ST.310, Spring 2024

Homework #9

Due on Wednesday, 4/17/24

***Directions:***

* Follow the ***homework format guidelines*** shown on the syllabus.
* Do ***not*** fill in your source code and answers on this problem set. Use the homework template.
* You may work with one partner if you wish.
* Upload a single Word file (saved as ***doc*** or ***docx***) on Moodle by ***11:59 PM*** on the due date.
* Late homework will ***not*** be accepted without any legitimate excuses.

Load the UsingR package. The package includes all textbook datasets. Show your R command lines and outputs for each question.

**Problem#5.1.** [Page 196]

For the UScereal data set, create a scatter plot of calories modeled by sugars using the shelf variable to create different plot characters. Add a legend to indicate the shelf number. Is there any pattern? Explain. [10 pts]

data("UScereal")

ggplot(UScereal, aes(x = sugars, y = calories, shape = as.factor(shelf))) +

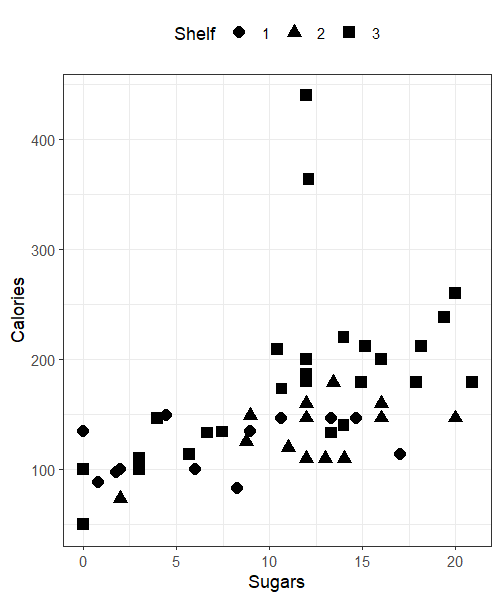
geom\_point(size = 3) +

labs(x = "Sugars", y = "Calories", shape = "Shelf") +

theme\_bw() +

theme(legend.position = "top")

There is a moderately linear pattern to this graph



**Problem#5.3.** [Page 196]

For the data set UScereal make a pairs plot of the numeric variables. Which correlation looks larger: fat and

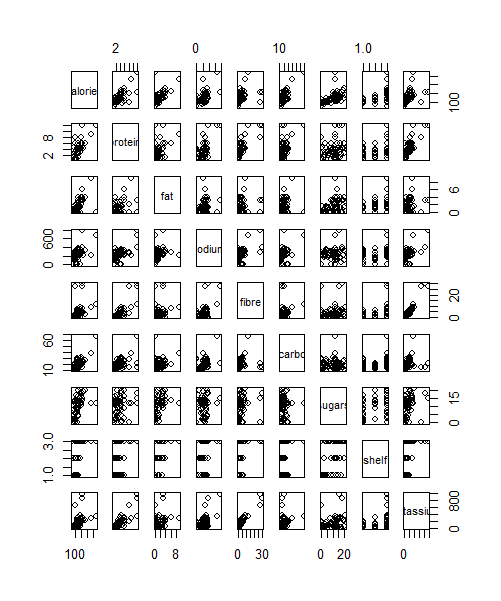
calories OR fat and sugars? Explain. [10 pts]

data("UScereal")

numericCheck<- UScereal[, sapply(UScereal, is.numeric)]

pairs(numericCheck)

Fat and calories because in compared to calories , sugar’s plots are more scattered than calories. While calories point are forming a line meaning the variables are more correlated to each other than sugar.



**Problem#11.4.** [Page 368]

The galton data set contains data collected by Francis Galton in 1885 concerning the influence a parent’s height has on a child’s height.

