Project #2

Dustin Leatherman

11/3/2019

1

Complete the following for the model $E(taste) = \beta_0 + \beta_1 Acetic \beta_2 H2S + \beta_3 Lactic$

a

Find the estimated regression model

```
cheese <- read.csv("~/Downloads/cheese.csv")
cheese.model1 <- lm(taste ~ Acetic + H2S + Lactic, data = cheese)
cheese.model1 %>% tidy %>% kable %>% kable_styling(latex_options = "hold_position")
```

term	estimate	std.error	statistic	p.value
(Intercept)	-28.8767696	19.735418	-1.4631952	0.1553991
Acetic	0.3277413	4.459757	0.0734886	0.9419798
H2S	3.9118411	1.248430	3.1334077	0.0042471
Lactic	19.6705434	8.629055	2.2795710	0.0310795

$\hat{Y}_i = -28.8768 + 0.3277$ Acetic $+ 3.9118$ H2S $+ 19.6705$ Lactic

b

Perform an overall F-test to determine if the model is useful for estimating taste. Use $\alpha = 0.05$

```
glance(cheese.model1) %>%
  mutate(numDf = df - 1) %>%
  select("F*" = statistic, numDf, "denDf" = df.residual, p.value) %>%
  kable %>% kable_styling(latex_options = "hold_position")
```

F*	numDf	denDf	p.value
16.22143	3	26	3.8e-06

There is **convincing** evidence that at least one of the parameters has a significant effect on determining taste score (Omnibus F-Test. p-value = 3.81e-06).

c

How does each of the predictor variables affect taste? Give the correct interpretation for each of the predictor variables using the estimated regression model. Make sure to fully explain your answers.

A one-unit increase in mean taste is associated with a 0.3277 increase in mean Acetic Acid, given that H2S and Lactic acid are fixed values.

A one-unit increase in mean taste is associated with a 3.9118 increase in mean H2S, given that Acetic acid and Lactic acid are fixed values.

A one-unit increase in mean taste is associated with a 19.6705 increase in mean Lactic Acid, given that H2S and Acetic acid are fixed values.

d

Use a t-test for acetic acid. Perform the same test using a partial F-test.

```
tidy(cheese.model1) %>% kable
```

term	estimate	std.error	statistic	p.value
(Intercept)	-28.8767696	19.735418	-1.4631952	0.1553991
Acetic	0.3277413	4.459757	0.0734886	0.9419798
H2S	3.9118411	1.248430	3.1334077	0.0042471
Lactic	19.6705434	8.629055	2.2795710	0.0310795

```
cheese.model.no_acetic <- lm(taste ~ H2S + Lactic, data = cheese)
anova(cheese.model1, cheese.model.no_acetic) %>% kable %>% kable_styling(latex_options = "hold_position
```

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
26	2668.411	NA	NA	NA	NA
27	2668.965	-1	-0.5542675	0.0054006	0.9419798

Per the table in (1a), given H2S and Lactic are in the model, there is **no** evidence that Acetic acid has an associated effect on taste (two-tailed t-test. p-value = 0.942). A partial F-Test for a model with and without Acetic acid supports this (p-value = 0.942).

e

The tests for acetic acid in the previous part resulted in large P-values; however, the same type of test performed for the model $E(taste) = \beta_0 + \beta_1$ Acetic had a low p-value. Why does this occur?

```
cheese.model.only_acetic <- lm(taste ~ Acetic, data = cheese)
cheese.model.only_acetic %>%
  tidy %>%
  kable %>%
  kable_styling(latex_options = "hold_position")
```

term	estimate	std.error	statistic	p.value
(Intercept)	-61.49861	24.846379	-2.475154	0.0196373
Acetic	15.64777	4.495773	3.480551	0.0016582

```
anova(cheese.model1, cheese.model.only_acetic) %>%
  kable %>% kable_styling(latex_options = "hold_position")
```

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
26	2668.411	NA	NA	NA	NA
28	5348.745	-2	-2680.334	13.05809	0.0001186

The two-tail t-test with a Type 3 Sums of Squares provides a mechanism for assessing the importance of predictors **given that other predictors are already in the model**. The previous tests describe the effect of Acetic acid when H2S and Lactic acid are already in the model. In a Simple Linear Model where Acetic

acid is the sole predictor, it is considered significant to the model even despite it not being significant in the presence of some other predictors.

