ST503 HW2 SHAN YANG

## 3.1

library(faraway)  
prostate<-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

fit the model

oprostate<-lm(lpsa~lcavol+lweight+age+lbph+svi+lcp+gleason+pgg45,data = prostate)  
sumary(oprostate)

## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 0.6693367 1.2963875 0.5163 0.606934  
## lcavol 0.5870218 0.0879203 6.6767 2.111e-09  
## lweight 0.4544674 0.1700124 2.6731 0.008955  
## age -0.0196372 0.0111727 -1.7576 0.082293  
## lbph 0.1070540 0.0584492 1.8316 0.070398  
## svi 0.7661573 0.2443091 3.1360 0.002329  
## lcp -0.1054743 0.0910135 -1.1589 0.249638  
## gleason 0.0451416 0.1574645 0.2867 0.775033  
## pgg45 0.0045252 0.0044212 1.0235 0.308860  
##   
## n = 97, p = 9, Residual SE = 0.70842, R-Squared = 0.65

(a)CI for age

cat("the 90% CI for age is:",(-0.0196372+c(-1,1)\*qt(.95,97-9)\*0.0111727),"and 95% CI for age is:",(-0.0196372+c(-1,1)\*qt(.975,97-9)\*0.0111727))

## the 90% CI for age is: -0.03821018 -0.001064217 and 95% CI for age is: -0.04184059 0.002566193

We can reject at (90% confidence level) since the interval does not includes 0, however, we cannot do it when (95% level).And we can see, the true p-value is 0.082293, which confirms our guess.

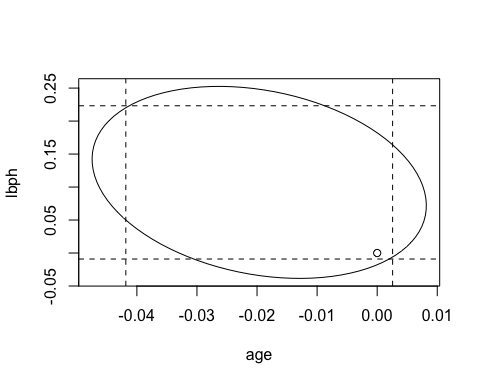
#draw the confidence region  
library(ellipse)

## Warning: package 'ellipse' was built under R version 3.4.3

##   
## Attaching package: 'ellipse'

## The following object is masked from 'package:graphics':  
##   
## pairs

plot(ellipse(oprostate,c(4,5)),type = "l")  
points(0,0)  
abline(v=confint(oprostate)[4,],lty=2)  
abline(h=confint(oprostate)[5,],lty=2)



The origin of the plot tests and . And the test result is we cannot reject the null because the point in in the region.

tstats<-4000  
set.seed(100)  
for (i in 1:tstats){  
 o<-lm(lpsa~lcavol+lweight+sample(age)+lbph+svi+lcp+gleason+pgg45,data = prostate) #generate permutation for age column  
 tstats[i]<-summary(o)$coef[4,3] #get the t value for each permutation  
}  
mean(abs(tstats)>abs(summary(oprostate)$coef[4,3])) #check the percentage of two tails t-stats larger than not permutated

## [1] 0.09225

Comparing with the p-value 0.082293 we got before, it is not so different.

## 7

punting<-punting  
head(punting)

## Distance Hang RStr LStr RFlex LFlex OStr  
## 1 162.50 4.75 170 170 106 106 240.57  
## 2 144.00 4.07 140 130 92 93 195.49  
## 3 147.50 4.04 180 170 93 78 152.99  
## 4 163.50 4.18 160 160 103 93 197.09  
## 5 192.00 4.35 170 150 104 93 266.56  
## 6 171.75 4.16 150 150 101 87 260.56

opunting<-lm(Distance~RStr+LStr+RFlex+LFlex,data = punting)  
summary(opunting)

##   
## Call:  
## lm(formula = Distance ~ RStr + LStr + RFlex + LFlex, data = punting)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -23.941 -8.958 -4.441 13.523 17.016   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) -79.6236 65.5935 -1.214 0.259  
## RStr 0.5116 0.4856 1.054 0.323  
## LStr -0.1862 0.5130 -0.363 0.726  
## RFlex 2.3745 1.4374 1.652 0.137  
## LFlex -0.5277 0.8255 -0.639 0.541  
##   
## Residual standard error: 16.33 on 8 degrees of freedom  
## Multiple R-squared: 0.7365, Adjusted R-squared: 0.6047   
## F-statistic: 5.59 on 4 and 8 DF, p-value: 0.01902

we can read from the Pr column, at .05 level, none of the predictors is significant.

