hw3

顾格非 3210103528

1

a. 这时候$(x-\xi)^3=0$，所以$a_1=\beta_0, b_1=\beta_1, c_1=\beta_2, d_1=\beta_3$

b. 这时候$f(x)=\beta_0+\beta_1 x+\beta_2 x^2+\beta_3 x^3+\beta_4(x-\xi)^3$

化简得$f(x)=(\beta_4+\beta_3)x^3+(\beta_2-3\beta_4\xi)x^2+(\beta_1+3\beta_4\xi^2)x+\beta_4\xi^3+\beta_0$

$a_1=\beta_4\xi^3+\beta_0, b_1=\beta_1+3\beta_4\xi^2, c_1=\beta_1+3\beta_4\xi^2, d_1=\beta_4+\beta_3$

c. $f_1(\xi)=\beta_0+\beta_1\xi+\beta_2\xi^2+\beta_3\xi^3$

$f_2(\xi)=(\beta_0-\beta_4\xi^3)+(\beta_1+3\xi^2\beta_4)\xi+(\beta_2-3\beta_4\xi)\xi^2+(\beta_3+\beta_4)\xi^3=\beta_0+\beta_1\xi+\beta_2\xi^2+\beta_3\xi^3.$

所以$f_1(\xi)=f_2(\xi)$，即$f(x)$在ξ连续。

d. $f_1'(\xi)=\beta_1+2\beta_2\xi+3\beta_3\xi^2$

$f_2'(\xi)=\beta_1+3\xi^2\beta_4+2(\beta_2-3\beta_4\xi)\xi+3(\beta_3+\beta_4)\xi^2=\beta_1+2\beta_2\xi+3\beta_3\xi^2.$

所以$f_1'(\xi)=f_2'(\xi)$，即$f'(x)$在ξ连续。

e.$f_1''(\xi)=2\beta_2+6\beta_3\xi$

$f_2''(\xi)=2(\beta_2-3\beta_4\xi)+6(\beta_3+\beta_4)\xi=2\beta_2+6\beta_3\xi.$

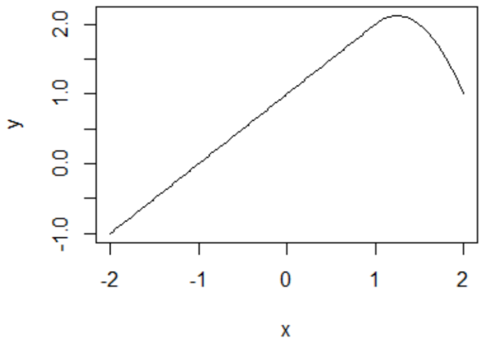
所以$f_1''(\xi)=f_2'(\xi)$，即$f''(x)$在ξ连续。

2

- $\lambda=\infty, m=0$：此时第二项的比重无穷大，所以$g(x)=0$
- $\lambda=\infty, m=1$：此时第二项的比重无穷大，所以$g'(x)=0\Rightarrow g(x)=c$, 所以要使 $h(c)=\sum(y_i-c)^2$最小，求导得$h'(c)=2\sum(c-y_i)=0$, 于是解得$g(x)=\frac{\sum_{i=1}^{n}y_i}{n}$
- $\lambda=\infty, m=2$：此时第二项的比重无穷大，所以$g''(x)=0\Rightarrow g'(x)=c\Rightarrow g(x)=cx+d$, 此时就是普通的线性回归，$c=\frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sum_{i=1}^{n}(x_i-\overline{x})^2}$,$d=\overline{y}-c\overline{x}$
- $\lambda=\infty, m=3$：此时第二项的比重无穷大，所以 $g''(x)=0\Rightarrow g''(x)=c\Rightarrow g'(x)=cx,\Rightarrow g(x)=\frac{c}{2}x^2+bx+d$ 所以是形如二次方程的拟合。
- $\lambda=0, m=3$ 这时候惩罚项不起作用，就是普通的线性拟合,拟合方程同（3）时。

