STT465Hw4

Sam Isken

October 27, 2019

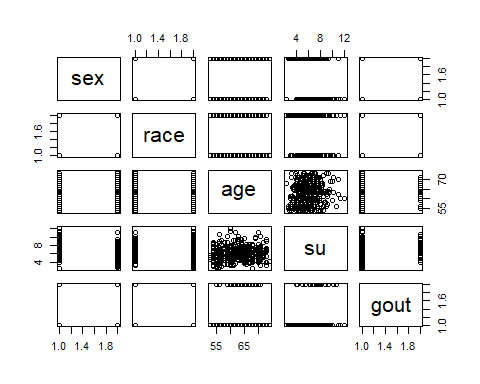
STT 465 Fall 2019 Homework 4 Due 10/30/2019 (In Class)

Instruction: When using R in any problem, copy the code and results onto your word document under that question number and add any required comments. You will lose points if I do not see your codes.

You should present a stapled document when multiple pages are used. The grader will not be held responsible for any loss of pages.

The gout data set (in D2L) contains information on the gout status of individual with respect to their sex, race, age and serum urate (su). After downloading onto your computer, import the data into R using:

GOUT <- read.table("gout.txt",header=TRUE)  
## Transform qualitative variables into factors in R, for example:  
GOUT$sex <- factor(GOUT$sex,levels=c('M','F'))  
pairs(GOUT)



summary(GOUT)

## sex race age su gout   
## M:175 B: 92 Min. :53.0 Min. : 2.400 N:370   
## F:225 W:308 1st Qu.:58.0 1st Qu.: 4.900 Y: 30   
## Median :63.0 Median : 6.100   
## Mean :62.7 Mean : 6.063   
## 3rd Qu.:67.0 3rd Qu.: 7.000   
## Max. :74.0 Max. :12.100

head(GOUT)

## sex race age su gout  
## 1 M W 67 8.3 N  
## 2 F W 72 8.6 N  
## 3 F W 70 7.3 N  
## 4 F W 63 6.2 N  
## 5 F W 55 4.3 N  
## 6 M W 63 7.0 N

1. Using the gout data set estimate the effects of sex, race and age on serum urate using the lm() function.

Model1\_Gout <- lm(su~sex+race+age, data=GOUT)  
Model1\_Gout

##   
## Call:  
## lm(formula = su ~ sex + race + age, data = GOUT)  
##   
## Coefficients:  
## (Intercept) sexF raceW age   
## 5.84828 -1.52853 -0.78212 0.02674

-Given a 1 unit (year) increase in age there is predicted to be a .02674 increase in su. -Given the sex of a patient is female there is predicted to be a 1.52853 DECREASE (-1.52853) in su. -Given the race of a patient is white there is predicted to be a .78212 DECREASE (-.78212) in su.

1. Report your results.

Model1\_Gout

##   
## Call:  
## lm(formula = su ~ sex + race + age, data = GOUT)  
##   
## Coefficients:  
## (Intercept) sexF raceW age   
## 5.84828 -1.52853 -0.78212 0.02674

summary(Model1\_Gout)

##   
## Call:  
## lm(formula = su ~ sex + race + age, data = GOUT)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -4.4843 -0.9717 -0.1829 0.8276 5.4296   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 5.84828 0.82314 7.105 5.64e-12 \*\*\*  
## sexF -1.52853 0.14306 -10.684 < 2e-16 \*\*\*  
## raceW -0.78212 0.16932 -4.619 5.22e-06 \*\*\*  
## age 0.02674 0.01299 2.058 0.0402 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.413 on 396 degrees of freedom  
## Multiple R-squared: 0.2504, Adjusted R-squared: 0.2447   
## F-statistic: 44.09 on 3 and 396 DF, p-value: < 2.2e-16

coefficients(Model1\_Gout)

## (Intercept) sexF raceW age   
## 5.84828010 -1.52852797 -0.78211876 0.02673734

The initial model is:

After calculating predicted values we can show our model to be:

1. Summarize your findings.

-Given a 1 unit (year) increase in age there is predicted to be a .02674 increase in su. -Given the sex of a patient is female there is predicted to be a 1.52853 DECREASE (-1.52853) in su. -Given the race of a patient is white there is predicted to be a .78212 DECREASE (-.78212) in su.

The initial model is:

After calculating predicted values we can show our model to be:

All of the predictors are significant with age being the least significant.

