Assignment3

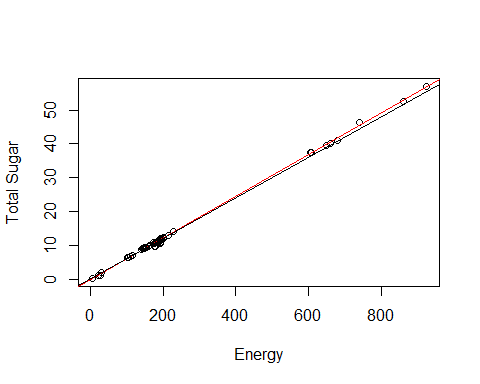
Syed Muhammad Adeel Ibrahim

May 12, 2019

## Question 1

You may have been told that juice drinks are “nothing but sugar.” Read in the dataset “cordial-simple.csv”, which has the nutrition information for several diﬀerent cordial brands and ﬂavours. Consider predicting Energy in kJ with grams of Total Sugar. 1. Suppose these drinks are indeed “nothing but sugar”, and sugar contains 16.7 kJ per gram. Formulate an appropriate null hypothesis for β0 and β1 that reﬂects this status quo. Make a scatter plot of the data; add both the line of best ﬁt and your hypothesized line.

juiceDrinks <- read.csv("cordial-simple.csv")  
plot(x = juiceDrinks$Energy, y = juiceDrinks$TotalSugar, xlab = "Energy", ylab = "Total Sugar")  
  
juiceDrinksMd <- lm(juiceDrinks$TotalSugar~juiceDrinks$Energy, data = juiceDrinks)  
juiceDrinksCoeff <- juiceDrinksMd$coefficients  
abline(juiceDrinksCoeff[1], juiceDrinksCoeff[2], col="Red")  
  
# Ho : beta0 = 0   
# Ha : beta0 != 0  
  
# Ho : beta1 = 16.7   
# Ha : beta1 != 16.7  
  
hypotesisData <- data.frame(TotalSugar = c(0, 1), Energy = c(0, 16.7))  
abline(lm(TotalSugar~Energy, data = hypotesisData))



# Juice Drinks are same as sugar

1. Fit a linear model and report the output of the “summary” function.

modJuice <- lm(Energy~TotalSugar, data=juiceDrinks)  
  
(modJuiceSummary <- summary(modJuice))

##   
## Call:  
## lm(formula = Energy ~ TotalSugar, data = juiceDrinks)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -12.701 -3.454 -1.811 3.018 18.592   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 4.57233 0.93856 4.872 7.09e-06 \*\*\*  
## TotalSugar 16.19326 0.04828 335.395 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 5.482 on 67 degrees of freedom  
## Multiple R-squared: 0.9994, Adjusted R-squared: 0.9994   
## F-statistic: 1.125e+05 on 1 and 67 DF, p-value: < 2.2e-16

1. Use the output to test the hypotheses in (1) against two sided alternatives. State your conclusions.

#Ho:beta1=1 vs Ha:Beta0 != 0  
  
coeffJuice <- modJuiceSummary$coefficients  
  
b1 <- coeffJuice[2, 1]  
seb1 <- coeffJuice[2, 2]  
  
test <- (b1 - 1) / seb1  
  
df <- nrow(juiceDrinks) - 2  
  
# 100 - 95 / 2 = 1.25  
(qtpoint <- qt(.975, df))

## [1] 1.996008

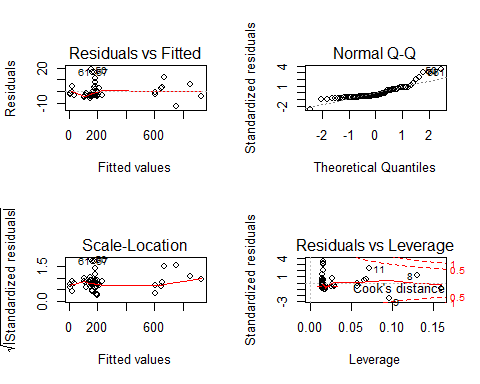
(ptJuice <- 2 \* pt(test, df))

## [1] 2

if(round(ptJuice, 3) == 0) {  
 paste("We can reject our null hypothesis since its not equal to zero")  
}

1. Comment on how close the relationship is between Total Sugar and energy, citing portions of your output in (2).
2. R square is approximately 1. There is close relationship. Deterministic relationship is seen between total sugar and energy. 0 to 20 total sugar, we see that the points lie on the line.
3. Produce and include diagnostic plots. What assumptions are they checking? Brieﬂy comment on your conclusions.

par(mfrow=c(2,2))  
plot(modJuice)

 A. 1.Residual vs Fitted - not too bad with just low variation. 2.Normal Q-Q - plot here helps to check whether sugar variable is normally distributed. Our graph is not very straight. Normality is clearly violation. 3.Scale location - It is not equally spread. Influential point lies in the trend. 4.The residuals and leverage - points are clustered in the left side excluding some points. This tells us that the slope coefficient changes.