3

在[-2,1]时，$Y=1+X+\epsilon$, 在[1,2]时，$Y=1+X+(X-1)^2$, 画出草图如下：



5

a. 由于惩罚项的作用很大，所以g_1和g_2分别是二次和三次多项式，$\hat{g}_2$

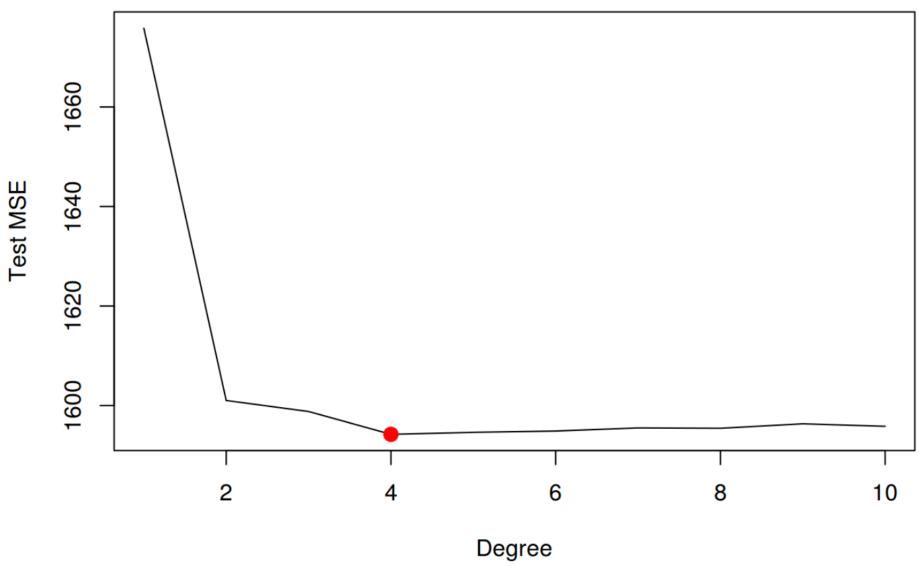
次数更高，能更好的拟合训练数据，所以training RSS更加小。

b. $\hat{g}_2$可能出现过拟合的问题，所以可能$\hat{g}_1$可能有更好的testing RSS（不过我觉的跟具体数据有关，具体情况具体考虑比较好）

c. 此时惩罚项不起作用，所以两个拟合是一样。

6

```
library(ISLR)
library(boot)
set.seed(1)
deltas <- rep(NA, 10)
for (i in 1:10) {
    fit <- glm(wage ~ poly(age, i), data = Wage)
    deltas[i] <- cv.glm(Wage, fit, K = 10)$delta[1]
}
plot(1:10, deltas, xlab = "Degree", ylab = "Test MSE", type = "l")
d.min <- which.min(deltas)
points(which.min(deltas), deltas[which.min(deltas)], col = "red", cex = 2, pch = 20)
```



