ST503 HW1 SHAN YANG

## 2.4

library(faraway)  
data<-prostate  
head(prostate)

## lcavol lweight age lbph svi lcp gleason pgg45 lpsa  
## 1 -0.5798185 2.7695 50 -1.386294 0 -1.38629 6 0 -0.43078  
## 2 -0.9942523 3.3196 58 -1.386294 0 -1.38629 6 0 -0.16252  
## 3 -0.5108256 2.6912 74 -1.386294 0 -1.38629 7 20 -0.16252  
## 4 -1.2039728 3.2828 58 -1.386294 0 -1.38629 6 0 -0.16252  
## 5 0.7514161 3.4324 62 -1.386294 0 -1.38629 6 0 0.37156  
## 6 -1.0498221 3.2288 50 -1.386294 0 -1.38629 6 0 0.76547

get the regression of each variables added in

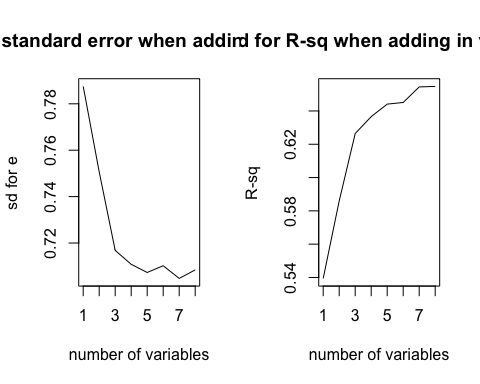
olca<-lm(lpsa~lcavol,data=prostate)  
olwei<-lm(lpsa~lcavol+lweight,data=prostate)  
osvi<-lm(lpsa~lcavol+lweight+svi,data=prostate)  
olbph<-lm(lpsa~lcavol+lweight+svi+lbph,data=prostate)  
oage<-lm(lpsa~lcavol+lweight+svi+lbph+age,data=prostate)  
olcp<-lm(lpsa~lcavol+lweight+svi+lbph+age+lcp,data=prostate)  
opgg<-lm(lpsa~lcavol+lweight+svi+lbph+age+lcp+pgg45,data=prostate)  
oglea<-lm(lpsa~lcavol+lweight+svi+lbph+age+lcp+pgg45+gleason,data=prostate)

get their standard errors for residuals

sd<-c(summary(olca)$sigma,summary(olwei)$sigma,summary(osvi)$sigma,summary(olbph)$sigma,summary(oage)$sigma,summary(olcp)$sigma,summary(opgg)$sigma,summary(oglea)$sigma)  
##Can not use sd(lm$residuals), the df is wrong  
##Same as sqrt(deviance(oglea)/df.residual(oglea))  
  
#get the r-squares for regression  
rsquared<-c(summary(olca)$r.squared,summary(olwei)$r.squared,summary(osvi)$r.squared,summary(olbph)$r.squared,summary(oage)$r.squared,summary(olcp)$r.squared,summary(opgg)$r.squared,summary(oglea)$r.squared)

plot the trend

par(mfrow=c(1, 2))  
plot(sd,main="Trend for standard error when adding in variables",xlab="number of variables",ylab="sd for e",type="l")  
plot(rsquared,main="Trend for R-sq when adding in variables",xlab="number of variables",ylab="R-sq",type="l")

 ##2.6

cheddar<-cheddar  
head(cheddar)

## taste Acetic H2S Lactic  
## 1 12.3 4.543 3.135 0.86  
## 2 20.9 5.159 5.043 1.53  
## 3 39.0 5.366 5.438 1.57  
## 4 47.9 5.759 7.496 1.81  
## 5 5.6 4.663 3.807 0.99  
## 6 25.9 5.697 7.601 1.09

(a)Regression and get the coefficients

ocheddar<-lm(taste~Acetic+H2S+Lactic,data=cheddar)  
ocheddar

##   
## Call:  
## lm(formula = taste ~ Acetic + H2S + Lactic, data = cheddar)  
##   
## Coefficients:  
## (Intercept) Acetic H2S Lactic   
## -28.8768 0.3277 3.9118 19.6705