1. the p-value for F-stat is .019, means we can reject the . Which means they are collectively has influence on the response

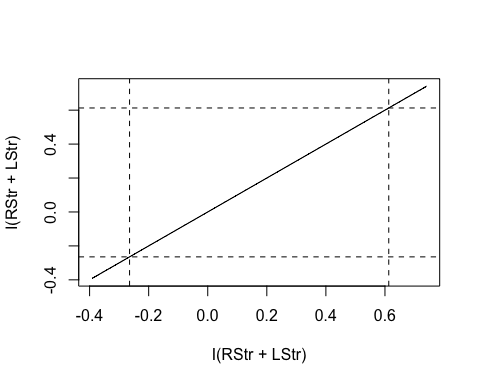
opuntingsub<-lm(Distance~I(RStr+LStr)+RFlex+LFlex,data = punting)  
anova(opuntingsub,opunting)

## Analysis of Variance Table  
##   
## Model 1: Distance ~ I(RStr + LStr) + RFlex + LFlex  
## Model 2: Distance ~ RStr + LStr + RFlex + LFlex  
## Res.Df RSS Df Sum of Sq F Pr(>F)  
## 1 9 2287.4   
## 2 8 2132.6 1 154.72 0.5804 0.468

the p-value for F-stat is 0.468, means we cannot reject the

We can see the region is kind of downward, when right leg strength impact is high, the left leg strength impact will tend to be low. Which means this two variables are correlated. So actually, the sum of them influences the response.

#draw the confidence region  
library(ellipse)  
plot(ellipse(opuntingsub,c(2,2)),type = "l")  
abline(v=confint(opuntingsub)[2,],lty=2)  
abline(h=confint(opuntingsub)[2,],lty=2)

 We can see from the plot that the ellipse is collapsed into a line，and their confidence region is a line now, which satisfies the (C), we cannot reject the H0:two betas are equal—-corr=1. ##4

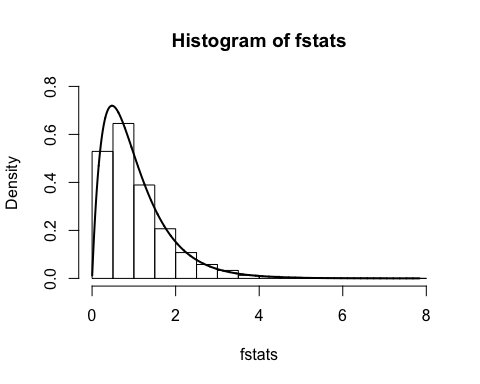
teengamb<-teengamb  
head(teengamb)

## sex status income verbal gamble  
## 1 1 51 2.00 8 0.0  
## 2 1 28 2.50 8 0.0  
## 3 1 37 2.00 6 0.0  
## 4 1 28 7.00 4 7.3  
## 5 1 65 2.00 8 19.6  
## 6 1 61 3.47 6 0.1

1. Because the model will become under null. With e~N(0,) and as fixed value. y will be N(,).

fstats<-5000  
set.seed(100)  
for (i in 1:fstats){  
 y<-rnorm(47)  
 o<-lm(y~sex+status+income+verbal,data = teengamb) #generate permutation for age column  
 fstats[i]<-summary(o)$fstat[1] #get the t value for each permutation  
}

hist(fstats,ylim=c(0,.8),freq = FALSE)  
xfit<-seq(min(fstats), max(fstats), length = 5000)  
yfit<-df(xfit, df1=4,df2=42)   
lines(xfit, yfit,col = "black", lwd = 2)



Yes, it follows