Week 6

6.3 已知如下数据，由表 *6.19* 所示.

表 6.19: 数据表

序号	X	Y	序号	X	Y	序号	X	Y
1	1	0.6	11	4	3.5	21	8	17.5
2	1	1.6	12	4	4.1	22	8	13.4
3	1	0.5	13	4	5.1	23	8	4.5
4	1	1.2	14	5	5.7	24	9	30.4
5	2	2.0	15	6	3.4	25	11	12.4
6	2	1.3	16	6	9.7	26	12	13.4
7	2	2.5	17	6	8.6	27	12	26.2
8	3	2.2	18	7	4.0	28	12	7.4
9	3	2.4	19	7	5.5			
10	3	1.2	20	7	10.5			

(1) 画出数据的散点图，求回归直线 $y = \hat{\beta}_0 + \hat{\beta}_1 x$，同时将回归直线也画在散点图上；

(2) 分析 T 检验和 F 检验是否通过；

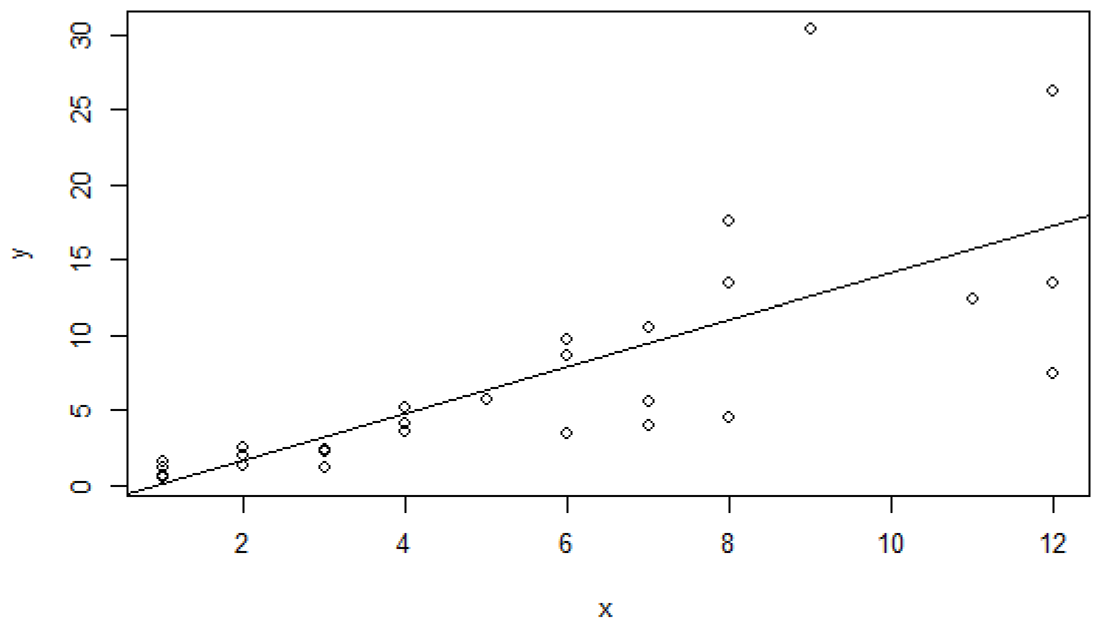
(3) 画出残差 *(*普通残差和标准化残差*)* 与预测值的残差图，分析误差是否是等方差的；

(4) 修正模型. 对响应变量 Y 作开方，再完成 *(1)–(3)* 的工作.

```
#(1)
d <- data.frame(
  x = c(1, 1, 1, 1, 2, 2, 2, 3, 3, 3, 4, 4, 4, 5, 6, 6, 6, 7, 7, 7, 8, 8, 8, 9, 11, 12,
12, 12),
  y = c(0.6, 1.6, 0.5, 1.2, 2, 1.3, 2.5, 2.2, 2.4, 1.2, 3.5, 4.1, 5.1, 5.7, 3.4, 9.7,
8.6, 4, 5.5, 10.5, 17.5, 13.4, 4.5, 30.4, 12.4, 13.4, 26.2, 7.4)
)

x = c(1, 1, 1, 1, 2, 2, 2, 3, 3, 3, 4, 4, 4, 5, 6, 6, 6, 7, 7, 7, 8, 8, 8, 9, 11, 12,
12, 12)
y = c(0.6, 1.6, 0.5, 1.2, 2, 1.3, 2.5, 2.2, 2.4, 1.2, 3.5, 4.1, 5.1, 5.7, 3.4, 9.7,
8.6, 4, 5.5, 10.5, 17.5, 13.4, 4.5, 30.4, 12.4, 13.4, 26.2, 7.4)
lm.sol <- lm(y~x)
summary(lm.sol)
```

```
plot(y~x)
abline(lm.sol)
```



