Logit模型与回归诊断

宋歌 2015080086 数52

5/16/2018

1 实验目的

给定数据集pgaBinary.txt，创建哑变量进行建模，并判断多重共线性性，进行回归诊断。

2 实验过程及结果讨论

2.1 创建哑变量并分析sex和Area是否对admit有影响

2.1.1 创建哑变量

```
#读取原始数据
dat <- read.table('pgaBinary.txt', header = TRUE)
summary(dat)


#创建哑变量
dat$rank <- factor(dat$rank)
dat$sex <- factor(dat$sex)
dat$Area <- factor(dat$Area)
```

在R中，将变量化为因子类型的变量后，在回归时会选择默认的基准变量并当成哑变量进行回归，即设计矩阵会变为如下形式：

```
X <- model.matrix( ~ gpa + gre + rank + sex + Area, dat)
head(X)
```

```
##   (Intercept)  gpa gre rank2 rank3 rank4 sexmale AreaC
## 1           1 3.61 380     0     1     0       1     1
## 2           1 3.67 660     0     1     0       0     1
## 3           1 4.00 800     0     0     0       1     1
## 4           1 3.19 640     0     0     1       0     1
## 5           1 2.93 520     0     0     1       0     1
## 6           1 3.00 760     1     0     0       1     1
```

容易看到，rank1，sexfemale，AreaA这三者被当作了基准变量，即取值为0.

2.1.2 判断是否有影响

```
#进行logistic回归
mylogit <- glm(admit ~ gre + gpa + rank + sex + Area, data = dat, binomial(link = 'logit'))
summary(mylogit)
```

```
##
## Call:
## glm(formula = admit ~ gre + gpa + rank + sex + Area, family = binomial(link = "logit"),
##     data = dat)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5956  -0.8806  -0.6307   1.1195   2.0881
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.794992   1.157048  -3.280 0.001038 **
## gre          0.002315   0.001098   2.109 0.034952 *
## gpa          0.790545   0.333118   2.373 0.017636 *
## rank2       -0.682479   0.316605  -2.156 0.031113 *
## rank3       -1.330628   0.345627  -3.850 0.000118 ***
## rank4       -1.534254   0.418169  -3.669 0.000244 ***
## sexmale     -0.005967   0.228759  -0.026 0.979190
## AreaC       -0.251343   0.249224  -1.009 0.313214
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 499.98  on 399  degrees of freedom
## Residual deviance: 457.51  on 392  degrees of freedom
## AIC: 473.51
##
## Number of Fisher Scoring iterations: 4
```

可以看到sex和Area变量均不显著，故尝试删去两个变量后进行回归：

```
#删去sex和Area后进行logistic回归
mylogit0 <- glm(admit ~ gre + gpa + rank, data = dat, binomial(link = 'logit'))
summary(mylogit0)
```