1. Are there any drinks where the kJ and Total Sugar amounts given are inconsistent, allowing that there may be other sources of kJ? A. From scatter plot, It is known that there is inconsistency in few points to 16.7 KJ.

## Question 2

Your friend conjectures that citrus/blackcurrant ﬂavored juice drinks have more added sugar because of the natural tartness of these ﬂavors. 1. Perform a t-test comparing the added sugar for citrus/blackcurrant drinks vs other ﬂavors. State your null and alternative hypotheses, pvalue, and ﬁnal conclusion. State which test (equal or unequal variance) you used and why.

citrus <- juiceDrinks[juiceDrinks$Flavour=="Citrus-Blackcurrant",]  
others <- juiceDrinks[juiceDrinks$Flavour=="Other",]  
currantSugar = citrus$TotalSugar  
otherSugar = others$TotalSugar  
t.test(currantSugar, otherSugar, alternative="greater")

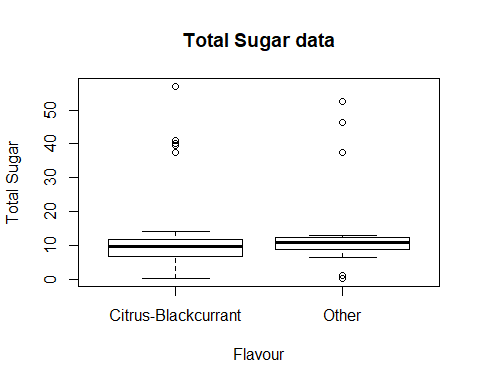
##   
## Welch Two Sample t-test  
##   
## data: currantSugar and otherSugar  
## t = -0.57396, df = 64.176, p-value = 0.716  
## alternative hypothesis: true difference in means is greater than 0  
## 95 percent confidence interval:  
## -7.51367 Inf  
## sample estimates:  
## mean of x mean of y   
## 12.95789 14.88065

var.test(currantSugar,otherSugar)

##   
## F test to compare two variances  
##   
## data: currantSugar and otherSugar  
## F = 0.9946, num df = 37, denom df = 30, p-value = 0.978  
## alternative hypothesis: true ratio of variances is not equal to 1  
## 95 percent confidence interval:  
## 0.4911561 1.9637947  
## sample estimates:  
## ratio of variances   
## 0.9946013

1. Produce a visualization suitable for comparing the added sugar in the two groups of drinks.

boxplot(TotalSugar~Flavour, data=juiceDrinks, main="Total Sugar data", xlab="Flavour", ylab="Total Sugar")

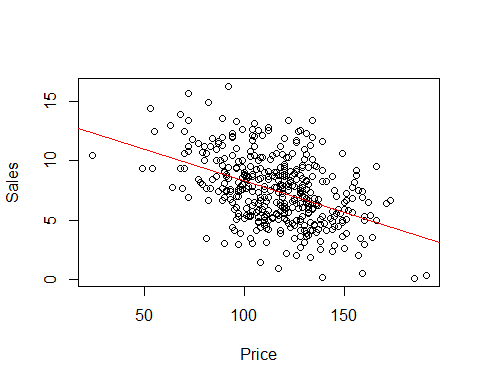


1. Are there serious deviations from the assumptions of the t-test? Explain. A. There is no serious deviation. As sample size is greater than 30.
2. Comment on whether the mean is a good way to characterize the typical amount of added sugar for these two ﬂavour groups. A. As we have a skewed data. Mean is not a great measure.