1. Report the code you used to obtain the results in (a) and (b)

Model1\_Gout <- lm(su~sex+race+age, data=GOUT)  
Model1\_Gout

##   
## Call:  
## lm(formula = su ~ sex + race + age, data = GOUT)  
##   
## Coefficients:  
## (Intercept) sexF raceW age   
## 5.84828 -1.52853 -0.78212 0.02674

summary(Model1\_Gout)

##   
## Call:  
## lm(formula = su ~ sex + race + age, data = GOUT)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -4.4843 -0.9717 -0.1829 0.8276 5.4296   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 5.84828 0.82314 7.105 5.64e-12 \*\*\*  
## sexF -1.52853 0.14306 -10.684 < 2e-16 \*\*\*  
## raceW -0.78212 0.16932 -4.619 5.22e-06 \*\*\*  
## age 0.02674 0.01299 2.058 0.0402 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.413 on 396 degrees of freedom  
## Multiple R-squared: 0.2504, Adjusted R-squared: 0.2447   
## F-statistic: 44.09 on 3 and 396 DF, p-value: < 2.2e-16

coefficients(Model1\_Gout)

## (Intercept) sexF raceW age   
## 5.84828010 -1.52852797 -0.78211876 0.02673734

LaTeX Code:

# Initial Model   
\text{su}=\beta\_0 + \text{sexF}\*x\_1 + \text{raceW}\*x\_2 + \text{age} \* x\_3  
# Filled in Model  
Y = -1.52853 x\_1 + -0.78212 x\_2 + 0.02674 x\_3 +5.84828

1. Produce a table equal to summary( ) using matrix operations. Your table must include parameter estimate, SE, t-statistic and p-value.

# Incidence matrix for intercept and effects of Sex, Race, Age,  
X=cbind(1,GOUT$sex,GOUT$race, GOUT$age)   
X1 <- model.matrix(~sex+race+age,data=GOUT)  
head(X)

## [,1] [,2] [,3] [,4]  
## [1,] 1 1 2 67  
## [2,] 1 2 2 72  
## [3,] 1 2 2 70  
## [4,] 1 2 2 63  
## [5,] 1 2 2 55  
## [6,] 1 1 2 63

y=GOUT$su  
head(y)

## [1] 8.3 8.6 7.3 6.2 4.3 7.0

dim (X)

## [1] 400 4

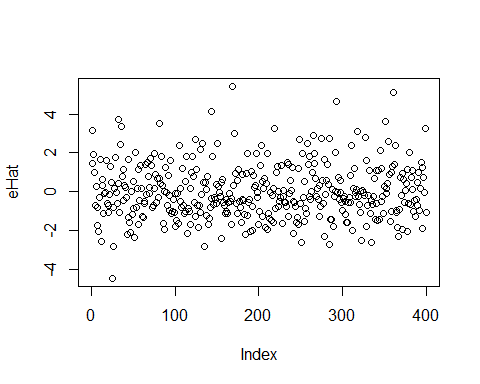
y = as.matrix(y)  
dim(y)

## [1] 400 1

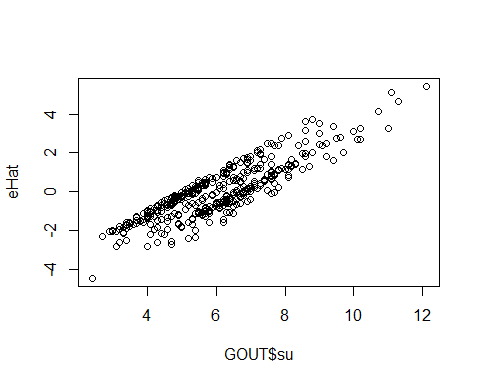
# OLS equations  
Xy=t(X)%\*%y  
XtX=t(X)%\*%X  
# Estimates, compare with fm$coef  
bHat2=solve(XtX,Xy)  
bHat2

## [,1]  
## [1,] 8.15892684  
## [2,] -1.52852797  
## [3,] -0.78211876  
## [4,] 0.02673734

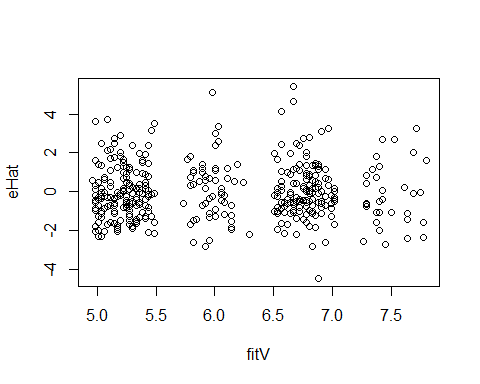
# To get SEs we need an estimate of the error variance  
fitV = X%\*%bHat2 ## Fitted values from ourmodel  
eHat=y-fitV ### Residual  
plot(eHat) ## Sequence plot of residual



plot(eHat ~ GOUT$su) ### ehat vs x



plot(eHat ~ fitV) ### Residual vs fitted value plot



vEHat=sum(eHat^2)/(nrow(GOUT)-ncol(X)) # Sum of squares of errors divide by n-rank(X)  
SE=sqrt(diag(solve(XtX)\*vEHat)) ## SE of bHat  
t\_stat=bHat2/SE  
## P-values (under normality first, and under t then...)  
pvaluesN=pnorm(abs(t\_stat),lower.tail=F)\*2  
pvaluesT=pt(abs(t\_stat),df=length(y)-ncol(X),lower.tail=F)\*2  
  
summary(Model1\_Gout)