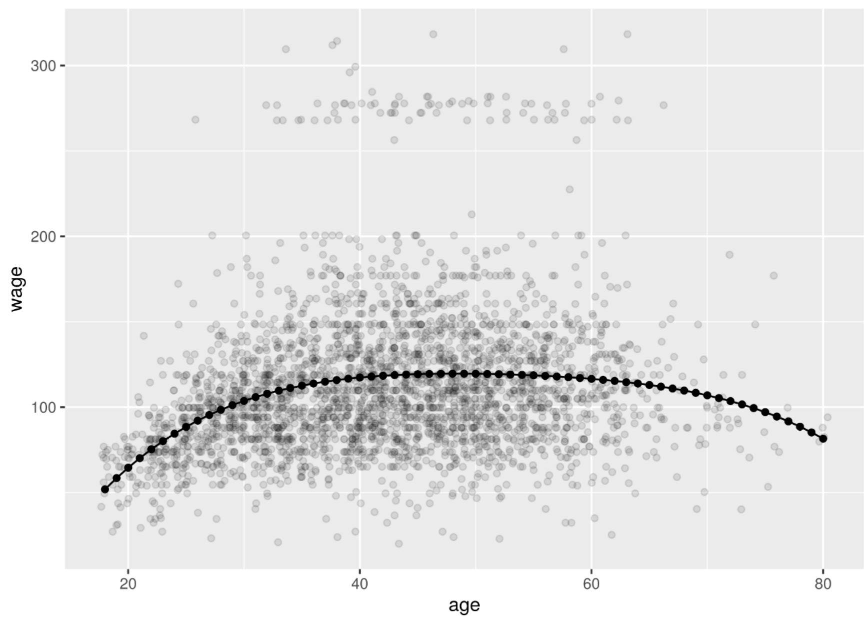
```
fit1 <- lm(wage ~ age, data = Wage)
fit2 <- lm(wage ~ poly(age, 2), data = Wage)
fit3 <- lm(wage ~ poly(age, 3), data = Wage)
fit4 <- lm(wage ~ poly(age, 4), data = Wage)
fit5 <- lm(wage ~ poly(age, 5), data = Wage)
anova(fit1, fit2, fit3, fit4, fit5)
```

```
## Analysis of Variance Table
##
## Model 1: wage ~ age
## Model 2: wage ~ poly(age, 2)
## Model 3: wage ~ poly(age, 3)
## Model 4: wage ~ poly(age, 4)
## Model 5: wage ~ poly(age, 5)
##   Res.Df     RSS Df Sum of Sq        F    Pr(>F)
## 1   2998 5022216
## 2   2997 4793430  1    228786 143.5931 < 2.2e-16 ***
## 3   2996 4777674  1     15756   9.8888  0.001679 **
## 4   2995 4771604  1      6070   3.8098  0.051046 .
## 5   2994 4770322  1      1283   0.8050  0.369682
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

从图一中我们发现4个自由度比较合适，再运用ANOVA假设检验发现4次很好的拟合了模型，我们用4次拟合：

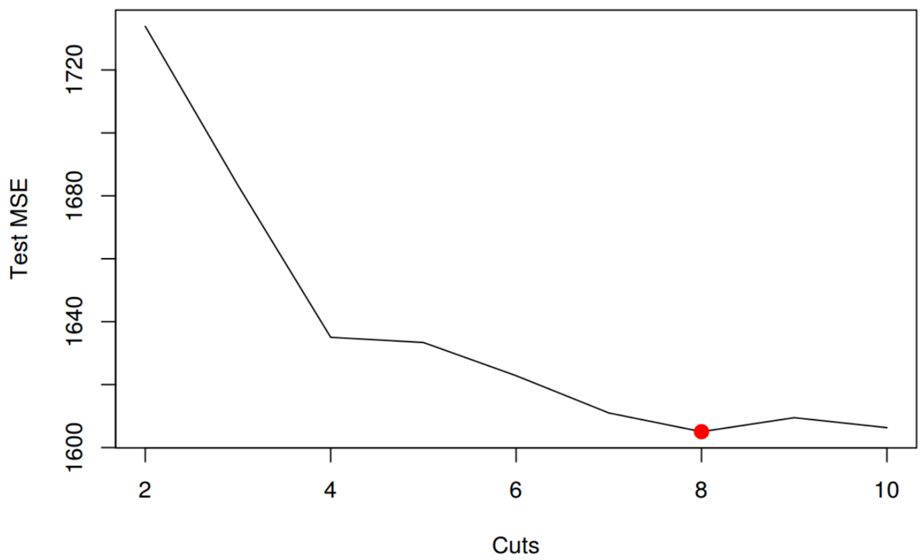
```
wage_model <- lm(wage ~ poly(age, 4), data = Wage)