The low p-value for the partial F-test suggests that the more complicated model is a better fit for the data than the Acetic-only model.

f

> For the mean values of acetic acid, hydrogen sulfide, and lactic acid, estimate the value of taste and find a 95% confidence interval for the mean response. Compare your answer here to what we be obtained using the Acetic only model. Discuss why the difference in Confidence Intervals occur.

```
# get mean values for predictors
newData <- cheese %>%
  select("Acetic", "H2S", "Lactic") %>%
  colMeans %>%
  enframe %>%
  pivot_wider

# set alpha
myAlpha <- 0.05
g <- newData %>% length

# bonferroni adjusted confidence level
myLevel <- 1 - myAlpha / g

# C.I for a mean response using mean responses
predict(cheese.model1, newdata = newData, interval = "confidence", level = myLevel) %>%
  as_tibble %>%
  select(fit, lwr, upr) %>%
  kable %>% kable_styling(latex_options = "hold_position")
```

fit	lwr	upr
24.53333	19.8003	29.26637

```
cheese.model.only_acetic <- lm(taste ~ Acetic, data = cheese)

predict(cheese.model.only_acetic, newdata = data.frame(Acetic = mean(cheese$Acetic)), interval = "confi
  as_tibble %>%
  select(fit, lwr, upr) %>%
  kable %>% kable_styling(latex_options = "hold_position")
```

fit	lwr	upr
24.53333	19.36438	29.70229

It is estimated that the mean taste score is 24.533 when Acetic Acid is 5.498, H2S is 5.9418, and Lactic Acid is 1.442. With 95% confidence, the mean taste score for the aforementioned predictor values is between 19.8003 and 29.2664 after adjusting for multiple variables with the Bonferonni adjustment.

The estimates mean taste score in the Acetic-only model is 24.5333. With 95% confidence, the Acetic-only model suggests that the mean taste score falls between 19.3644 and 29.7023.

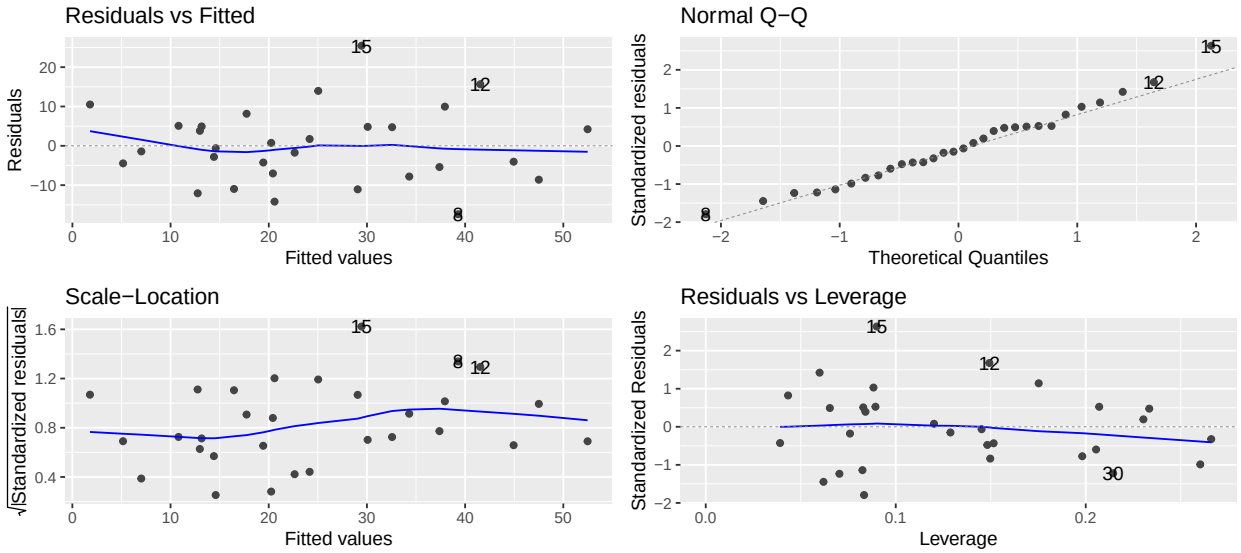
The confidence intervals are similar but differ slightly for a couple reasons. The first being that there is only one predictor in the Acetic-only model so the bonferroni adjustment to the confidence level is not needed.

Using the Bonferroni adjustment is a good approximation to a particular confidence level, but it is not perfect. The second reason being that H2S and Lactic are not accounted for in the Acetic-only model meaning that the data will not be present in the model meaning that the numbers will be different.

g

Comment on 1) linearity of the regression model, 2) constant-error variance, 3) outliers, and 4) normality of ϵ_i

```
autoplot(cheese.model1)
```



Linearity of the Regression Model

There are no discernible patterns in the Residual plot indicating that a linear model is appropriate for this data. This is confirmed by the smoother line hovering near 0.

