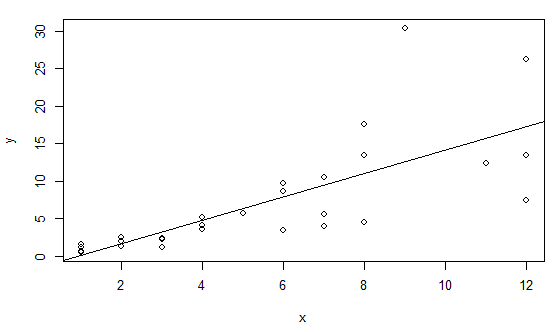
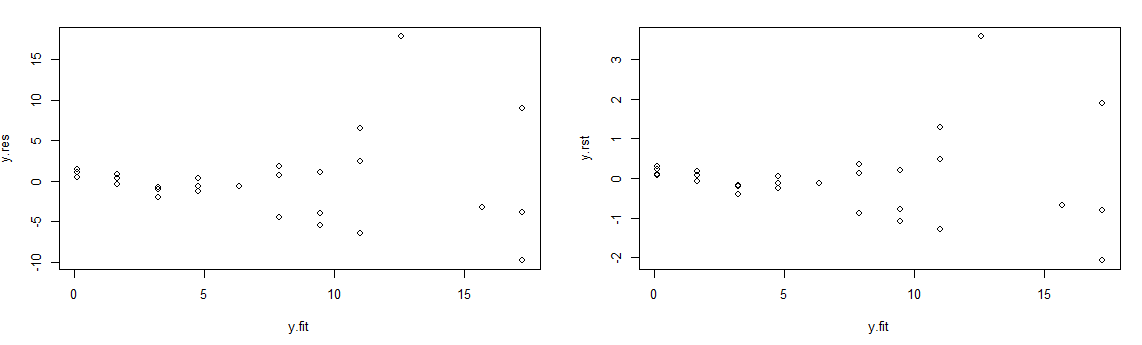
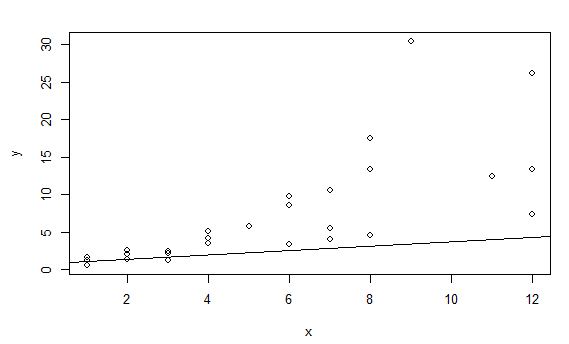
**Week 6**

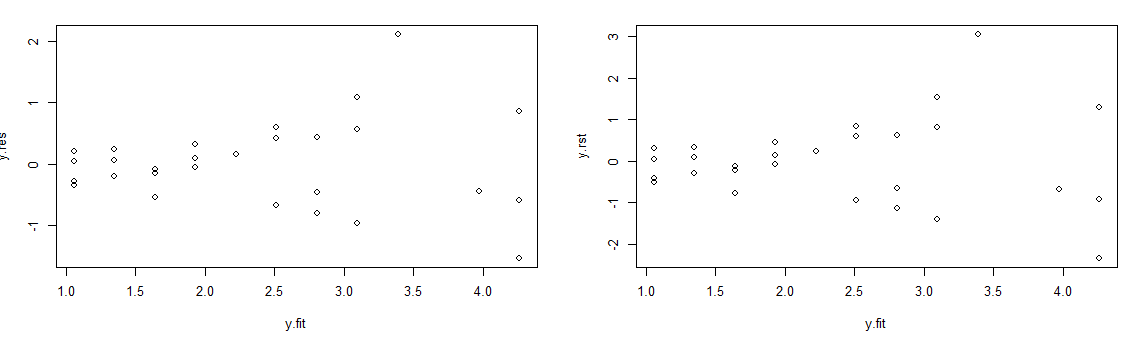


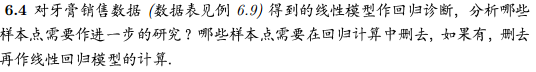
#(1)  
d <- data.frame(  
  x = c(1, 1, 1, 1, 2, 2, 2, 3, 3, 3, 4, 4, 4, 5, 6, 6, 6, 7, 7, 7, 8, 8, 8, 9, 11, 12, 12, 12),  
  y = c(0.6, 1.6, 0.5, 1.2, 2, 1.3, 2.5, 2.2, 2.4, 1.2, 3.5, 4.1, 5.1, 5.7, 3.4, 9.7, 8.6, 4, 5.5, 10.5, 17.5, 13.4, 4.5, 30.4, 12.4, 13.4, 26.2, 7.4)      
)  
  