tibble(age = range(wage$age)[1]:range(wage$age)[2]) %>%
    mutate(wage = predict(wage_model, newdata = .)) %>%
    ggplot(aes(age, wage)) +
    geom_jitter(data = wage, mapping = aes(age, wage), alpha = .1) +
    geom_line() +
    geom_point()
```



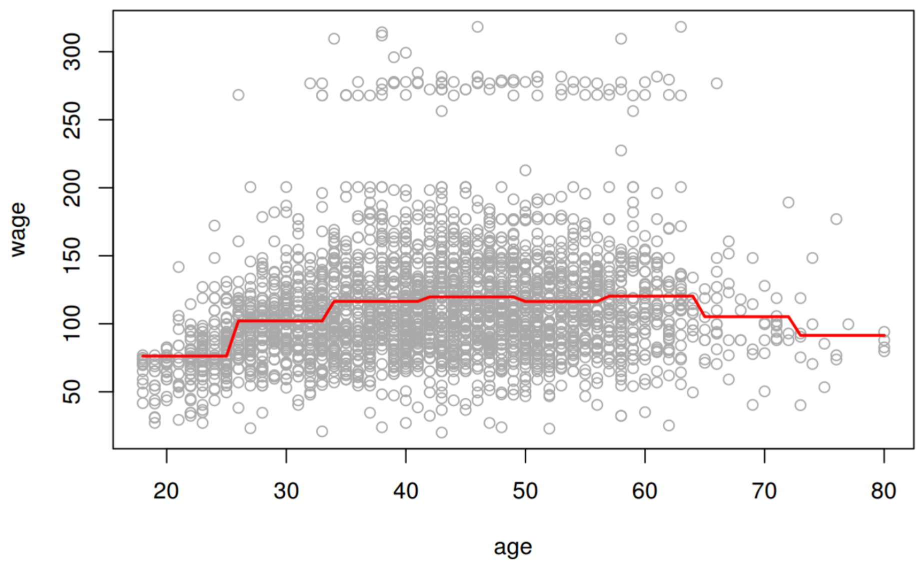
(b)

```
cvs <- rep(NA, 10)
for (i in 2:10) {
    Wage$age.cut <- cut(Wage$age, i)
    fit <- glm(wage ~ age.cut, data = Wage)
    cvs[i] <- cv.glm(Wage, fit, K = 10)$delta[1]
}
plot(2:10, cvs[-1], xlab = "Cuts", ylab = "Test MSE", type = "l")
d.min <- which.min(cvs)
points(which.min(cvs), cvs[which.min(cvs)], col = "red", cex = 2, pch = 20)
```



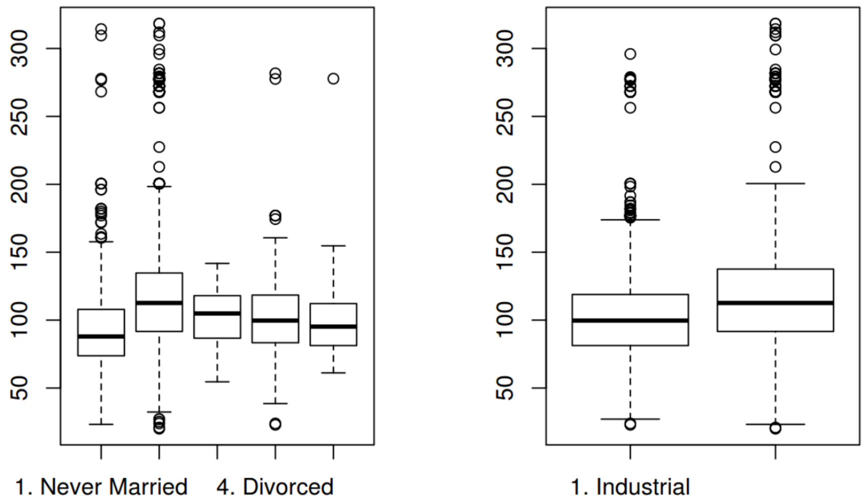
用k=10的交叉验证得到cuts=8比较合适，用阶梯函数拟合如下：