#(2)

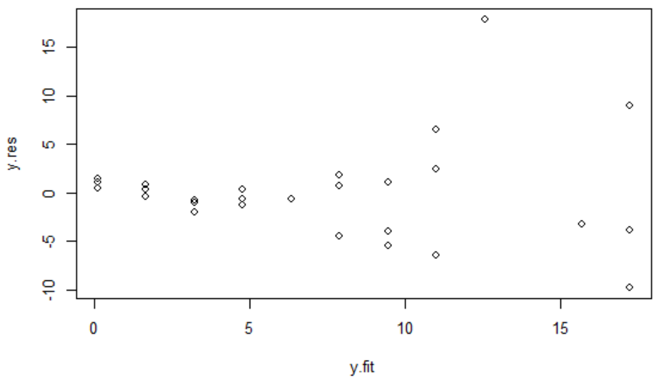
#P-值=0.436>0.05, Intercept 没有通过显著性检验(T 检验)

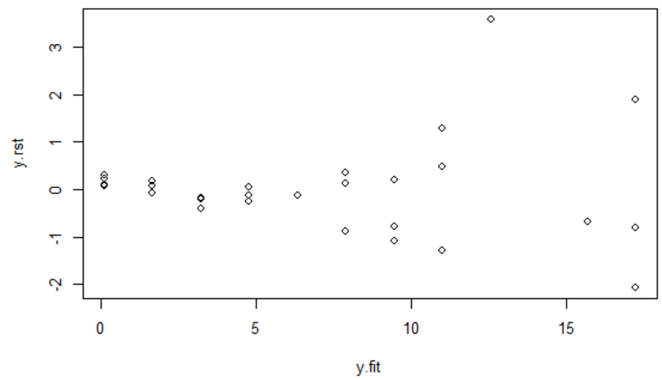
#P-值=7.93e-06<0.05, x 通过显著性检验(T 检验)

#P-值=7.931e-06<0.05, 方程整体通过 F 检验

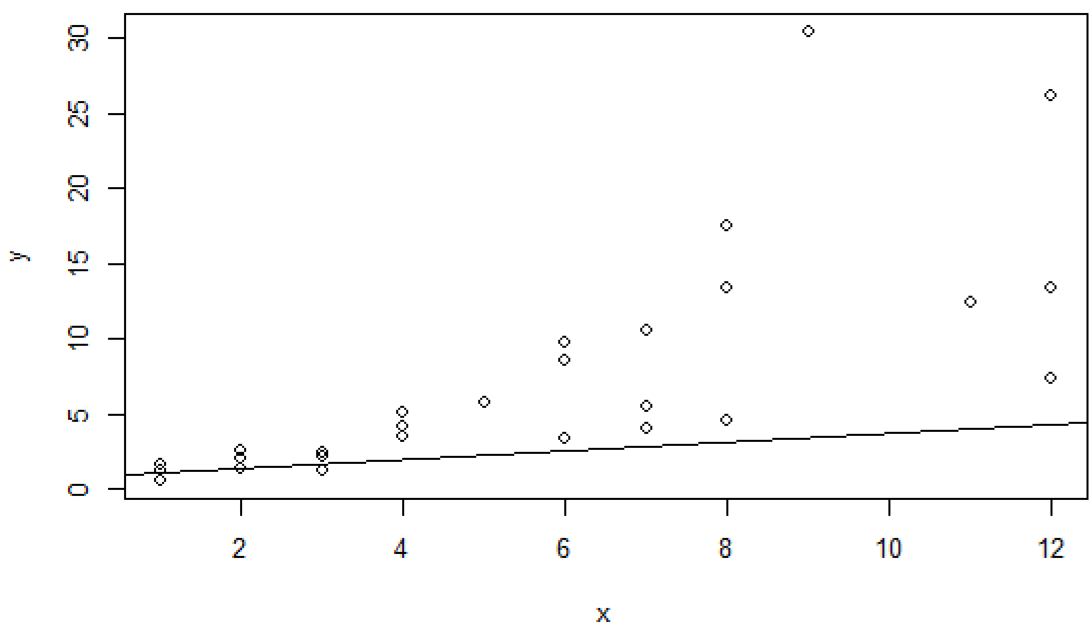
```
#(3)
y.res<-residuals(lm.sol)
y.fit<-predict(lm.sol)
y.rst<-rstandard(lm.sol)
plot(y.res~y.fit)
plot(y.rst~y.fit)
```

