Julia Nelson

Professor Li

December 7, 2019

“I pledge my honor that I have abided by the Steven’s Honor System.”

Final Project

Simple and Multiple Linear Regression

In a study of cheddar cheese from the LaTrobe Valley of Victoria, Australia, samples of cheese were analyzed for their chemical compositions and received taste tests. As a student enrolled in MA331 Intermediate Statistics at Stevens Institute of Technology, I was asked to investigate the data of CHEESE for the different present levels of chemicals effect on Taste. This report discusses my findings.

Data Set

LaTrobe Valley collected 30 cases for their data set. In this set, there are three chemical’s concentrations taken into account for the explanatory variables; the natural log of acetic acid (Acetic), the natural log of hydrogen sulfide (H2S), and the concentration of lactic acid (Lactic). My analysis is attached to this report in case you need further details.

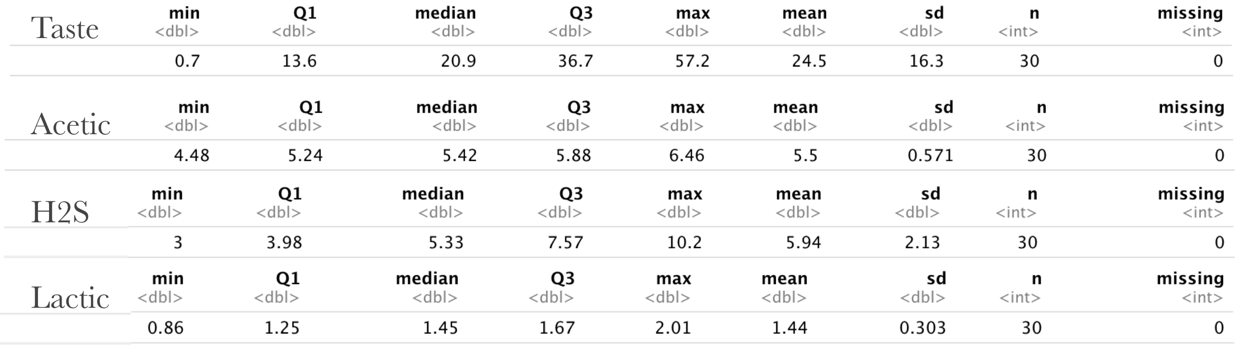
Software

My analysis was performed entirely by RStudio, using the mosaic package add-in.

Analysis

To begin my analysis, I first found the mean, median, standard deviation and interquartile ranges for the 4 variables, taste , acetic, h2s, and lactic shown in Figure 1.

Figure 1:



Taste

min	Q1	median	Q3	max	mean	sd	n	missing
<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<int>	<int>
0.7	13.6	20.9	36.7	57.2	24.5	16.3	30	0

Acetic

min	Q1	median	Q3	max	mean	sd	n	missing
<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<int>	<int>
4.48	5.24	5.42	5.88	6.46	5.5	0.571	30	0

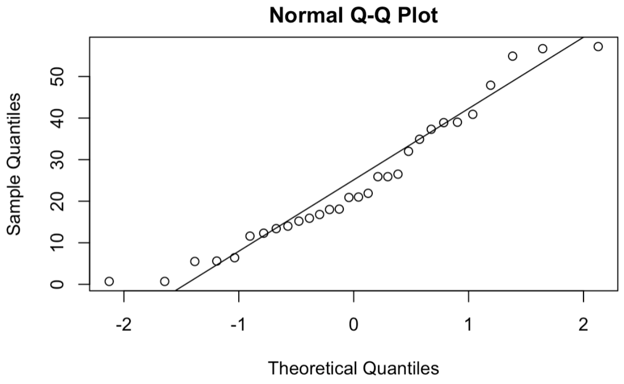
H2S

min	Q1	median	Q3	max	mean	sd	n	missing
<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<int>	<int>
3	3.98	5.33	7.57	10.2	5.94	2.13	30	0