```
plot(wage ~ age, data = Wage, col = "darkgrey")
agelims <- range(Wage$age)
age.grid <- seq(from = agelims[1], to = agelims[2])
fit <- glm(wage ~ cut(age, 8), data = Wage)
preds <- predict(fit, data.frame(age = age.grid))
lines(age.grid, preds, col = "red", lwd = 2)
```



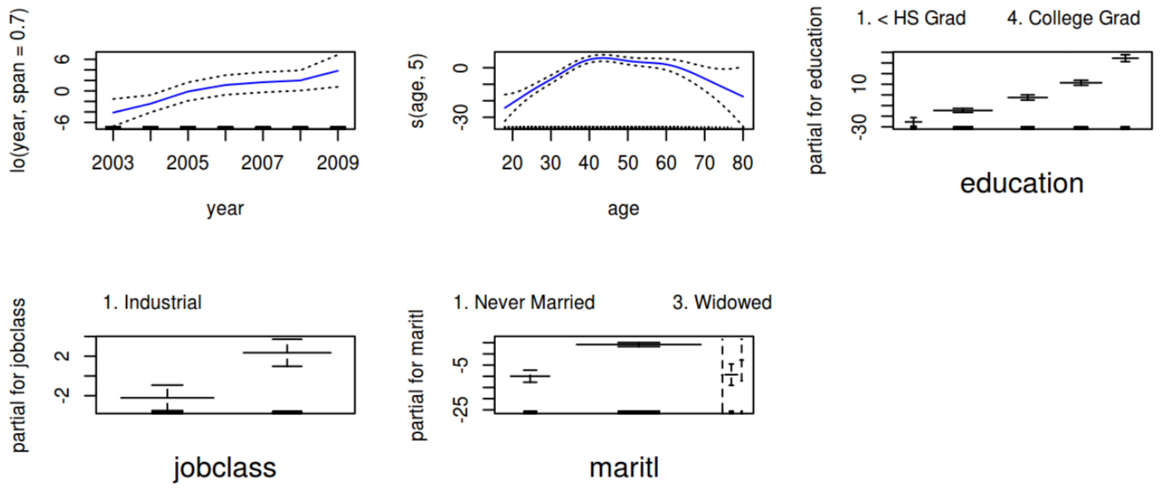
7

```
par(mfrow = c(1, 2))
plot(Wage$maritl, Wage$wage)
plot(Wage$jobclass, Wage$wage)
```




```
library(gam)
fit0 <- gam(wage ~ lo(year, span = 0.7) + s(age, 5) + education, data = Wage)
fit1 <- gam(wage ~ lo(year, span = 0.7) + s(age, 5) + education + jobclass, data = Wage)
fit2 <- gam(wage ~ lo(year, span = 0.7) + s(age, 5) + education + maritl, data = Wage)
fit3 <- gam(wage ~ lo(year, span = 0.7) + s(age, 5) + education + jobclass + maritl,
data = Wage)
anova(fit0, fit1, fit2, fit3)
par(mfrow = c(3, 3))
plot(fit3, se = T, col = "blue")
```

```
## Analysis of Deviance Table
##
## Model 1: wage ~ lo(year, span = 0.7) + s(age, 5) + education
## Model 2: wage ~ lo(year, span = 0.7) + s(age, 5) + education + jobclass
## Model 3: wage ~ lo(year, span = 0.7) + s(age, 5) + education + maritl
## Model 4: wage ~ lo(year, span = 0.7) + s(age, 5) + education + jobclass +
##     maritl
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1    2987.1    3691855
## 2    2986.1    3679689  1    12166 0.0014637 **
## 3    2983.1    3597526  3    82163  9.53e-15 ***
## 4    2982.1    3583675  1    13852 0.0006862 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