#从图形上看，误差并不呈现等方差,总体呈现喇叭口的样子

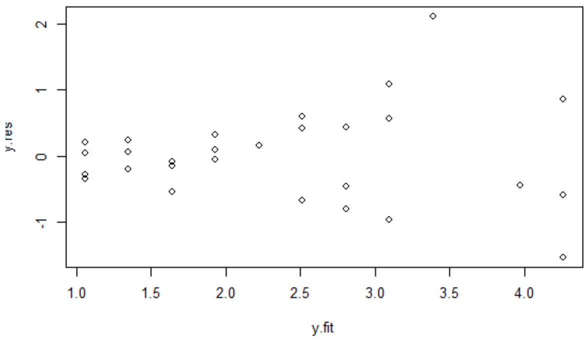


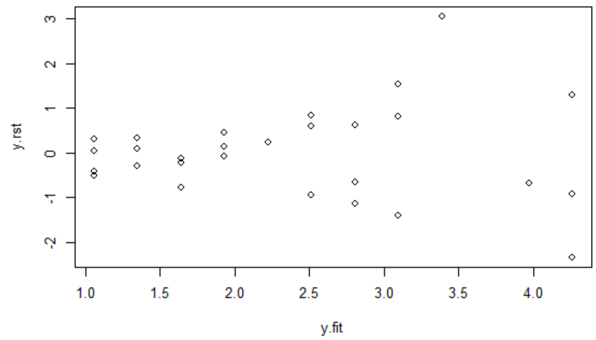



```
#(4)
lm.new<-update(lm.sol, sqrt(.)~.)
summary(lm.new)
plot(y~x)
abline(lm.new)
```




```
y.res<-residuals(lm.new)
y.fit<-predict(lm.new)
y.rst<-rstandard(lm.new)
plot(y.res~y.fit)
plot(y.rst~y.fit)
```





6.4 对牙膏销售数据 *(数据表见例 6.9)* 得到的线性模型作回归诊断，分析哪些样本点需要作进一步的研究？哪些样本点需要在回归计算中删去，如果有，删去再作线性回归模型的计算.

```
toothpaste<-data.frame(
  X1=c(-0.05, 0.25,0.60,0, 0.25,0.20, 0.15,0.05,-0.15, 0.15,
       0.20, 0.10,0.40,0.45,0.35,0.30, 0.50,0.50, 0.40,-0.05,
       -0.05,-0.10,0.20,0.10,0.50,0.60,-0.05,0, 0.05, 0.55),
  X2=c( 5.50,6.75,7.25,5.50,7.00,6.50,6.75,5.25,5.25,6.00,
        6.50,6.25,7.00,6.90,6.80,6.80,7.10,7.00,6.80,6.50,
        6.25,6.00,6.50,7.00,6.80,6.80,6.50,5.75,5.80,6.80),
  Y =c( 7.38,8.51,9.52,7.50,9.33,8.28,8.75,7.87,7.10,8.00,
        7.89,8.15,9.10,8.86,8.90,8.87,9.26,9.00,8.75,7.95,
        7.65,7.27,8.00,8.50,8.75,9.21,8.27,7.67,7.93,9.26)
)

attach(toothpaste)

lm.sol<-lm(Y ~ X1 + X2 + I(X2^2) + X1:X2, data=toothpaste)
summary(lm.sol)
```

```
Call:
lm(formula = Y ~ X1 + X2 + I(X2^2) + X1:X2, data = toothpaste)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.43725 -0.11754  0.00489  0.12263  0.38410 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  29.1133     7.4832   3.890 0.000656 ***
X1           11.1342     4.4459   2.504 0.019153 *  
X2           -7.6080     2.4691  -3.081 0.004963 ** 
I(X2^2)       0.6712     0.2027   3.312 0.002824 ** 
X1:X2        -1.4777     0.6672  -2.215 0.036105 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2063 on 25 degrees of freedom
Multiple R-squared:  0.9209, Adjusted R-squared:  0.9083 
F-statistic: 72.78 on 4 and 25 DF,  p-value: 2.107e-13
```

```
lm.sol1 <- lm(Y~X1, data=toothpaste)
lm.sol2 <- lm(Y~X2, data=toothpaste)

source("Reg_Diag.R")
Reg_Diag(lm.sol)
```