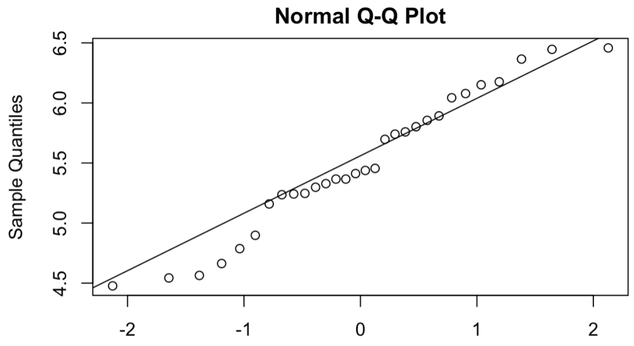
Lactic

min	Q1	median	Q3	max	mean	sd	n	missing
<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<int>	<int>
0.86	1.25	1.45	1.67	2.01	1.44	0.303	30	0

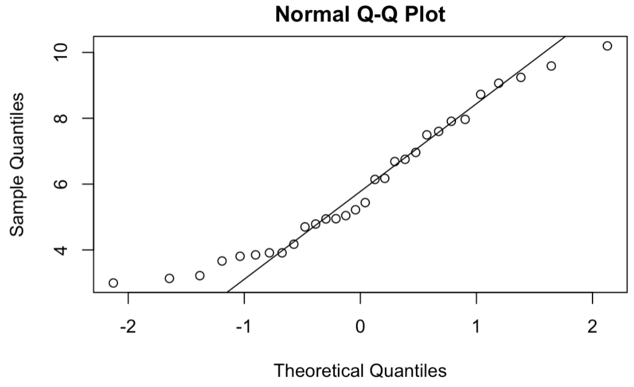
These variables can also visually be compared by their Distributions. Below are the QQ-Plots of Taste, Acetic, H2S, and Lactic.



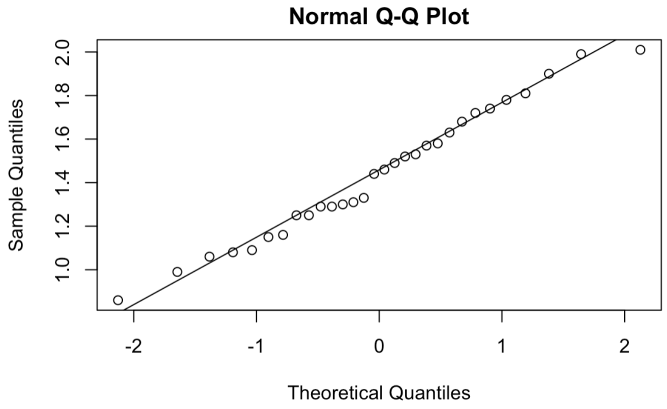
For Taste(left), I found the distribution is skewed slightly right, but it is still approximately Normal.



For Acetic (left), I found that the distribution is approximately Normal.



For H2S (left), I found the distribution to be skewed slightly right but approximately Normal.

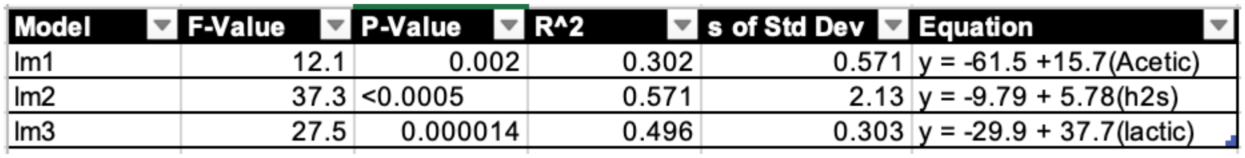


For Lactic (left), I found the distribution to be Normal.