x = c(1, 1,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



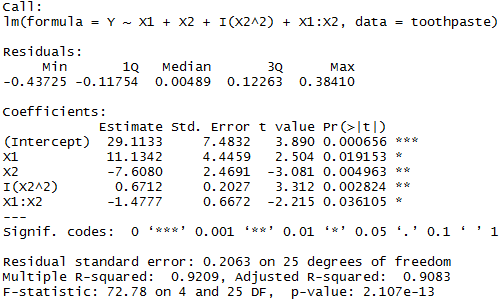
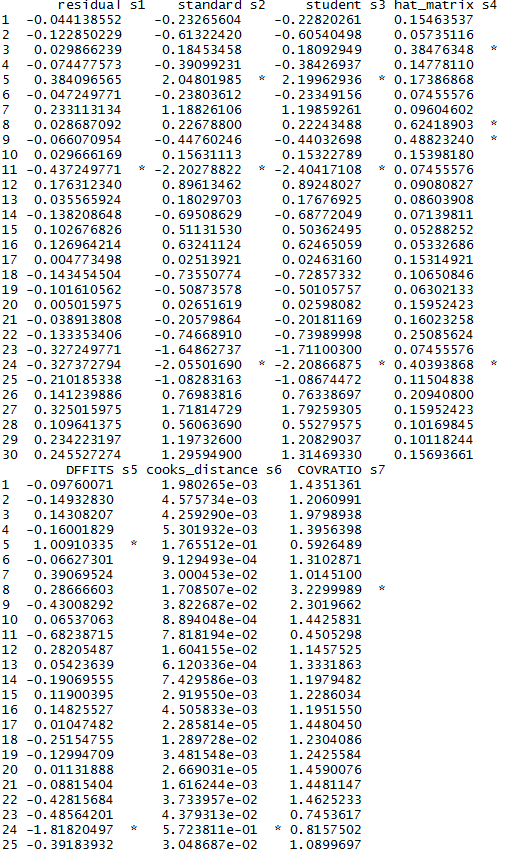
#(2)  
#P-值=0.436>0.05, Intercept没有通过显著性检验(T检验)  
#P-值=7.93e-06<0.05, x通过显著性检验(T检验)  
#P-值=7.931e-06<0.05, 方程整体通过F检验  
  
#(3)  
y.res<-residuals(lm.sol)  
y.fit<-predict(lm.sol)  
y.rst<-rstandard(lm.sol)  
plot(y.res~y.fit)  
plot(y.rst~y.fit)  
  
#从图形上看，误差并不呈现等方差,总体呈现喇叭口的样子  
  
  
#(4)  
lm.new<-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)

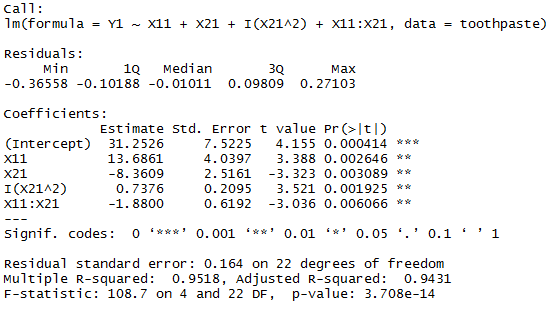




toothpaste<-data.frame(  
  X1=c(-0.05, 0.25,0.60,0, 0.25,0.20, 0.15,0.05,-0.15, 0.15,  
       0.20, 0.10,0.40,0.45,0.35,0.30, 0.50,0.50, 0.40,-0.05,  
       -0.05,-0.10,0.20,0.10,0.50,0.60,-0.05,0, 0.05, 0.55),  
  X2=c( 5.50,6.75,7.25,5.50,7.00,6.50,6.75,5.25,5.25,6.00,  
        6.50,6.25,7.00,6.90,6.80,6.80,7.10,7.00,6.80,6.50,  
        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)

  
  
lm.sol1 <- lm(Y~X1, data=toothpaste)  
lm.sol2 <- lm(Y~X2, data=toothpaste)  
  
source("Reg\_Diag.R")  
Reg\_Diag(lm.sol)  
  
  
#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

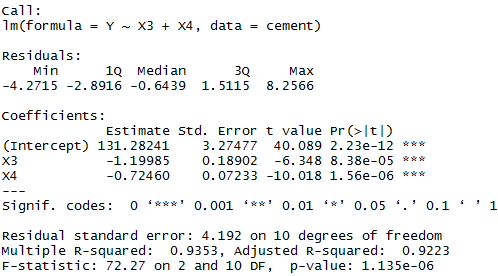
#尝试在回归计算中删除第5，11和24号样本点  
  
Y1 <- Y[-c(5, 11, 24)]  
X21 <- X2[-c(5, 11, 24)]  
X11 <- X1[-c(5, 11, 24)]  
  
lm.sol1 <- lm(Y1~X11, data=toothpaste)  
lm.sol2 <- lm(Y1~X21, data=toothpaste)  
  
plot(Y1~X11)  
abline(lm.sol1)  
plot(Y1~X21)  
abline(lm.sol2)  
  
lm.sol<-lm(Y1 ~ X11 + X21 + I(X21^2) + X11:X21, data=toothpaste)  
summary(lm.sol)



#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)

C:\Users\XU6\AppData\Local\Temp\enhtmlclip\Image(9).png

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)



**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)