	residual	s1	standard	s2	student	s3	hat_matrix	s4
1	-0.044138552		-0.23265604		-0.22820261		0.15463537	
2	-0.122850229		-0.61322420		-0.60540498		0.05735116	
3	0.029866239		0.18453458		0.18092949		0.38476348	*
4	-0.074477573		-0.39099231		-0.38426937		0.14778110	
5	0.384096565		2.04801985	*	2.19962936	*	0.17386868	
6	-0.047249771		-0.23803612		-0.23349156		0.07455576	
7	0.233113134		1.18826106		1.19859261		0.09604602	
8	0.028687092		0.22678800		0.22243488		0.62418903	*
9	-0.066070954		-0.44760246		-0.44032698		0.48823240	*
10	0.029666169		0.15631113		0.15322789		0.15398180	
11	-0.437249771	*	-2.20278822	*	-2.40417108	*	0.07455576	
12	0.176312340		0.89613462		0.89248027		0.09080827	
13	0.035565924		0.18029703		0.17676925		0.08603908	
14	-0.138208648		-0.69508629		-0.68772049		0.07139811	
15	0.102676826		0.51131530		0.50362495		0.05288252	
16	0.126964214		0.63241124		0.62465059		0.05332686	
17	0.004773498		0.02513921		0.02463160		0.15314921	
18	-0.143454504		-0.73550774		-0.72857332		0.10650846	
19	-0.101610562		-0.50873578		-0.50105757		0.06302133	
20	0.005015975		0.02651619		0.02598082		0.15952423	
21	-0.038913808		-0.20579864		-0.20181169		0.16023258	
22	-0.133353406		-0.74668910		-0.73989998		0.25085624	
23	-0.327249771		-1.64862737		-1.71100300		0.07455576	
24	-0.327372794		-2.05501690	*	-2.20866875	*	0.40393868	*
25	-0.210185338		-1.08283163		-1.08674472		0.11504838	
26	0.141239886		0.76983816		0.76338697		0.20940800	
27	0.325015975		1.71814729		1.79259305		0.15952423	
28	0.109641375		0.56063690		0.55279575		0.10169845	
29	0.234223197		1.19732600		1.20829037		0.10118244	
30	0.245527274		1.29594900		1.31469330		0.15693661	

	DFFITS	s5	cooks_distance	s6	COVRATIO	s7
1	-0.09760071		1.980265e-03		1.4351361	
2	-0.14932830		4.575734e-03		1.2060991	
3	0.14308207		4.259290e-03		1.9798938	
4	-0.16001829		5.301932e-03		1.3956398	
5	1.00910335	*	1.765512e-01		0.5926489	
6	-0.06627301		9.129493e-04		1.3102871	
7	0.39069524		3.000453e-02		1.0145100	
8	0.28666603		1.708507e-02		3.2299989	*
9	-0.43008292		3.822687e-02		2.3019662	
10	0.06537063		8.894048e-04		1.4425831	
11	-0.68238715		7.818194e-02		0.4505298	
12	0.28205487		1.604155e-02		1.1457525	
13	0.05423639		6.120336e-04		1.3331863	
14	-0.19069555		7.429586e-03		1.1979482	
15	0.11900395		2.919550e-03		1.2286034	
16	0.14825527		4.505833e-03		1.1951550	
17	0.01047482		2.285814e-05		1.4480450	
18	-0.25154755		1.289728e-02		1.2304086	
19	-0.12994709		3.481548e-03		1.2425584	
20	0.01131888		2.669031e-05		1.4590076	
21	-0.08815404		1.616244e-03		1.4481147	
22	-0.42815684		3.733957e-02		1.4625233	
23	-0.48564201		4.379313e-02		0.7453617	
24	-1.81820497	*	5.723811e-01	*	0.8157502	
25	-0.39183932		3.048687e-02		1.0899697	