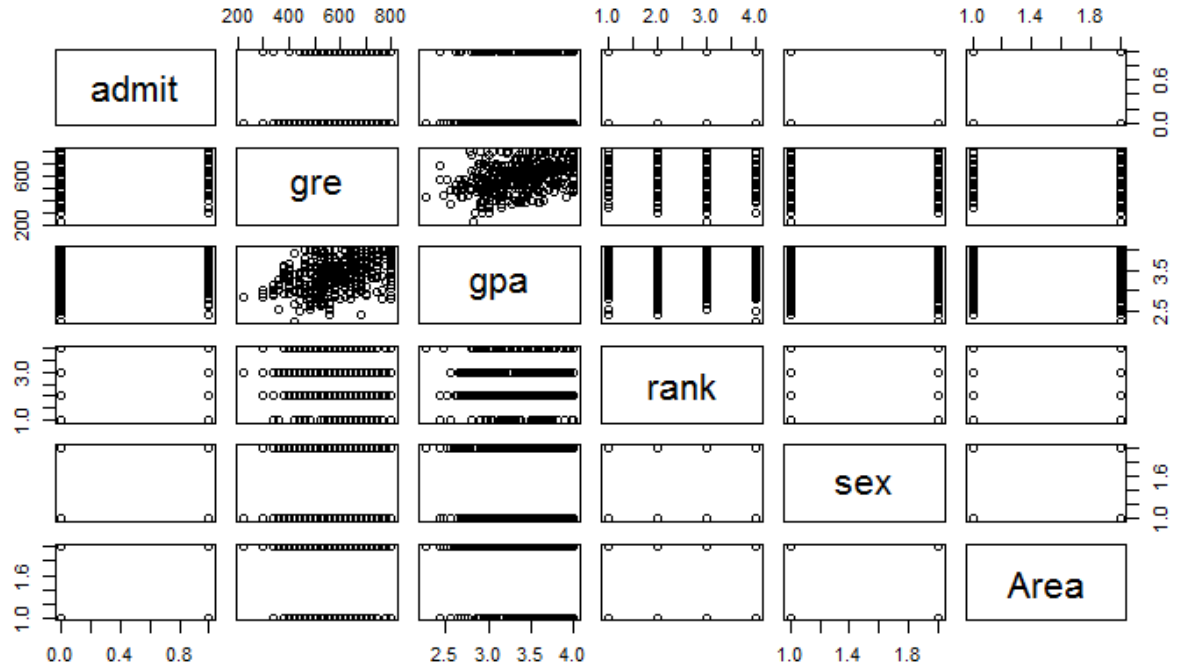
```
##
## Call:
## glm(formula = admit ~ gre + gpa + rank, family = binomial(link = "logit"),
##     data = dat)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6268  -0.8662  -0.6388   1.1490   2.0790
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.989979   1.139951  -3.500 0.000465 ***
## gre          0.002264   0.001094   2.070 0.038465 *
## gpa          0.804038   0.331819   2.423 0.015388 *
## rank2       -0.675443   0.316490  -2.134 0.032829 *
## rank3       -1.340204   0.345306  -3.881 0.000104 ***
## rank4       -1.551464   0.417832  -3.713 0.000205 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 499.98  on 399  degrees of freedom
## Residual deviance: 458.52  on 394  degrees of freedom
## AIC: 470.52
##
## Number of Fisher Scoring iterations: 4
```

可以看到删去sex和Area后回归系数均是显著的，且mylogit0的回归结果比起mylogit没有太大变化。

综上所述，可以认为sex和Area对admit没有影响。

2.2 判断多重共线性性

2.2.1 画散点图粗略判断



分类型变量之间的关系不易从图中得出，但可从图中看出连续型变量gre与gpa之间可能有一定的线性关系。

2.2.2 用条件数考查多重共线性性

- 将分类变量与连续型变量分开考虑

```
#分开考虑分类变量和连续型变量
X <- model.matrix( ~ gpa + gre + rank + sex + Area, dat)
con <- c(1,2,3)
dum <- c(1,4,5,6,7,8)
Y <- X[,con]
Z <- X[,dum]
```

- 用条件数考查连续变量之间的多重共线性性

```
#连续变量之间
lambda <- eigen(t(Y)%*%Y)
kY <- max(lambda$values) / min(lambda$values)
kY
```

```
## [1] 32096770
```

```
aY <- lambda$vectors[,which.min(lambda$values)]
aY
```

```
## [1]  0.9668461215 -0.2553596556 -0.0001538148
```

条件数$kY >> 1000$，可见连续变量之间存在严重的共线性，从其最小特征值对应的特征向量可以看出截距项与gre，gpa之间可能存在共线性:

$$0.9668 \times 1 - 0.2554\text{gre} - 0.0002\text{gpa} = 0$$

- 用条件数考查分类变量之间的多重共线性性

```
#分类变量之间
mu <- eigen(t(Z)%*%Z)
kZ <- max(mu$values) / min(mu$values)
kZ
```

```
## [1] 63.00134
```

```
aZ <- mu$vectors[,which.min(mu$values)]
aZ
```

```
## [1]  0.49169790 -0.47761197 -0.48392927 -0.53980057 -0.03319166 -0.05870350
```

条件数$kZ < 100$，可见分类变量之间共线性程度很小。

- 尝试消除连续变量之间的多重共线性性

```
#检验截距项与gre,gpa之间的共线性
c <- X[,1]
ltest1 <- lm(c ~ 0 + gre + gpa, data = dat)
summary(ltest1)
```

```
##
## Call:
## lm(formula = c ~ 0 + gre + gpa, data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.18744 -0.07229  0.01090  0.08875  0.33557
##
## Coefficients:
##      Estimate Std. Error t value Pr(>|t|)
## gre 0.0002023  0.0000504   4.013 7.15e-05 ***
## gpa 0.2564076  0.0088486  28.977  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1095 on 398 degrees of freedom
## Multiple R-squared:  0.9881, Adjusted R-squared:  0.988
## F-statistic: 1.648e+04 on 2 and 398 DF,  p-value: < 2.2e-16
```