1. Fit a linear model for a child’s height (as the response) modeled by his parent’s height (as the predictor). Report the summary table of the model. [10 pts]\

model <- lm(child~ parent, data = galton)

summary(model)

Call:  
lm(formula = child ~ parent, data = galton)  
  
Residuals:  
 Min 1Q Median 3Q   
-7.8050 -1.3661 0.0487 1.6339   
 Max   
 5.9264   
  
Coefficients:  
 Estimate Std. Error  
(Intercept) 23.94153 2.81088  
parent 0.64629 0.04114  
 t value Pr(>|t|)   
(Intercept) 8.517 <2e-16 \*\*\*  
parent 15.711 <2e-16 \*\*\*  
---  
Signif. codes:   
 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05  
 ‘.’ 0.1 ‘ ’ 1  
  
Residual standard error: 2.239 on 926 degrees of freedom  
Multiple R-squared: 0.2105, Adjusted R-squared: 0.2096

1. Make a scatterplot for the two variables, including the regression line. [10 pts]

ggplot(galton, aes(x = parent, y = child)) +

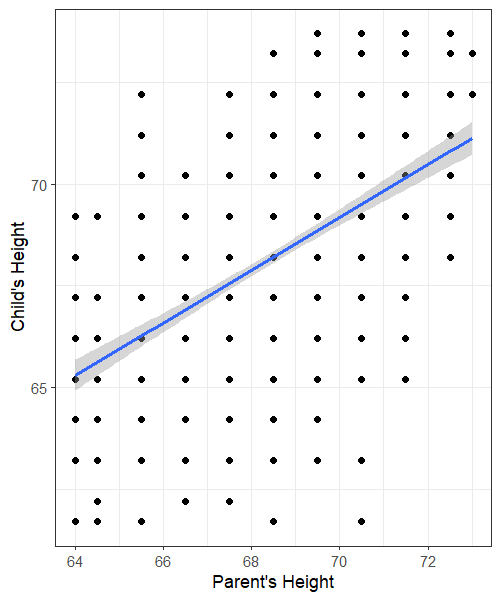
geom\_point() +

ylab("Child's Height") +

xlab("Parent's Height") +

geom\_smooth(method = 'lm') +

theme\_bw()



1. What is the value of *β*ˆ1, and why is this of interest? [10 pts]

model <- lm(child ~ parent, data = galton)

summary(model)

Call:

lm(formula = child ~ parent, data = galton)

Residuals:

Min 1Q Median 3Q

-7.8050 -1.3661 0.0487 1.6339

Max

5.9264

Coefficients:

Estimate Std. Error

(Intercept) 23.94153 2.81088

parent 0.64629 0.04114

t value Pr(>|t|)

(Intercept) 8.517 <2e-16 \*\*\*

parent 15.711 <2e-16 \*\*\*

---

Signif. codes:

0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05

‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.239 on 926 degrees of freedom

Multiple R-squared: 0.2105, Adjusted R-squared: 0.2096

F-statistic: 246.8 on 1 and 926 DF, p-value: < 2.2e-16

The B1 is 0.64629 which is the average change. b1 is of interest because it tells us the expected height of a child for every unit of change in parents height.

**Problem#11.10.** [Page 387]

For the homedata data set, find the regression equation to predict the year-2000 value of a home from its year-1970 value.

1. Make a prediction for an $80,000 home in 1970. [10 pts]

model <- lm(y2000 ~ y1970, data = homedata)

summary(model)

prediction <- predict(model, newdata = data.frame(y1970 = 80000))

print(prediction)

Call:

lm(formula = y2000 ~ y1970, data = homedata)

Residuals:

Min 1Q Median 3Q

-416665 -36308 809 34372

Max

536605

Coefficients:

Estimate Std. Error

(Intercept) -1.040e+05 2.337e+03

y1970 5.258e+00 3.147e-02

t value Pr(>|t|)

(Intercept) -44.51 <2e-16 \*\*\*

y1970 167.07 <2e-16 \*\*\*

---

Signif. codes:

0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05

‘.’ 0.1 ‘ ’ 1

Residual standard error: 58000 on 6839 degrees of freedom

Multiple R-squared: 0.8032, Adjusted R-squared: 0.8032

F-statistic: 2.791e+04 on 1 and 6839 DF, p-value: < 2.2e-16

Prediction

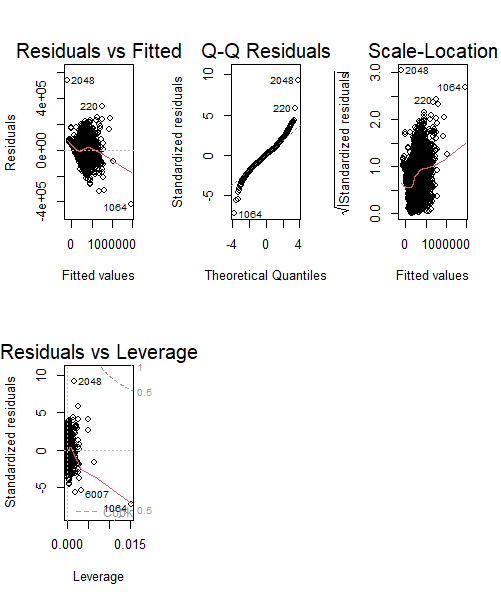
1   
316633.1

1. Draw the six diagnostic plots. Comment on the appropriateness of the regression model by investigating the plots. [10 pts]

data(homedata)

par(mfrow = c(2, 3))

plot(model)



**Problem#6.19.** [Page 235]

Assume that the average finger length for females is 3*.*20 inches, with a standard deviation of 0*.*35 inches, and that the distribution of lengths is normal. If a glove manufacturer makes a glove that fits fingers with lengths between 3*.*5 and 4 inches, what percent of the population will the glove fit? [10 pts]

**Note.** Use pnorm which is the normal distribution function of R.

mean <- 3.20

sd <- 0.35

lowerBound <- 3.5

upperBound <- 4

zLower <- (lowerBound - mean) / sd

zUpper <- (upperBound - mean) / sd

percentage <- pnorm(zUpper) - pnorm(zLower)

percentagePop <- percentage \* 100

print(percentagePop)

[1] 18.45475

The percent population is %1 8.45475

**Problem#6.22.** [Page 235]

For the fheight variable in the father.son data set, compute what percent of the data is within 1, 2, and 3 standard deviations from the mean. Compare to the percentages 68%, 95%, and 99*.*7%. [10 pts]

fheight <- father.son$fheight

mean <- mean(fheight)

sd <- sd(fheight)

percentageSd <- function(n) {

lowerBound <- mean- n \* sd

upperBound <- mean + n \* sd

percentage <- pnorm(upperBound, mean = mean, sd = sd) -

pnorm(lowerBound, mean = mean, sd = sd)

return(percentage \* 100)

}

percentage\_1<- percentageSd (1)

percentage\_2 <- percentageSd (2)

percentage\_3 <- percentageSd(3)

cat("Percentage within 1 standard deviation:", round(percentage\_1, 2), "%\n")

cat("Percentage within 2 standard deviations:", round(percentage\_2, 2), "%\n")

cat("Percentage within 3 standard deviations:", round(percentage\_3, 2), "%\n")

Percentage within 1 standard deviation: 68.27 %

Percentage within 2 standard deviations: 95.45 %

Percentage within 3 standard deviations: 99.73 %

**Problem#6.31.** [Page 240]

An elevator can safely hold 3500 pounds. A sign in the elevator limits the passenger count to 15. If the adult population has a mean weight of 180 pounds with a 25-pound standard deviation, how unusual would it be, if the central limit theorem applied, that an elevator holding 15 people would be carrying more than 3500 pounds? [10 pts]

**Note.** Use pnorm which is the normal distribution function of R.

meanWeight <- 180

sdWeight<- 25

capacity <- 3500

passenger\_limit <- 15

mean\_total\_weight <- meanWeight \* passenger\_limit

sd\_total\_weight <- sqrt(passenger\_limit) \* sdWeight

probability <- 1 - pnorm(capacity, mean = mean\_total\_weight, sd = sd\_total\_weight)

cat( round(probability \* 100, 2), "%.\n")

The probability that an elevator carrying 15 people would be carrying more than 3500 pounds, if the central limit theorem applied, is approximately 0 %. Which means that there's a 0% that 15 people in one evaluator will weight the equalivant amount of 3500 pounds.