#从结果来看：
#第 11 号向本 residual 最大，且 standard 和 student 残差绝对值大于 2
#第 5，24 号的 standard, student 和 DFFITS 统计量超过规定指标
#经过分析，第 5，11 和 24 号样本点需要进一步研究
#尝试在回归计算中删除第 5，11 和 24 号样本点

```
Y1 <- Y[-c(5, 11, 24)]
X21 <- X2[-c(5, 11, 24)]
X11 <- X1[-c(5, 11, 24)]

lm.sol1 <- lm(Y1~X11, data=toothpaste)
lm.sol2 <- lm(Y1~X21, data=toothpaste)

plot(Y1~X11)
abline(lm.sol1)
plot(Y1~X21)
abline(lm.sol2)

lm.sol<-lm(Y1 ~ X11 + X21 + I(X21^2) + X11:X21, data=toothpaste)
summary(lm.sol)
```

```
Call:
lm(formula = Y1 ~ X11 + X21 + I(X21^2) + X11:X21, data = toothpaste)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.36558 -0.10188 -0.01011  0.09809  0.27103 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  31.2526     7.5225   4.155 0.000414 ***
X11          13.6861     4.0397   3.388 0.002646 ** 
X21          -8.3609     2.5161  -3.323 0.003089 ** 
I(X21^2)      0.7376     0.2095   3.521 0.001925 ** 
X11:X21      -1.8800     0.6192  -3.036 0.006066 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.164 on 22 degrees of freedom
Multiple R-squared:  0.9518, Adjusted R-squared:  0.9431 
F-statistic: 108.7 on 4 and 22 DF,  p-value: 3.708e-14
```

#可以看出，残差标准差从 0.2063 降低到 0.164，相关系数的平方从 0.9209 提高到 0.9518

detach(toothpaste)

6.5 诊断水泥数据 (数据见例 6.10) 是否存在多重共线性, 分析例 6.10 中 step() 函数去掉的变量是否合理.

```
cement<-data.frame(
  X1=c( 7, 1, 11, 11, 7, 11, 3, 1, 2, 21, 1, 11, 10),
  X2=c(26, 29, 56, 31, 52, 55, 71, 31, 54, 47, 40, 66, 68),
  X3=c( 6, 15, 8, 8, 6, 9, 17, 22, 18, 4, 23, 9, 8),
  X4=c(60, 52, 20, 47, 33, 22, 6, 44, 22, 26, 34, 12, 12),
  Y =c(78.5, 74.3, 104.3, 87.6, 95.9, 109.2, 102.7, 72.5,
       93.1,115.9, 83.8, 113.3, 109.4)
)

lm.sol<-lm(Y ~ X1+X2+X3+X4, data=cement)
summary(lm.sol)

#最优结果
lm.opt<-lm(Y ~ X1+X2, data=cement)
summary(lm.opt)

XX<-cor(cement[1:4])
kappa(XX,exact=TRUE)
#得到 1376.881〉1000， 认为存在严重的多重共线性
eigen(XX)
#取最小的 value, 得到的系数(0.2410522, 0.6417561, 0.2684661, 0.676734)
#（0.2410522, 0.6417561）和（0.2684661, 0.676734）存在线性相关，认为step()中去掉的变量（X3, X4）是合理的。如果
#如果采用 X3, X4 也能取得不错的线性回归公式
lm.opt<-lm(Y ~ X3+X4, data=cement)
summary(lm.opt)
```

```
Call:
lm(formula = Y ~ X3 + X4, data = cement)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.2715 -2.8916 -0.6439  1.5115  8.2566 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 131.28241    3.27477  40.089 2.23e-12 ***
X3           -1.19985    0.18902  -6.348 8.38e-05 ***
X4           -0.72460    0.07233 -10.018 1.56e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.192 on 10 degrees of freedom
Multiple R-squared:  0.9353, Adjusted R-squared:  0.9223 
F-statistic: 72.27 on 2 and 10 DF,  p-value: 1.135e-06
```

