ST 503 Hw 2

Robin Baldeo

### Question 1(Exercise 3.1 in LMR)

### (A)

#full model  
lmod<-lm(lpsa~lcavol+lweight+age+lbph+ svi+lcp+ gleason+ pgg45, data=prostate )  
  
#summary  
summary(lmod)

##   
## Call:  
## lm(formula = lpsa ~ lcavol + lweight + age + lbph + svi + lcp +   
## gleason + pgg45, data = prostate)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.7331 -0.3713 -0.0170 0.4141 1.6381   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.669337 1.296387 0.516 0.60693   
## lcavol 0.587022 0.087920 6.677 2.11e-09 \*\*\*  
## lweight 0.454467 0.170012 2.673 0.00896 \*\*   
## age -0.019637 0.011173 -1.758 0.08229 .   
## lbph 0.107054 0.058449 1.832 0.07040 .   
## svi 0.766157 0.244309 3.136 0.00233 \*\*   
## lcp -0.105474 0.091013 -1.159 0.24964   
## gleason 0.045142 0.157465 0.287 0.77503   
## pgg45 0.004525 0.004421 1.024 0.30886   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.7084 on 88 degrees of freedom  
## Multiple R-squared: 0.6548, Adjusted R-squared: 0.6234   
## F-statistic: 20.86 on 8 and 88 DF, p-value: < 2.2e-16

#a  
#at the 90% Ci reject Ho because 0 is not within the interval  
confint(lmod, "age", level=.90)

## 5 % 95 %  
## age -0.0382102 -0.001064151

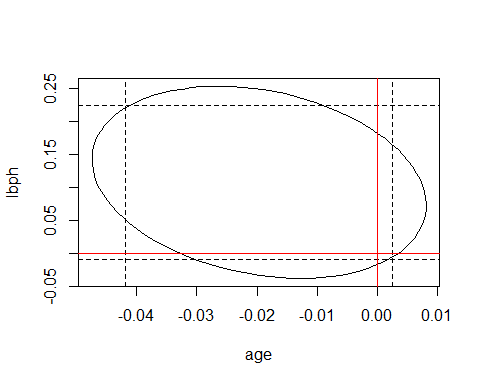
#at the 95% Ci fail to reject H0 becuause 0 is within interval  
confint(lmod, "age")

## 2.5 % 97.5 %  
## age -0.04184062 0.002566267

At the 90% level we reject the null hypothesis of age being insignificant because 0 is not within the CI, However, at 95% level we fail to reject the null because 0 is within the interval. This also corresponds to the p-value for age in the summary output. Age p-value is greater than .05 because of this we fail to reject the null.So Both the summary ouput p-value and Ci at the 95% level gives the same conclusion.

### (B)

ci<-confint(lmod)  
#ellipse  
plot(ellipse(lmod, c(4,5)), type = "l")  
#CI  
abline(v=ci[4,], h=ci[5,], lty=2)  
#marking orgin with red lines  
abline(h= 0, v = 0, col = "red")



#partial f test to confirm  
lmodR<-lm(lpsa~lcavol+lweight+ svi+lcp+ gleason+ pgg45, data = prostate)  
anova(lmodR, lmod)

## Analysis of Variance Table  
##   
## Model 1: lpsa ~ lcavol + lweight + svi + lcp + gleason + pgg45  
## Model 2: lpsa ~ lcavol + lweight + age + lbph + svi + lcp + gleason +   
## pgg45  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 90 46.768   
## 2 88 44.163 2 2.6048 2.5951 0.08034 .  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

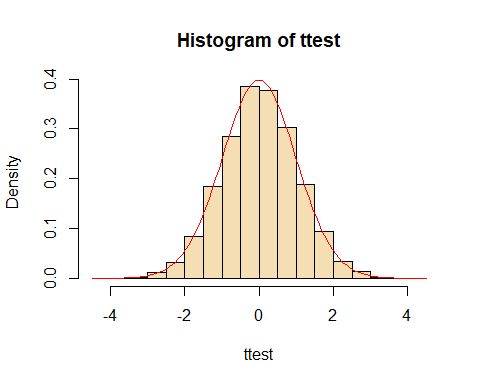
Here we fail to reject the null because the origin(red lines) falls within the ellipse. Therefore, these variables age and ibph can be removed from the model because of no significant contribution to the quality of the model.This is also confirmed with a partial f-test. Here p-value> .05 confirming that the values age and lbph don’t contribute to the model in a significant way.

### (C)

set.seed(123)  
n<- 10000  
  
ttest<-numeric(n)  
  
#looping n times   
for(i in 1:n){  
 #storing all t values into the ttest vector  
 ttest[i]= (summary(lm(lpsa~lcavol+lweight+sample(age)+lbph+ svi+lcp+ gleason+ pgg45, data=prostate )))$coefficients[4,3]  
}  
  
#simulated p-valu for t-test  
2 \* mean(ttest<summary(lmod)$coefficients[4,3])

## [1] 0.0778

hist(ttest, freq = F, col = "wheat")  
x<-ttest  
curve(dt(x, length(x)), add=TRUE, col = "red")



The permutation method corresponds to the p-value from the summary output. The simulated p-value for age is 0.0778 which closely matches the summary value of 0.0822932.