由回归结果知系数均为显著的，可能的线性模型为

$$1 = 0.0002\text{gre} + 0.2564\text{gpa}$$

该结果与用条件数检验出的共线性结果

$$0.9668 \times 1 - 0.2554\text{gre} - 0.0002\text{gpa} = 0$$

十分相近，故可以认为三者之间是用共线性的，从而可以删去截距项做进一步回归。

```
#删去截距项后回归
Y1 <- Y[,-1]
lambda1 <- eigen(t(Y1)%*%Y1)
kY1 <- max(lambda1$values) / min(lambda1$values)
kY1
```

```
## [1] 936809.9
```

```
aY1 <- lambda1$vectors[,which.min(lambda1$values)]
aY1
```

```
## [1] -0.999984315  0.005600932
```

由以上结果知删去截距项回归后条件数依然远大于1000，即gre与gpa之间可能存在共线性性。

```
#检验gre与gpa之间的共线性性
ltest2 <- lm(gpa ~ 0 + gre, data = dat)
summary(ltest2)
```

```
##
## Call:
## lm(formula = gpa ~ 0 + gre, data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.70081 -0.29255  0.09136  0.47737  1.59778
##
## Coefficients:
##      Estimate Std. Error t value Pr(>|t|)
## gre 5.601e-03  5.173e-05   108.3   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6196 on 399 degrees of freedom
## Multiple R-squared:  0.9671, Adjusted R-squared:  0.967
## F-statistic: 1.173e+04 on 1 and 399 DF,  p-value: < 2.2e-16
```

该检验结果说明gre与gpa之间确实存在显著的线性关系，故可以删去任意一个变量做进一步回归。

- 消除共线性与无关变量后重新回归

```
#消除共线性与无关变量后重新回归
newlogit <- glm(admit ~ gpa + rank, data = dat, binomial(link = 'logit'))
summary(newlogit)
```

```
##
## Call:
## glm(formula = admit ~ gpa + rank, family = binomial(link = "logit"),
##     data = dat)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5055  -0.8663  -0.6590   1.1505   2.0913
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -3.4636     1.1003  -3.148 0.001645 **
## gpa           1.0521     0.3102   3.392 0.000694 ***
## rank2        -0.6810     0.3141  -2.168 0.030181 *
## rank3        -1.3919     0.3419  -4.071 4.68e-05 ***
## rank4        -1.5943     0.4152  -3.840 0.000123 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 499.98  on 399  degrees of freedom
## Residual deviance: 462.88  on 395  degrees of freedom
## AIC: 472.88
##
## Number of Fisher Scoring iterations: 4
```

以上newlogit回归结果中变量均显著，且AIC值略小于最初mylogti模型的AIC值。

2.3 建立新的线性模型并进行回归诊断

2.3.1 新的响应变量与预测变量

将gpa作为响应变量，其余变量作为预测变量。
将新加入的预测变量admit转化为哑变量。

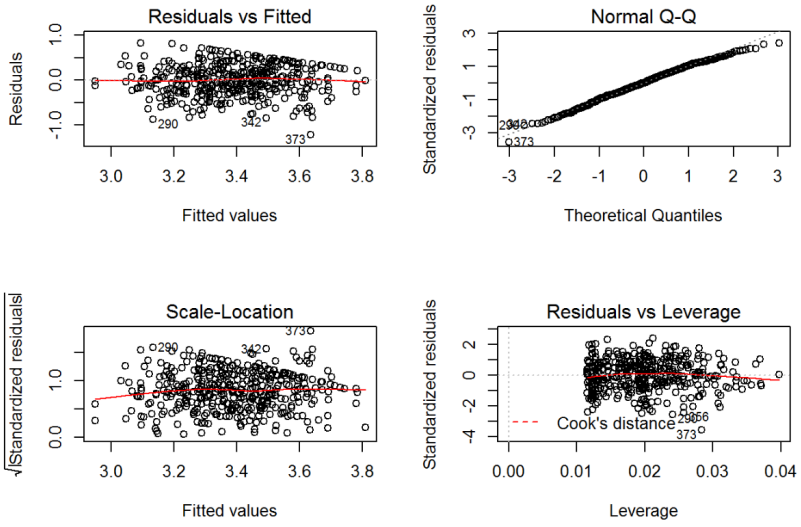
```
#新的响应变量与预测变量
dat1 <- dat
dat1$admit <- factor(dat1$admit)
```