6.6

```
X1 = c(1, 1, 1, 1, 0, 0, 0, 0)  #是否用抗生素
X2 = c(1, 1, 0, 0, 1, 1, 0, 0)  #是否有危险因子
X3 = c(1, 0, 1, 0, 1, 0, 1, 0)  #事先有计划
success = c(1, 11, 0, 0, 28, 23, 8, 0)  #有感染
fail = c(17, 87, 2, 0, 30, 3, 32, 9)  #无感染
infection = data.frame(X1, X2, X3, success, fail)
infection$Ymat = cbind(infection$success, infection$fail)

glm.sol = glm(Ymat ~ X1 + X2 + X3, family = binomial, data = infection)
summary(glm.sol)

#感染的回归模型是 P=exp(-0.82-3.2544X1+2.0299X2-1.072X3)/1+exp(-0.82-3.2544X1+2.0299X2-1.072X3)
#根据上述模型，我们认为使用抗生素并有计划，将很大可能无感染。而有危险因子将很大可能有感染
```

6.8

```
X1 = c(70, 60, 70, 40, 40, 70, 70, 80, 60, 30, 80, 40, 60, 40, 20, 50, 50, 40,
       80, 70, 60, 90, 50, 70, 20, 80, 60, 50, 70, 40, 30, 30, 40, 60, 80, 70,
       30, 60, 80, 70)  # 生活行为能力
X2 = c(64, 63, 65, 69, 63, 48, 48, 63, 63, 53, 43, 55, 66, 67, 61, 63, 66, 68,
```

```
    41, 53, 37, 54, 52, 50, 65, 52, 70, 40, 36, 44, 54, 59, 69, 50, 62, 68,
    39, 49, 64, 67)  # 年龄
X3 = c(5, 9, 11, 10, 58, 9, 11, 4, 14, 4, 12, 2, 25, 23, 19, 4, 16, 12, 12,
    8, 13, 12, 8, 7, 21, 28, 13, 13, 22, 36, 9, 87, 5, 22, 4, 15, 4, 11, 10,
    18)  # 诊断到直入研究时间
X4 = c(rep(1, 7), rep(2, 7), rep(3, 2), rep(0, 4), rep(1, 8), rep(2, 4), rep(3,
                                                        3), rep(0, 5))  # 肿瘤类
型
X5 = c(rep(1, 21), rep(0, 19))  # 化疗方法
Y = c(1, rep(0, 11), 1, rep(0, 5), 1, 1, 0, 1, 1, 1, 0, 1, rep(0, 12), 1, 1)
lung.df = data.frame(X1, X2, X3, X4, X5, Y)
lung.df

lung.glm1 = glm(Y ~ X1 + X2 + X3 + X4 + X5, family = binomial, data =
lung.df)
summary(lung.glm1)



#肺癌生存时间的模型是 P=exp(-7.0114+0.0999X1+0.01415X2+0.01749X3-
1.083X4-0.613X5)/1+exp(-7.0114+0.0999X1+0.01415X2+0.01749X3-
1.083X4-0.613X5)
#X1~X5 对 P(Y=1)的综合影响不够显著。X4 肿瘤类型是最主要的影响因素，
但不够显著

#计算病人的生存概率
lung.pre1 = predict(lung.glm1, lung.df[1:5])
p.lung.pre1 = exp(lung.pre1)/(1 + exp(lung.pre1))
p.lung.pre1

#逐步回归选取自变量并计算病人生存概率
lung.glm2 = step(lung.glm1)

summary(lung.glm2)
```

```
lung.pre2 = predict(lung.glm2, lung.df[1:5])
p.lung.pre2 = exp(lung.pre2)/(1 + exp(lung.pre2))
p.lung.pre2
```

#比较两个模型，从估计病人生存时间角度，使用简化模型更方便，且在仅考虑的 X1，X4 两个因素更显著

Plot

第一张为残差散点图，能够看出哪些点拟合的不够好

第二张为标准化残差QQ图，判断标准化残差是否方为正态分布

第三张为标准化残差开方的散点图，能够显示出异常点

第四张图为标准化残差与杠杆值

四张图分别对应plot(lm.sol,1), plot(lm.sol,2), plot(lm.sol,3), plot(lm.sol,5)