## Question 3

Consider the data in the ﬁle “carseats-simple.csv.” The response is the number of carseats of a particular brand sold (in thousands) at diﬀerent locations over the course of a year. The predictor is the price in dollars of the carseat at that location. The data has been collected by a consumer agency that has no role in setting the price. 1. Produce a scatterplot of the data, including the line of best ﬁt.

carseats <- read.csv("carseats-simple.csv")  
  
plot(x = carseats$Price, y = carseats$Sales, xlab = "Price", ylab = "Sales")  
  
carseatsModel <- lm(Sales~Price, data = carseats)  
carseatsCoeff <- carseatsModel$coefficients  
  
abline(carseatsCoeff[1], carseatsCoeff[2], col="Red")



1. Test for a relationship between the sales and the price. Give your conclusion, referring to relevant portions of your output.

(carseatsSummary <- summary(carseatsModel))

##   
## Call:  
## lm(formula = Sales ~ Price, data = carseats)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -6.5227 -1.8454 -0.1464 1.6496 7.5113   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 13.638582 0.632718 21.556 <2e-16 \*\*\*  
## Price -0.053042 0.005354 -9.908 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.532 on 398 degrees of freedom  
## Multiple R-squared: 0.1978, Adjusted R-squared: 0.1958   
## F-statistic: 98.16 on 1 and 398 DF, p-value: < 2.2e-16

A. Based on R square, it is very small. Correlation coefficient is negative. The relationship between sales and the price is moderate negative correlated.

1. What is our uncertainty about the average number of sales for locations where the price is $150? Give a conﬁdence interval for this quantity.

predict(carseatsModel, data.frame(Price = 150), interval = "confidence")

## fit lwr upr  
## 1 5.682271 5.24461 6.119931

1. Suppose we observe a location has set the price at $150. Give a prediction interval for the sales at that location. Explain why it is larger than the interval in (4).

predict(carseatsModel, data.frame(Price = 150), interval="prediction")

## fit lwr upr  
## 1 5.682271 0.685397 10.67914

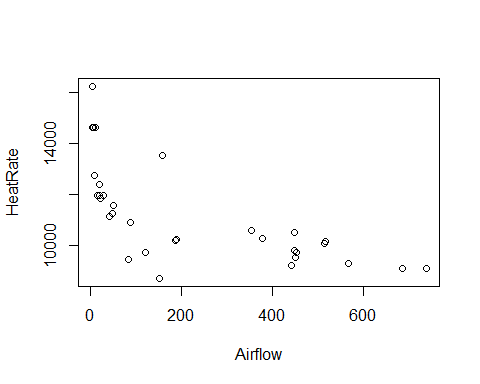
A. The standard error on the prediction is larger therefore the prediction interval will be wider than the confidence interval.

1. Suppose we change the price at a location previously charging $200 to $150. Should we use the interval above to predict sales at the new price? Explain any reservations you have. A. Observational value is not same as experimental data. In observational value we cannot know the causality. May be there are not many people and the price is set for $200 to cover the cost. Different levels may be set for different places. If we change the price we get other, this we get from experimental set up.

## Question 4

Consider the data in the ﬁle “turbine-simple.csv.” A gas turbine is an internal combustion engine used to generate electricity. A key measure of how well the turbine works is the heat rate, the energy input divided by the electricity generated (with lower heat rate being better). We are interested in how it is related to the airﬂow through the turbine. 1. Produce a scatterplot of the data. Describe the relationship between airﬂow and heat rate.

turbine <- read.csv("turbine-simple.csv")  
plot(HeatRate~Airflow, data=turbine)

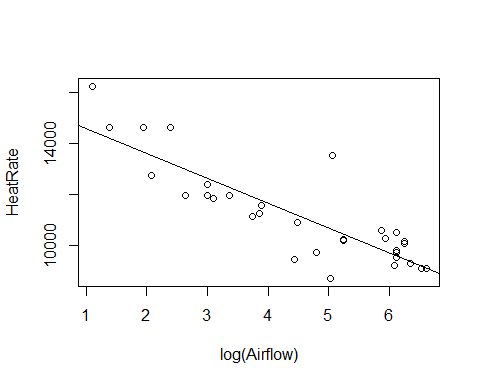


turbineModel <- lm(HeatRate~Airflow, data=turbine)  
summary(turbineModel)

##   
## Call:  
## lm(formula = HeatRate ~ Airflow, data = turbine)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2860.3 -799.4 -139.4 540.6 3840.2   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 12419.486 357.940 34.697 < 2e-16 \*\*\*  
## Airflow -5.561 1.116 -4.982 2.45e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1435 on 30 degrees of freedom  
## Multiple R-squared: 0.4527, Adjusted R-squared: 0.4345   
## F-statistic: 24.82 on 1 and 30 DF, p-value: 2.452e-05

1. Try a few diﬀerent transformations of airﬂow, with the aim of making the relationship more linear. Show plots of the diﬀerent transformations you have tried (at least three) and indicate which you like the best.