2.3.2 以gpa为响应变量进行线性回归

```
#以gpa为响应变量进行线性回归
ll <- lm(gpa ~ admit + gre + rank + sex + Area, data = dat1)
summary(ll)
```

```
##
## Call:
## lm(formula = gpa ~ admit + gre + rank + sex + Area, data = dat1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.21578 -0.22779  0.00355  0.24818  0.82593
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.6807380  0.1073864  24.963  < 2e-16 ***
## admit1       0.0919966  0.0391045   2.353   0.0191 *
## gre          0.0012102  0.0001539   7.866  3.6e-14 ***
## rank2       -0.0576830  0.0532089  -1.084   0.2790
## rank3        0.0536100  0.0560250   0.957   0.3392
## rank4       -0.0445219  0.0632471  -0.704   0.4819
## sexmale      0.0400816  0.0349642   1.146   0.2523
## AreaC       -0.0504463  0.0393540  -1.282   0.2006
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3473 on 392 degrees of freedom
## Multiple R-squared:  0.1816, Adjusted R-squared:  0.167
## F-statistic: 12.42 on 7 and 392 DF,  p-value: 2.166e-14
```

```
par(mfrow=c(2,2))
plot(ll)
```



可见该线性模型中，rank,sex,Area变量均不显著，拟合优度很小。
从图中还可看到拟合值与残差值之间具有一定的相关性，说明该回归模型还需改进。

2.3.3 用方差膨胀因子考查多重共线性性

```
#用方差膨胀因子考查多重共线性
vif(11)
```

```
##           GVIF Df GVIF^(1/(2*Df))
## admit 1.098578  1        1.048130
## gre   1.044728  1        1.022119
## rank  1.088478  3        1.014230
## sex   1.007553  1        1.003770
## Area  1.012034  1        1.005999
```

从方差膨胀因子的结果中并未观察到预测变量之间有明显的共线性性，但其实这并不能说明预测变量与截距项之间是否存在共线性性。
实际上，在该模型中只有gre是连续型变量，剩余的都是分类型变量，而在上一题中已经判断过，这些分类型变量之间（包括截距项）的共线性程度很小。故该模型中暂时没有发现需要消除的共线性性。

2.3.4 变量的显著性检验

```
#变量的显著性检验
wald.test(b = coef(11), Sigma = vcov(11), Terms = 4:6)
```

```
## Wald test:
## ----------
##
## Chi-squared test:
## X2 = 7.5, df = 3, P(> X2) = 0.057
```