When taking all of variables, it is important to look for correlations between any of them. To do so, I paired all of the variable and created their scatterplots for a visual representation of the calculated correlations I found. The strongest correlation found was between Taste and H2S, followed by Taste and Lactic, with the weakest being between Taste and Acetic. This correlation will be very helpful in the decision for which linear regression model is best to use.

I then calculated the simple linear regressions between Taste as the Response Variable and the chemical variables as the explanatory variables to find their equations.

$Taste = -61.5 + 15.7Acetic$
$Taste = -9.79 + 5.78H2S$
$Taste = -29.9 + 37.7Lactic$
The table of the comparison of the 3 simple linear regressions is below.



Model	F-Value	P-Value	R^2	s of Std Dev	Equation
lm1	12.1	0.002	0.302	0.571	y = -61.5 +15.7(Acetic)
lm2	37.3	<0.0005	0.571	2.13	y = -9.79 + 5.78(h2s)
lm3	27.5	0.000014	0.496	0.303	y = -29.9 + 37.7(lactic)

Summary

After I calculated the multiple linear regressions of Taste with Acetic and H2S, Acetic and Lactic, H2S and Lactic, and finally Acetic, H2S and Lactic. I found that the acetic variable linear regression provides little detail compared to the others, due to its correlation with H2S and its high P value indicating a very close to zero coefficient.

Based off of the outputs of the linear regression models, the best model to use is the Multiple Linear Regression of Taste with explanatory variable H2S and Lactic because of its high R-squared vale (variation of Taste) and its lack of unnecessary variables like Acetic.