(b)get the corr and square it

fittedy<-ocheddar$fitted.values  
corsq<-cor(x=cheddar$taste,y=fittedy)^2  
corsq

## [1] 0.6517747

summary(ocheddar)

##   
## Call:  
## lm(formula = taste ~ Acetic + H2S + Lactic, data = cheddar)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -17.390 -6.612 -1.009 4.908 25.449   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -28.8768 19.7354 -1.463 0.15540   
## Acetic 0.3277 4.4598 0.073 0.94198   
## H2S 3.9118 1.2484 3.133 0.00425 \*\*  
## Lactic 19.6705 8.6291 2.280 0.03108 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 10.13 on 26 degrees of freedom  
## Multiple R-squared: 0.6518, Adjusted R-squared: 0.6116   
## F-statistic: 16.22 on 3 and 26 DF, p-value: 3.81e-06

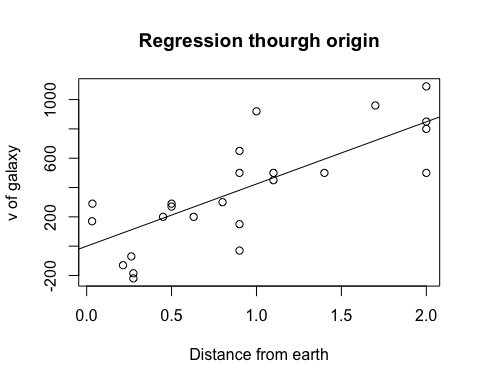
So the corr-sq is the Multiple R-squared in the regression.

Problem 4(b). Regression through the origin, Hubble data

x is the distance from earth, y is the v of a galaxy

x <- c(0.032, 0.034, 0.214, 0.263, 0.275, 0.275, 0.450, 0.500, 0.500, 0.630, 0.800, 0.900,   
 0.900, 0.900, 0.900, 1.000, 1.100, 1.100, 1.400, 1.700, 2.000, 2.000, 2.000, 2.000)  
y <- c(170, 290, -130,-70, -185, -220, 200, 290, 270, 200, 300, -30,   
 650, 150, 500, 920, 450, 500, 500, 960, 500, 850, 800, 1090)

ohubble<-lm(y~0+x) #do the regression and force it go through the origin  
plot(x,y,xlab = "Distance from earth",ylab = "v of galaxy",main = "Regression thourgh origin") #get the scatter plot  
abline(ohubble) #add regression line to it

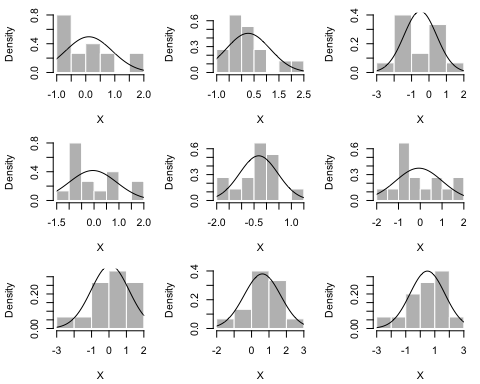


fittedv<-ohubble$fitted.values  
cor(y,fittedv)^2

## [1] 0.6235305

We can read the correlation is .6235, which means a relatively strong positive relationship between distance and velocity

n <- 15  
op <- par(pty="m", mfrow=c(3, 3), mar=c(4.2, 4.2, 1, 1))  
for(i in 1:9) {  
   
 X <- rnorm(n)  
 hist(X, freq=FALSE, col="gray", border="white", main="")  
 curve(dnorm(x, mean(X), sd(X)), add=TRUE)  
   
}



par(op)

4 of them are kind or “normal”. The first one is right-skewed, the second is approximately normal but with really heavy tail, third has two “tops”, forth is almost normal, fifth also has two tops, sixth right-skewed, seventh seems normal, eighth almost normal, ninth almost normal.