该检验进一步说明了rank变量在该模型中是不显著的。

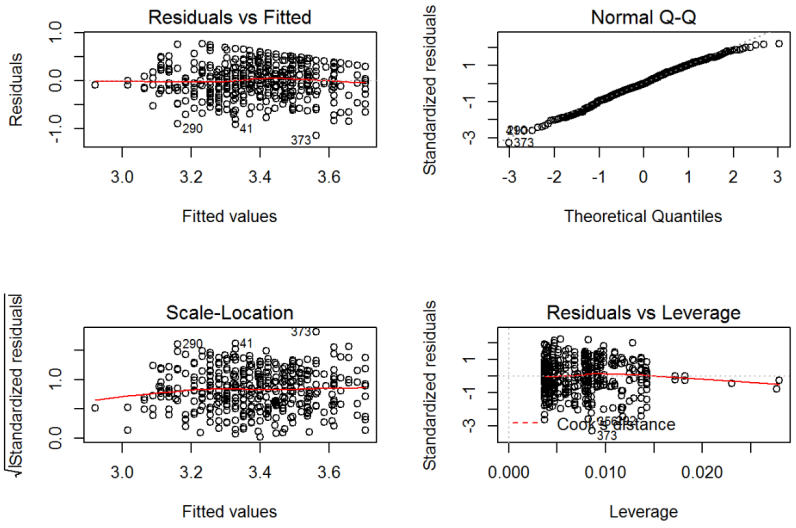
2.3.5 向后回归法删去不显著的变量

通过以上结果知，rank,sex,Area这三个变量在该回归模型中均不显著，故可用向后回归法的思想，删去这三个变量，作进一步的线性回归：

```
#向后回归法删去不显著的变量
12 <- lm(gpa ~ admit + gre, data = dat1)
summary(12)
```

```
##
## Call:
## lm(formula = gpa ~ admit + gre, data = dat1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.14244 -0.21935  0.00115  0.25494  0.76798
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.6567711  0.0908985  29.228  < 2e-16 ***
## admit1      0.0907314  0.0382235   2.374   0.0181 *
## gre         0.0011984  0.0001542   7.771 6.75e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3498 on 397 degrees of freedom
## Multiple R-squared:  0.1596, Adjusted R-squared:  0.1554
## F-statistic: 37.69 on 2 and 397 DF,  p-value: 1.027e-15
```

```
par(mfrow=c(2,2))
plot(12)
```



删去三个变量后，l2的回归结果与l1的回归结果没有显著的差异，也可进一步说明这三个变量在该模型中对gpa无影响。

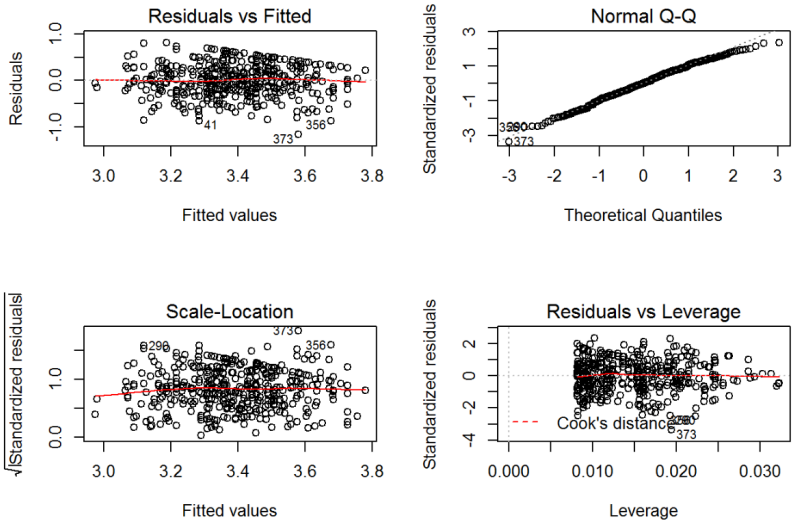
2.3.6 用AIC和BIC作变量选择

- 用AIC作变量选择

```
#用AIC作变量选择
lm.aic <- step(lm(gpa ~ admit + gre + rank + sex + Area, data = dat1), trace = F)
summary(lm.aic)
```

```
##
## Call:
## lm(formula = gpa ~ admit + gre + rank, data = dat1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.15947 -0.22492  0.00266  0.25623  0.81307
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.6596263  0.1032709  25.754  < 2e-16 ***
## admit1       0.0950650  0.0390943   2.432   0.0155 *
## gre          0.0012129  0.0001539   7.879 3.24e-14 ***
## rank2       -0.0548915  0.0532524  -1.031   0.3033
## rank3        0.0539972  0.0560317   0.964   0.3358
## rank4       -0.0498321  0.0632516  -0.788   0.4313
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3478 on 394 degrees of freedom
## Multiple R-squared:  0.1753, Adjusted R-squared:  0.1648
## F-statistic: 16.75 on 5 and 394 DF,  p-value: 5.21e-15
```

```
par(mfrow=c(2,2))
plot(lm.aic)
```