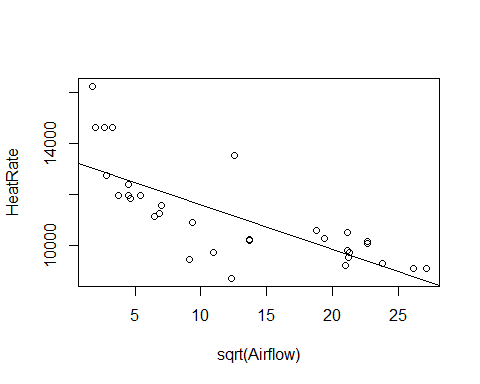
turbineModel1 <- lm(HeatRate~log(Airflow), data=turbine)  
plot(HeatRate~log(Airflow), data=turbine)  
abline(turbineModel1)



summary(turbineModel1)

##   
## Call:  
## lm(formula = HeatRate ~ log(Airflow), data = turbine)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1935.7 -578.7 -118.1 539.2 2911.2   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 15556.7 509.5 30.536 < 2e-16 \*\*\*  
## log(Airflow) -976.7 106.3 -9.191 3.14e-10 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 992.7 on 30 degrees of freedom  
## Multiple R-squared: 0.7379, Adjusted R-squared: 0.7292   
## F-statistic: 84.47 on 1 and 30 DF, p-value: 3.144e-10

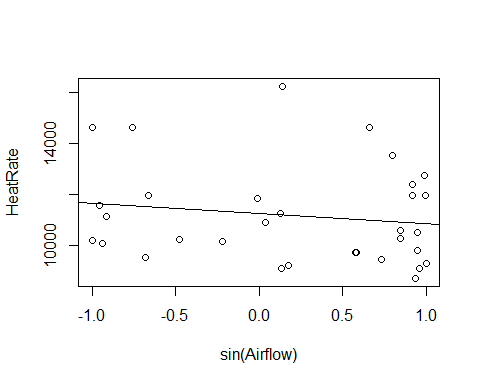
turbineModel2 <- lm(HeatRate~sqrt(Airflow), data=turbine)  
plot(HeatRate~sqrt(Airflow), data=turbine)  
abline(turbineModel2)



summary(turbineModel2)

##   
## Call:  
## lm(formula = HeatRate ~ sqrt(Airflow), data = turbine)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2498.4 -744.0 -117.1 559.8 3184.7   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 13360.00 410.30 32.561 < 2e-16 \*\*\*  
## sqrt(Airflow) -174.19 27.27 -6.387 4.76e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1262 on 30 degrees of freedom  
## Multiple R-squared: 0.5762, Adjusted R-squared: 0.5621   
## F-statistic: 40.79 on 1 and 30 DF, p-value: 4.756e-07

turbineModel3 <- lm(HeatRate~sin(Airflow), data=turbine)  
plot(HeatRate~sin(Airflow), data=turbine)  
abline(turbineModel3)



summary(turbineModel3)

##   
## Call:  
## lm(formula = HeatRate ~ sin(Airflow), data = turbine)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2153.2 -1466.1 -410.7 1080.4 5055.5   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 11244.6 351.9 31.95 <2e-16 \*\*\*  
## sin(Airflow) -404.3 464.5 -0.87 0.391   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1915 on 30 degrees of freedom  
## Multiple R-squared: 0.02463, Adjusted R-squared: -0.00788   
## F-statistic: 0.7576 on 1 and 30 DF, p-value: 0.391

1. Fit a linear model using the transformation you determined in (2). Show the prediction from this model as a curve on your scatterplot from (1). Create a prediction interval for the heat rate when the airﬂow is 300. Add this interval to your scatterplot as well.

par(c(1,2));

## NULL

predict(turbineModel1, data.frame(Airflow = 300), interval="prediction")

## fit lwr upr  
## 1 9985.567 7910.194 12060.94

predictValue <- predict(turbineModel1, data.frame(Airflow = seq(0, 800, .01)), interval="prediction")  
  
  
plot(HeatRate~Airflow, data=turbine, ylim=c(7000, 20000))  
lines(seq(0, 800, .01), predictValue[,1], col="blue")  
lines(seq(0, 800, .01), predictValue[,2], col="red")  
lines(seq(0, 800, .01), predictValue[,3], col="red")

