y ~ 1, data = dat, method = "probit", Hess = T)
probit1 = polr(y ~ admit + gpa + rank + sex + Area, data = dat, method = "probit", Hess = T)

anova(probit0, probit1)
```

```
## Likelihood ratio tests of ordinal regression models
##
## Response: y
##                              Model Resid. df Resid. Dev   Test    Df
## 1                                1       398   728.4434
## 2 admit + gpa + rank + sex + Area       391   667.4537 1 vs 2     7
##   LR stat.      Pr(Chi)
## 1
## 2 60.98964 9.574541e-11
```

可见p值很小，拒绝零假设，从而probit1模型是显著的。

- 拟合结果

```r
summary(probit1)
```

```
## Call:
## polr(formula = y ~ admit + gpa + rank + sex + Area, data = dat,
##     Hess = T, method = "probit")
##
## Coefficients:
##              Value Std. Error t value
## admit1   0.25983     0.1351  1.9236
## gpa      1.10015     0.1663  6.6173
## rank2    0.07232     0.1838  0.3935
## rank3   -0.19779     0.1922 -1.0289
## rank4   -0.11680     0.2145 -0.5446
## sexmale  0.07932     0.1192  0.6654
## AreaC   -0.03987     0.1345 -0.2964
##
## Intercepts:
##     Value   Std. Error t value
## a|b  2.1875  0.5809     3.7656
## b|c  3.9517  0.6008     6.5773
##
## Residual Deviance: 667.4537
## AIC: 685.4537
```

回归所得的分界点估计为$\hat{\theta}_1 = 2.1875, \hat{\theta}_2 = 3.9517$

3.3 有序logit线性模型

- 方差分析

```r
#建立有序logit线性模型
logit0 = polr(y ~ 1, data = dat, method = "logistic", Hess = T)
logit1 = polr(y ~ admit + gpa + rank + sex + Area, data = dat, method = "logistic", Hess = T)

anova(logit0, logit1)
```

```
## Likelihood ratio tests of ordinal regression models
##
## Response: y
##                               Model Resid. df Resid. Dev   Test    Df
## 1                                 1       398   728.4434
## 2 admit + gpa + rank + sex + Area       391   669.1026 1 vs 2     7
##    LR stat.       Pr(Chi)
## 1
## 2 59.34073 2.043713e-10
```

可见p值很小，拒绝零假设，从而logit1模型是显著的。

- 拟合结果

```r
summary(logit1)
```

```
## Call:
## polr(formula = y ~ admit + gpa + rank + sex + Area, data = dat,
##     Hess = T, method = "logistic")
##
## Coefficients:
##             Value Std. Error t value
## admit1    0.40438     0.2288  1.7673
## gpa       1.83569     0.2855  6.4296
## rank2     0.16904     0.3098  0.5457
## rank3    -0.27845     0.3240 -0.8593
## rank4    -0.17051     0.3637 -0.4688
## sexmale   0.10619     0.2033  0.5222
## AreaC    -0.06429     0.2302 -0.2793
##
## Intercepts:
##     Value  Std. Error t value
## a|b 3.5640 0.9922     3.5919
## b|c 6.6088 1.0344     6.3891
##
## Residual Deviance: 669.1026
## AIC: 687.1026
```

回归所得的分界点估计为$\hat{\theta}_1 = 3.5640, \hat{\theta}_2 = 6.6088$