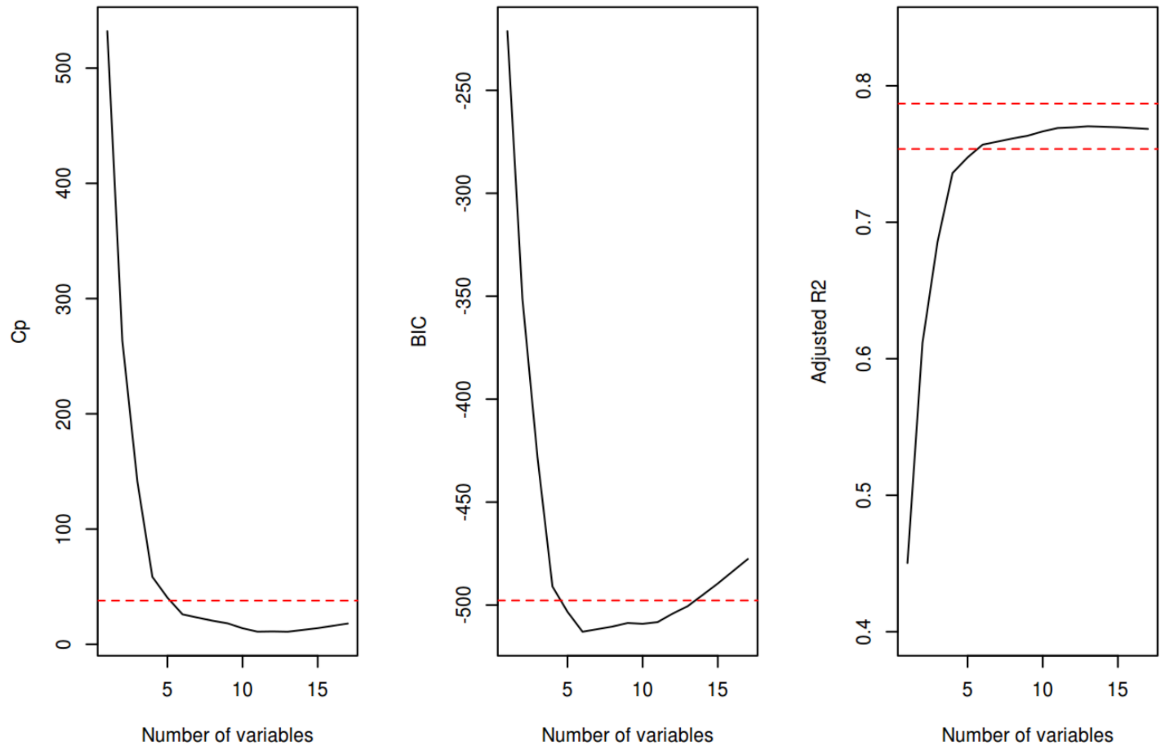


从上面的结果看，已婚的人和从事information工作的人有更高比例的人有更高的工资。

10

(a)

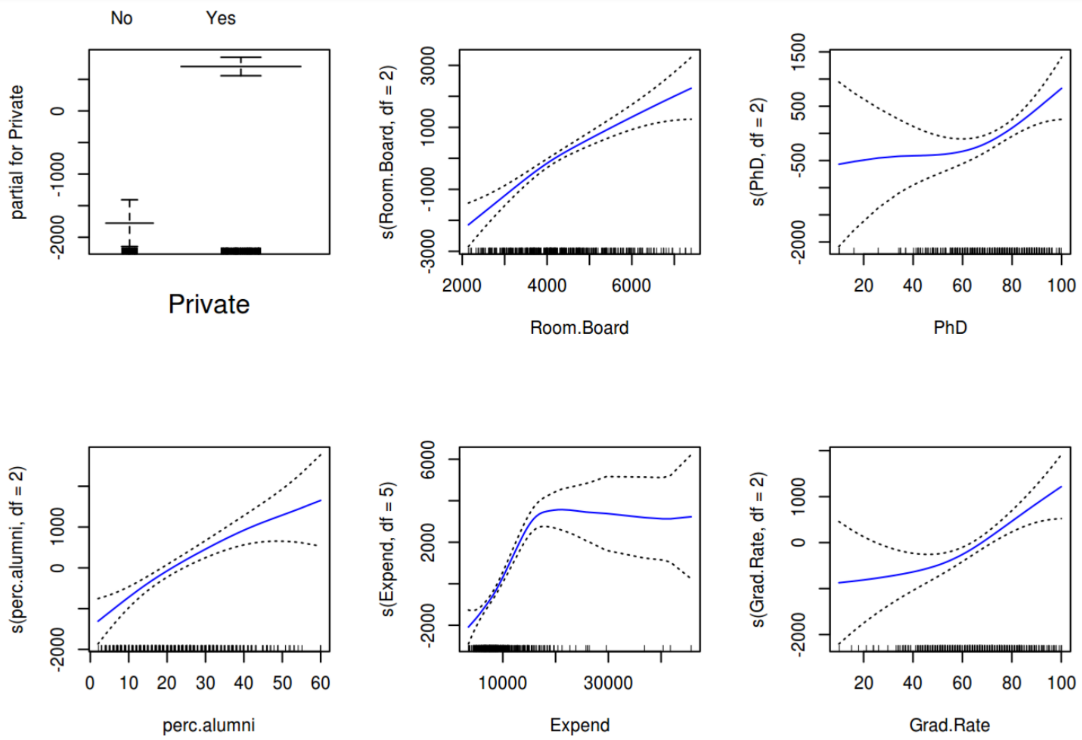
```
library(leaps)
set.seed(1)
attach(College)
train <- sample(length(Outstate), length(Outstate) / 2)
test <- -train
College.train <- College[train, ]
College.test <- College[test, ]
fit <- regsubsets(Outstate ~ ., data = College.train, nvmax = 17, method = "forward")
fit.summary <- summary(fit)
par(mfrow = c(1, 3))
plot(fit.summary$cp, xlab = "Number of variables", ylab = "Cp", type = "l")
min.cp <- min(fit.summary$cp)
std.cp <- sd(fit.summary$cp)
abline(h = min.cp + 0.2 * std.cp, col = "red", lty = 2)
abline(h = min.cp - 0.2 * std.cp, col = "red", lty = 2)
plot(fit.summary$bic, xlab = "Number of variables", ylab = "BIC", type='l')
min.bic <- min(fit.summary$bic)
std.bic <- sd(fit.summary$bic)
abline(h = min.bic + 0.2 * std.bic, col = "red", lty = 2)
abline(h = min.bic - 0.2 * std.bic, col = "red", lty = 2)
plot(fit.summary$adjr2, xlab = "Number of variables", ylab = "Adjusted R2", type = "l",
ylim = c(0.4, 0.84))
max.adjr2 <- max(fit.summary$adjr2)
std.adjr2 <- sd(fit.summary$adjr2)
abline(h = max.adjr2 + 0.2 * std.adjr2, col = "red", lty = 2)
abline(h = max.adjr2 - 0.2 * std.adjr2, col = "red", lty = 2)
```