The R Code and more details are stated on the following page.

```
require(mosaic)
trellis.par.set(theme=col.mosaic())
options(digits = 3)

View(ex11_53cheese)

# Question 11.53
  ds = ex11_53cheese
  names(ds)

  # Find mean, median, std, interquartile range
  favstats(ds$taste)  # Taste:  Mean= 24.5, Median= 20.9 StandardDev=
16.3, InterquartileRange= (13.6,36.7)
  favstats(ds$acetic) # Acetic: Mean= 5.50, Median= 5.42 StandardDev=
0.571, InterquartileRange= (5.24,5.88)
  favstats(ds$h2s)    # H2S:    Mean= 5.94, Median= 5.33 StandardDev=
2.13, InterquartileRange= (3.98,7.57)
  favstats(ds$lactic) # Lactic: Mean= 1.44, Median= 1.45 StandardDev=
0.303, InterquartileRange= (1.25,1.67)

  # Display each distribution's stemplot and Normal Quantile Plot
    # Taste:
      stem(ds$taste)
      qqnorm(ds$taste)
      qqline(ds$taste)
      # Normality: The taste Distribution is skewed slightly right but is
approximately Normal.
    # Acetic:
      stem(ds$acetic)
      qqnorm(ds$acetic)
      qqline(ds$acetic)
      # The acetic Distribution is approximately Normal.
    # H2S:
      stem(ds$h2s)
      qqnorm(ds$h2s)
      qqline(ds$h2s)
      # The H2S Distribution is skewed slightly right but is
approximately Normal.
    # Lactic:
      stem(ds$lactic)
      qqnorm(ds$lactic)
      qqline(ds$lactic)
      # The Lactic Distribution is approximately Normal




# Question 11.54

  # Scatterplot's of each pair of variables
      # taste & acetic
```

```
        plot(subset(ds, select = c("taste","acetic")), main =
"Scatterplot Taste vs. Acetic")
      # taste & h2s
        plot(subset(ds, select = c("taste","h2s")), main = "Scatterplot
Taste vs. H2S")
      # taste & lactic
        plot(subset(ds, select = c("taste","lactic")), main =
"Scatterplot Taste vs. Lactic")
      # acetic & h2s
        plot(subset(ds, select = c("acetic","h2s")), main = "Scatterplot
Acetic vs. H2S")
      # acetic & lactic
        plot(subset(ds, select = c("acetic","lactic")), main =
"Scatterplot Acetic vs. Lactic")
      # h2s & lactic
        plot(subset(ds, select = c("h2s","lactic")), main = "Scatterplot
H2S vs. Lactic")

  # Correlation & P-value for 0-pop correlation
    cor.test(ds$taste,ds$acetic)
    cor.test(ds$taste,ds$h2s)
    cor.test(ds$taste,ds$lactic)
    cor.test(ds$acetic,ds$h2s)
    cor.test(ds$acetic,ds$lactic)
    cor.test(ds$h2s,ds$lactic)

  # Outcome Summary:
    # Pair          Correlation  P-Values     (closer to 1 = stronger)
    # taste + acetic  = 0.550   = 0.002        the weakest linear
relationship
    # taste + h2s     = 0.756   = 0.000001     the strongest positive
linear relationship
    # taste + lactic  = 0.704   = 0.00001      positive linear
relationship
    # acetic + h2s    = 0.618   = 0.0003       positive linear
relationship
    # acetic + lactic = 0.604   = 0.0004       positive linear
relationship
    # h2s + lactic    = 0.645   = 0.0001       positive linear
relationship
      # p-values of zero pop correlation are very very small, can reject
hypothesis
      # that the correlations on pop are not zero



# Question 11.55
    # simple linear regression
    # taste - response variable
    # acetic - explanatory variable
    acetic = ds$acetic
    taste = ds$taste
    lm1 = lm(taste~acetic, data = ds)
    coef(lm1)
    rsquared(lm1)
    summary(lm1)
    anova(lm1)
    # plot of the data
      plot(acetic, taste, xlab = "Acetic", ylab = "Taste", main = "Plot 
Taste vs Acetic")
    # add the found least-squares regression line
      abline(coef(lm1), lwd =2, lty = 2, col="red")
    # Summary:
      # Equation: hatTaste = -61.5 + 15.7acetic
      # R-squared = 0.302
      # F-statistic = 12.1
      # acetic: coef= 15.7, t-value = 3.48,  std err = 4.5
      # intercept: coef= -61.5, t-value = -2.48,  std err = 24.9
      # P-value = 0.002

  # The residuals:
    # Residuals
        plot(lm1, which =1)
        plot(lm1, which=2)
      # The Residuals have a relatively Normal distribution
    # Plot of acetic Residuals vs H2S
      h2s = ds$h2s