### Question 2(Exercise 3.5 on page 50 of LMR.)

From class lecture notes:

Where and . As a result:

Now the F test from page 35 of text book is given as:

Using the expression which is plugged into the formula we get formula relating to the :

### Question 3(Exercise 3.7 on page 50 of LMR)

### (A)

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

## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) -79.6236456 65.5935321 -1.2138948 0.2594050  
## RStr 0.5116373 0.4855691 1.0536859 0.3228080  
## LStr -0.1861996 0.5129643 -0.3629875 0.7260179  
## RFlex 2.3745010 1.4373549 1.6519935 0.1371390  
## LFlex -0.5277338 0.8255156 -0.6392778 0.5405109

#none

none of the variables in the model are significant at 5% level.

### (B)

f<- summary(lmodD)$fstat  
1- pf(f[1], f[2], f[3])

## value   
## 0.01902482

we reject the null in favour of the alternative where at least one coefficient is non zero because p-value = 0.0190248) <.05.

### (C)

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

##   
## Call:  
## lm(formula = Distance ~ I(RStr + LStr) + RFlex + LFlex, data = punting)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -21.698 -9.494 -5.155 9.081 20.611   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) -71.2694 63.1447 -1.129 0.288  
## I(RStr + LStr) 0.1741 0.1940 0.898 0.393  
## RFlex 2.3137 1.4013 1.651 0.133  
## LFlex -0.5772 0.8035 -0.718 0.491  
##   
## Residual standard error: 15.94 on 9 degrees of freedom  
## Multiple R-squared: 0.7174, Adjusted R-squared: 0.6232   
## F-statistic: 7.615 on 3 and 9 DF, p-value: 0.00769

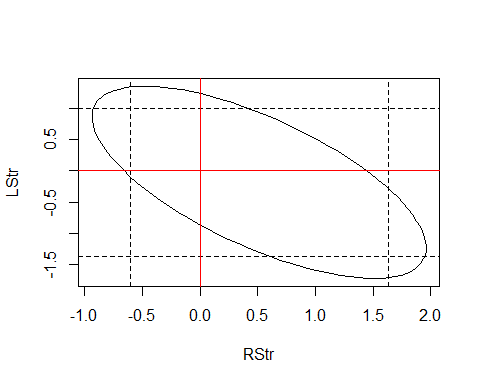
anova(lmodDR, lmodD)

## 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

Based on the p-values >.05, we fail to reject h0=B\_rstr=B\_lstr = 0, where the proposed simplification to the model may be justified.

### (D)

ci <- confint(lmodD)  
plot(ellipse(lmodD, c(2,3)), type ="l")  
abline(h =ci[3,], v = ci[2,], lty = 2 )  
abline(h= 0, v = 0, col = "red")



lmodDR1<-lm(Distance~ RFlex + LFlex, data = punting)  
anova(lmodDR1, lmodD)

## Analysis of Variance Table  
##   
## Model 1: Distance ~ RFlex + LFlex  
## Model 2: Distance ~ RStr + LStr + RFlex + LFlex  
## Res.Df RSS Df Sum of Sq F Pr(>F)  
## 1 10 2492.1   
## 2 8 2132.6 2 359.46 0.6742 0.5363

We fail to reject the null of B1\_rstr=B2\_lstr= 0. Because the origin falls within the ellipse. Since our test in c shows that RStr + LStr has the same effect as using the variables separately. We see from the test above that the two variables separately or combined might not have any significant contribution to the reponse. This is also confirmed with a partial f-test where p-values > .05.

### Question 4

### (A)

set.seed(123)  
  
nr <- nrow(teengamb)  
  
teen <-teengamb  
  
teen$y <- rnorm(nr)  
  
  
model<-lm(y~sex + status + income+ verbal, data = teen)  
summary(model)

##   
## Call:  
## lm(formula = y ~ sex + status + income + verbal, data = teen)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.06932 -0.65108 0.00983 0.65841 2.17125   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) -0.50355 0.73584 -0.684 0.498  
## sex 0.43498 0.35135 1.238 0.223  
## status 0.01191 0.01203 0.990 0.328  
## income 0.01404 0.04388 0.320 0.751  
## verbal -0.03669 0.09295 -0.395 0.695  
##   
## Residual standard error: 0.9709 on 42 degrees of freedom  
## Multiple R-squared: 0.03865, Adjusted R-squared: -0.05291   
## F-statistic: 0.4221 on 4 and 42 DF, p-value: 0.7918

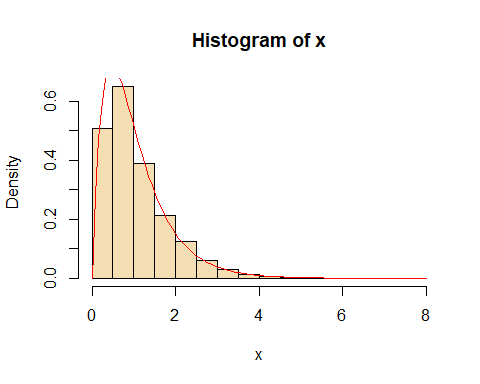
We would fail to reject the null because the p-value > .05.

### (B)

set.seed(123)  
  
B <- 5000  
x <- numeric(B)  
  
for(b in 1:B) {  
 teen <-teengamb  
 teen$y <- rnorm(nr)  
 o.full<-lm(y~sex + status + income+ verbal, data = teen)  
 x[b]<- summary(o.full)$fstat[1]  
   
}

### (C)

hist(x, freq=F, col= "wheat" )  
curve(df(x, 4, 42), add = T, col = "red")



Yes, the claim is true, the plot proves that the simulation of the 5000 f plotted gives the f-distribution.