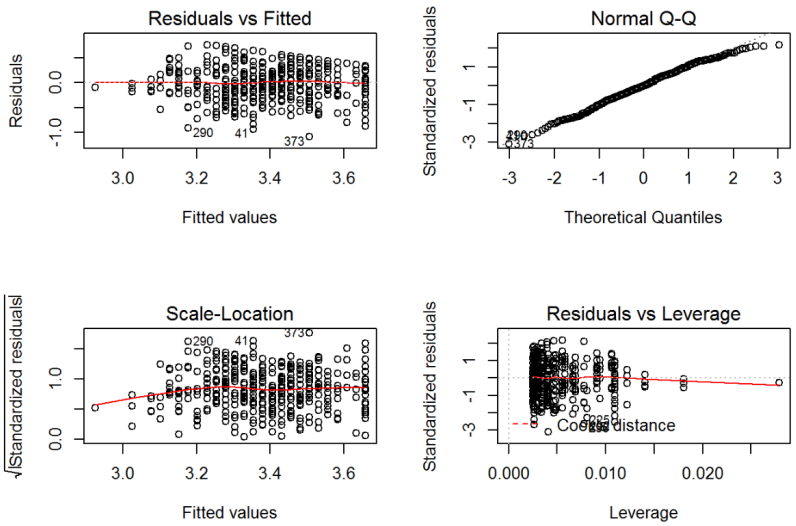
可见AIC选取了变量admit,gre,rank进行了有截距项的回归，其中rank的回归系数并不显著。从图中还可以看出残差与拟合值之间有轻微的相关性，接下来用BIC作更精确的选择。

- 用BIC作变量选择

```
#用BIC作变量选择
lm.bic <- step(lm(gpa ~ admit + gre + rank + sex + Area, data = dat1), k = log(400), trace = F)
summary(lm.bic)
```

```
##
## Call:
## lm(formula = gpa ~ gre, data = dat1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.08675 -0.22435 -0.00015  0.24809  0.76176
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.6458978  0.0913100  28.977  < 2e-16 ***
## gre         0.0012660  0.0001525   8.304  1.6e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3518 on 398 degrees of freedom
## Multiple R-squared:  0.1477, Adjusted R-squared:  0.1455
## F-statistic: 68.95 on 1 and 398 DF,  p-value: 1.596e-15
```

```
par(mfrow=c(2,2))
plot(lm.bic)
```

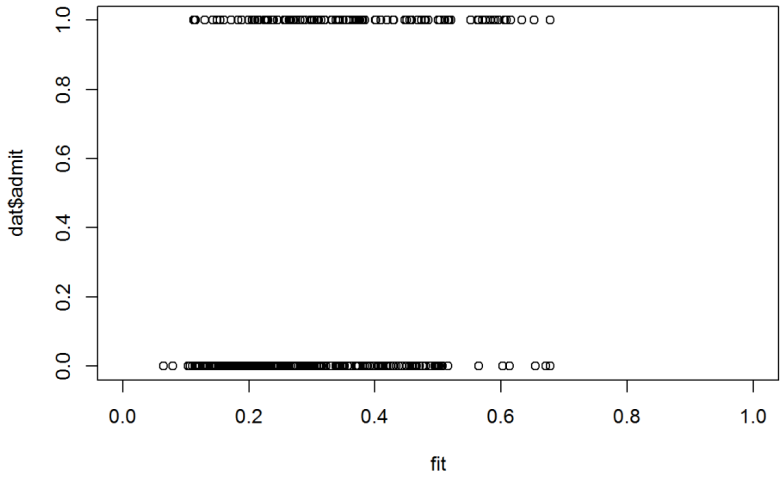


可见BIC只选取了变量gre进行了有截距项的回归，回归系数均为显著的。该结果也与上一题中所判断出的“gpa与gre存在线性关系”一致。从图中还可以看出残差与拟合值无明显的相关性，残差的分布基本符合正态性假设，故该回归结果是良好的。

2.4 思考

如何像线性模型中的响应图和残差图一样，对logit线性模型用图形方法诊断模型拟合的好坏。
以第一题中的newlogit模型为例：

```
#诊断newlogit模型
fit <- newlogit$fitted.values
plot(fit, dat$admit, xlim = 0:1, ylim = 0:1)
```



理想的图应该是当$admit = 0$时，fit的值集中在$(0, 0.5)$区间中；当$admit = 1$时，fit的值集中在$(0.5, 1)$区间中。
由此可见newlogit的模型在使得$admit = 0$的点上拟合得较好，在使得$admit = 1$的点上拟合得不好。