      xyplot(residuals(lm1) ~ h2s, type=c("p", "r", "smooth"), main = 
"Residuals v H2S")
      # Residuals and H2S are positively associated  with a more random 
pattern
    # Plot of acetic Residuals vs Lactic
      lactic = ds$lactic
      xyplot(residuals(lm1) ~ lactic, type=c("p", "r", "smooth"),  main = 
"Residuals v Lactic")
      # The residuals seem to be slightly positively associated with 
lactic but with a random patter
    # Residuals Summary:
      # Min: -29.64
      # 1Q: -7.44
      # Median: 2.08
      # 3Q: 6.60
      # Max: 26.58
      # res std error: 13.8 on 28 df




# Question 11.56
  # simple linear regression
  # taste - response var
  # h2s - explanatory var
  taste = ds$taste
  h2s = ds$h2s
  lm2 = lm(taste~h2s, data = ds)
  coef(lm2)
  rsquared(lm2)
  summary(lm2)
  anova(lm2)
  # plot of the data
    plot(h2s, taste, xlab = "H2S", ylab = "Taste", main = "Plot Taste vs
H2S")
  # add the found least-squares regression line
    abline(coef(lm2), lwd =2, lty = 2, col="red")
  # Summary:
    # Equation: hatTaste = -9.79 + 5.78h2s
    # R-squared = 0.571
    # F-statistic = 37.3
    # h2s: coef= 5.776, t-value = 6.11,  std err = 0.946
    # intercept: coef= -9.787, t-value = -1.64,  std err = 5.958
    # P-value = 0.00000137 = < 0.0005

  # The residuals:
    # Residuals
      plot(lm2, which=1)
      plot(lm2, which=2)
      # The residuals are of Normal Distribution
    # Plot of H2S residuals vs acetic
      acetic = ds$acetic
      xyplot(residuals(lm2) ~ acetic, type=c("p", "r", "smooth"), main =
"Residuals v Acetic")
      # There is no real pattern between the residuals and lactic
    # Plot of H2S residuals vs lactic
      lactic = ds$lactic
      xyplot(residuals(lm2) ~ lactic, type=c("p", "r", "smooth"),  main =
"Residuals v Lactic")
      # There is possibly a  pattern between the residuals and lactic
meaning the fit could be better
    # Residuals Summary:
      # Min: -15.43
      # 1Q: -7.61
      # Median: 3.49
      # 3Q: 6.42
      # Max: 25.69
      # res std error: 10.8 on 28 df




# Question 11.57
  # simple linear regression
  # taste - response var
  # lactic - explanatory var
  taste = ds$taste
  lactic = ds$lactic
  lm3 = lm(taste~lactic, data = ds)
  coef(lm3)
  rsquared(lm3)
  summary(lm3)
  anova(lm3)
    # plot of the data
    plot(lactic, taste, xlab = "Lactic", ylab = "Taste", main = "Plot
Taste vs Lactic")
    # add the found least-squares regression line
    abline(coef(lm3), lwd =2, lty = 2, col="red")
  # Summary:
    # Equation: hatTaste =  -29.9 + 37.7lactic
    # R-squared = 0.496
    # F-statistic = 27.5
    # lactic: coef= 37.72, t-value = 5.25,  std err = 7.19
    # intercept: coef= -29.86, t-value = -2.82,  std err = 10.58
    # P-value = 0.000014

  # The residuals:
    # Residuals
      plot(lm3, which=1)
      plot(lm3, which =2)
      # The residuals are of Normal Distribution
    # Plot of Lactic residuals vs acetic
      acetic = ds$acetic
      xyplot(residuals(lm3) ~ acetic, type=c("p", "r", "smooth"), main =
"Residuals v Acetic")
      # There pattern present between the residuals and acetic
    # Plot of Lactic residuals vs H2S
      h2s = ds$h2s
      xyplot(residuals(lm3) ~ h2s, type=c("p", "r", "smooth"),  main =
"Residuals v H2S")
      # There pattern present between the residuals and H2S
  # Residuals Summary:
      # Min: -19.94
      # 1Q: -8.68
      # Median: -0.11
      # 3Q: 9.00
      # Max: 27.43
      # res std error: 11.7 on 28 df



# Question 11.58
  # Comparing the  3 simple linear regressions
  # (Table Made in excel)
  # The 3 regression equations are:
      # hatTaste = -61.5 + 15.7Acetic
      # hatTaste = -9.79 + 5.78H2S
      # hatTaste =  -29.9 + 37.7Lactic
    # The intercepts all differ because they are comparing on differently
scaled variables
    #  the intercept changes depensing on when x = 0, and in each
equation, x has a different coefficient



# Question 11.59
    # multiple linear regression Taste using acetic + h2s
    acetic = ds$acetic
    taste = ds$taste
    h2s = ds$h2s
    lm4 = lm(taste~ acetic + h2s, data = ds)