从图上看，选择6个变量比较合适。

(b)

```
fit <- gam(Outstate ~ Private + s(Room.Board, df = 2) + s(PhD, df = 2) + s(perc.alumni,
df = 2) + s(Expend, df = 5) + s(Grad.Rate, df = 2), data=College.train)
par(mfrow = c(2, 3))
plot(fit, se = T, col = "blue")
```



(c)

```
preds <- predict(fit, College.test)
err <- mean((College.test$Outstate - preds)^2)
err
```

```
tss <- mean((College.test$Outstate - mean(College.test$Outstate))^2)
rss <- 1 - err / tss
rss
```

得到err是3745460的RSS是0.7696916.相比线性model，确实非线性的拟合效果更好（在test集合上表现更好）。

(d) `summary(fit)` 得到如下：

```
##
## Call: gam(formula = Outstate ~ Private + s(Room.Board, df = 2) + s(PhD,
##     df = 2) + s(perc.alumni, df = 2) + s(Expend, df = 5) + s(Grad.Rate,
##     df = 2), data = College.train)
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -4977.74 -1184.52    58.33  1220.04  7688.30
##
## (Dispersion Parameter for gaussian family taken to be 3300711)
##
##     Null Deviance: 6221998532 on 387 degrees of freedom
## Residual Deviance: 1231165118 on 373 degrees of freedom
## AIC: 6941.542
##
## Number of Local Scoring Iterations: 2
##
## Anova for Parametric Effects
##                          Df     Sum Sq    Mean Sq F value    Pr(>F)
## Private                   1 1779433688 1779433688 539.106 < 2.2e-16 ***
## s(Room.Board, df = 2)     1 1221825562 1221825562 370.171 < 2.2e-16 ***
## s(PhD, df = 2)            1  382472137  382472137 115.876 < 2.2e-16 ***
## s(perc.alumni, df = 2)    1  328493313  328493313  99.522 < 2.2e-16 ***
## s(Expend, df = 5)         1  416585875  416585875 126.211 < 2.2e-16 ***
## s(Grad.Rate, df = 2)      1   55284580   55284580  16.749 5.232e-05 ***
## Residuals               373 1231165118    3300711
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Anova for Nonparametric Effects
##                        Npar Df  Npar F     Pr(F)
## (Intercept)
## Private
## s(Room.Board, df = 2)        1  3.5562   0.06010 .
## s(PhD, df = 2)               1  4.3421   0.03786 *
## s(perc.alumni, df = 2)       1  1.9158   0.16715
## s(Expend, df = 5)            4 16.8636 1.016e-12 ***
## s(Grad.Rate, df = 2)         1  3.7208   0.05450 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

*“Outstate” and “Expend，Outstate" and “Grad.Rate*之间有较强的非线性关系。