##   
## Call:  
## lm(formula = su ~ sex + race + age, data = GOUT)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -4.4843 -0.9717 -0.1829 0.8276 5.4296   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 5.84828 0.82314 7.105 5.64e-12 \*\*\*  
## sexF -1.52853 0.14306 -10.684 < 2e-16 \*\*\*  
## raceW -0.78212 0.16932 -4.619 5.22e-06 \*\*\*  
## age 0.02674 0.01299 2.058 0.0402 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.413 on 396 degrees of freedom  
## Multiple R-squared: 0.2504, Adjusted R-squared: 0.2447   
## F-statistic: 44.09 on 3 and 396 DF, p-value: < 2.2e-16

cbind(bHat2,SE,t\_stat,pvaluesT)

## SE   
## [1,] 8.15892684 0.88565712 9.212286 1.893885e-18  
## [2,] -1.52852797 0.14306195 -10.684378 1.388923e-23  
## [3,] -0.78211876 0.16931945 -4.619190 5.217251e-06  
## [4,] 0.02673734 0.01298969 2.058352 4.021019e-02

1. Report your results and comment on any difference between this and results from 1(a)

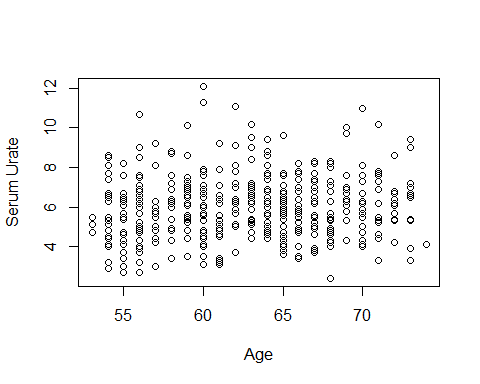
The table is shown above in the matrix with the only difference being the intercept values.

1. Report the code you used to produce the results reported in 2(a)

# Code is shown above.

1. Using the estimate effects reported in 1(a) above; produce a plot with age in the x-axis, expected serum urate in the y-axis and lines for male/white, female/white, male/black, female/black. Use a range from 30 to 70 for the x-axis values. Clearly distinguish the different links (Use different colors on different line types).

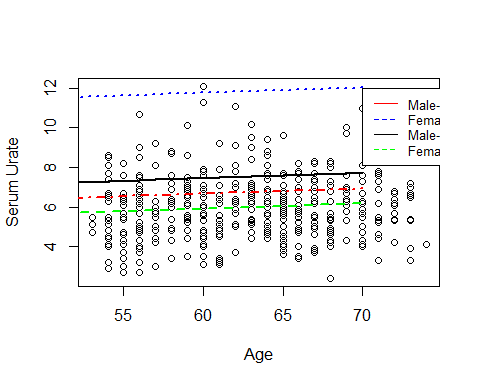
#Base Plot  
plot(x=GOUT$age ,y=GOUT$su,   
 xlab = "Age",  
 ylab = "Serum Urate")



## Prediction equations  
bHat1=coef(Model1\_Gout)  
bHat1

## (Intercept) sexF raceW age   
## 5.84828010 -1.52852797 -0.78211876 0.02673734

plot(su~age,data =GOUT,  
 xlab = "Age",  
 ylab = "Serum Urate")  
  
# Prediction Equation for Age: Base, no variables added  
#Int= bHat1[1]  
#Slope=bHat1[4]  
#Age=seq(from=30,to=70,by=1)  
#lines(x=Age,y=Int+Age\*Slope,col=4)  
  
# Prediction Equation for Age: Male, White  
Int= bHat1[1]+bHat1[3]  
Slope=bHat1[4]  
Age=seq(from=30,to=70,by=1)  
lines(x=Age,y=Int+Age\*Slope,col="red",lwd=2,lty="dotdash")  
  
# Prediction Equation for Age: Female, White  
Int= bHat1[1]+bHat1[1]+bHat1[2]  
Slope=bHat1[4]  
Age=seq(from=30,to=70,by=1)  
lines(x=Age,y=Int+Age\*Slope,col="blue",lwd=2,lty="dotted")  
  
# Prediction Equation for Age: Male, Black  
Int= bHat1[1]  
Slope=bHat1[4]  
Age=seq(from=30,to=70,by=1)  
lines(x=Age,y=Int+Age\*Slope,col="black",lwd=2,lty="solid")  
  
# Prediction Equation for Age: Female, Black   
Int= bHat1[1]+bHat1[2]  
Slope=bHat1[4]  
Age=seq(from=30,to=70,by=1)  
lines(x=Age,y=Int+Age\*Slope,col="green",lwd=2,lty="dashed")  
  
legend(70, 12, legend=c("Male-White", "Female-White","Male-Black","Female-Black"),  
 col=c("red", "blue","black","green"), lty=1:2, cex=0.8)



Slope

## age   
## 0.02673734

Clearly interpret the slope for eaach line.

The slope for all the lines is 0.02673734. What is changing is the intercept values.