    coef(lm4)
    rsquared(lm4)
    summary(lm4)
    anova(lm4)
    plot(lm4)
    # Summary:
      # Equation: hatTaste =  -26.94 + 3.80Acetic + 5.15H2S
      # R-squared = 0.551
      # F-statistic = 18.8
      # acetic: coef= 3.80, t-value = 0.84,  std err = 4.51, P = 0.406
      # h2s: coef= 5.15, t-value = 4.26,  std err = 1.21, P = 0.00022
      # intercept: coef= -26.94, t-value = -1.27,  std err = 21.19, P =
0.21454
    # Compared to the simple linear regression of Taste~Acetic,
    # where t = 3.48, P = 0.002, the Multiple linear regression including
H2S
    # gave Acetic t = 0.84, P= 0.406.
    # Because the two are correlated, not
    # I would prefer the multiple linear regression model however,
because
    # it only a slightly  better fit and provides 55.1% of the variation
in taste
    # versus the simple linear regression's 30.2% variation.



# Question 11.60
    # mult linear reg Taste using Lactic + h2s
    lactic = ds$lactic
    taste = ds$taste
    h2s = ds$h2s
    lm5 = lm(taste~ h2s + lactic, data = ds)
    coef(lm5)
    rsquared(lm5)
    summary(lm5)
    anova(lm5)
    plot(lm5)

  # Summary :
  # Compared to the results to the simple linear regressions using each
  #   of these variables alone, it is evident that a better result
  #   is obtained by using both predictors in a model.
    # In this model,
      # Intercept: coef= -27.59, std Error= 8.98, t-value= -3.07, Pr(>|
t|)= 0.0048
      # H2S: coef= 3.95, std Error= 1.14, t-value= 3.47, Pr(>|t|)= 0.0017
      # Lactic: coef= 19.89, std Error= 7.96, t-value= 2.50, Pr(>|t|)=
0.0188
    # In the individual simple linear regressions:
      # H2S: coef= 5.776,  std err = 0.946, t-value = 6.11, Pr(>|t|)=
0.0000014
```

```
      # Lactic: coef= 37.72, std err = 7.19, t-value = 5.25, Pr(>|
t|)=0.000014
    # Comparing the outputs, it is clear that for both variables the
multiple
    # linear regression model is better. You can see that the R-squared
value (= 0.652)
    # is also higher than the individual regression's (0.571, 0.496)
providing more information
    # on the variation of taste. In addition, both have higher P values
in the
    # multiple linear regression signifying the coeffients are
significantly different than 0.



# Question 11.61
    # multiple linear regression - all
    lactic = ds$lactic
    taste = ds$taste
    h2s = ds$h2s
    acetic = ds$acetic

    lmALL = lm(taste~ acetic+ h2s + lactic, data = ds)
    coef(lmALL)
    rsquared(lmALL)
    summary(lmALL)
    anova(lmALL)
    plot(lmALL)
    # write a short summary of your results, including an examination
    # of the residuals. Based on all the regression analyses you
    # have carried out on these data, which model do you prefer and
    # why?
    # Summary:
      # Equation: hatTaste =  -28.877 + 0.328Acetic + 3.912H2S +
19.671Lactic
      # R-squared = 0.652
      # F-statistic = 16.2
      # acetic:  t-value = 0.07, P = 0.9420
      # h2s:  t-value = 3.13, P = 0.0042
      # lactic:  t-value = 2.28, P = 0.0331
      # intercept:  t-value = -1.46, P = 0.1554
    # Residuals:
      plot(lmALL, which=1)
      plot(lmALL, which =2)
      # The residuals seem to be Normally distributed

      # scatter Plot of  residuals vs acetic
        xyplot(residuals(lmALL) ~ acetic, type=c("p", "r", "smooth"),
main = "Residuals v Acetic")
      # scatter Plot of  residuals vs H2S
        xyplot(residuals(lmALL) ~ h2s, type=c("p", "r", "smooth"),  main
= "Residuals v H2S")
      # scatter Plot of  residuals vs Lactic
```

```
        xyplot(residuals(lmALL) ~ lactic, type=c("p", "r", "smooth"),
main = "Residuals v Lactic")
      # No significant patterns seen in any of the residuals v variables
scatterplots

      # Comparing the R-squared of lmALL to lm5, they both = 0.652, but
because Acetic is very close to 0 with P= 0.9420,
        # making it almost unnecessary to include in the model. The best
model to choose would be lm5, with H2S and Lactic,
        # and not lmALL because of the unnecessary information of acetic.
```