Constant-error Variance

There is one main outlier on the residual plot where the residual is greater than 20. Barring that, there is a constant spread around y = 0 indicating that there is constant error variance. This is confirmed by the Brown-Forsythe test (p-value = 0.2232). It is also be confirmed by the Breusch-Pagan Test since the data are normal (p-value = 0.2387).

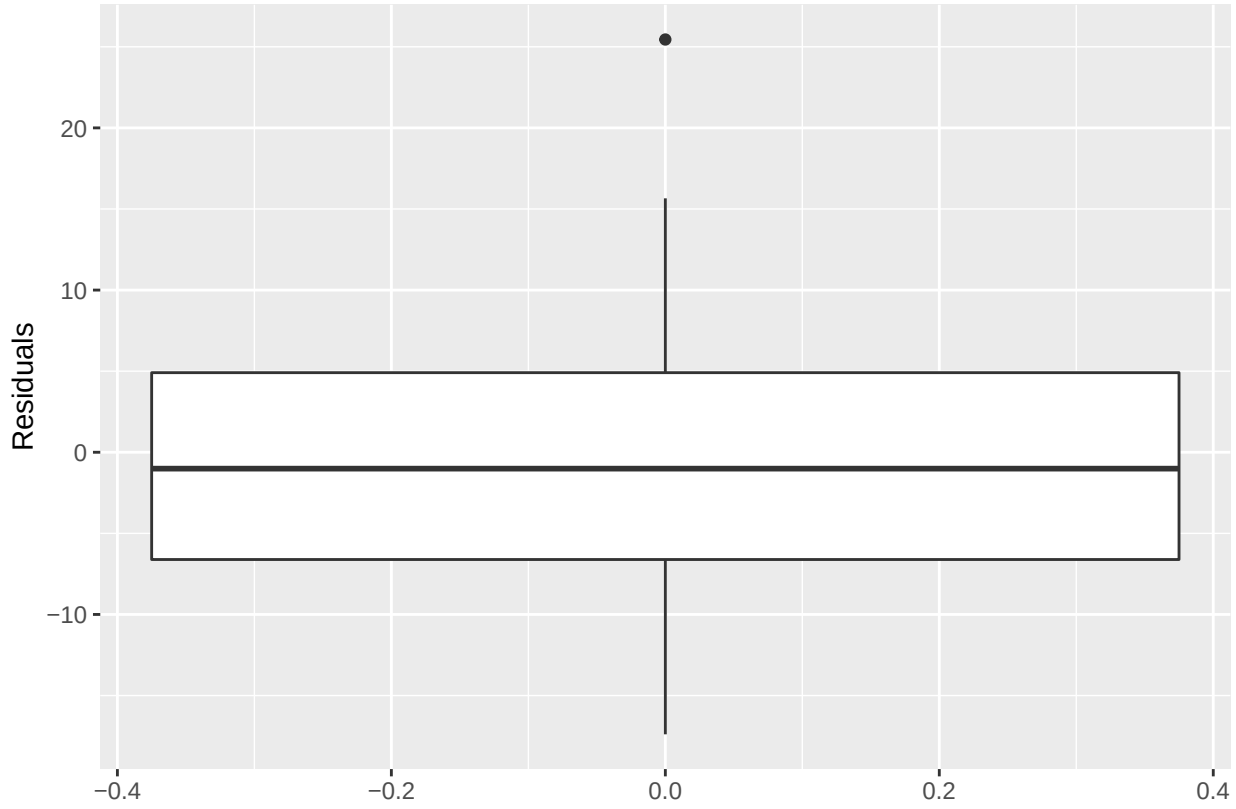
```
levene.test(cheese.model1 %>% residuals, group = cheese$taste <= median(cheese$taste)) %>%
  tidy %>%
  # concat breusch-pagan test
  bind_rows(
    lmtest::bptest(cheese.model1) %>% tidy
  ) %>%
  select(method, statistic, p.value) %>%
  kable %>%
  kable_styling(latex_options = "hold_position", full_width = F) %>%
  column_spec(1, width = "5in")
```

method	statistic	p.value
Modified robust Brown-Forsythe Levene-type test based on the absolute deviations from the median	1.551937	0.2231735
studentized Breusch-Pagan test	4.219300	0.2387367

Outliers

The residual plot has a fairly obvious outlier where one value is greater than 20. This can also be seen with a boxplot of the Residual values

```
qplot(y = .resid, geom = "boxplot", data = cheese.model1, ylab = "Residuals")
```



Normality of ϵ_i

The data are generally close to the line on the Normal Probability plot with the exception of a few values on the right tail. A Shapiro-Wilk test confirms that there is not enough evidence to suggest the data are non-normal (p-value = 0.8312)

```
shapiro.test(cheese.model1$residuals) %>%
  tidy %>%
  select(method, statistic, p.value) %>%
  kable %>% kable_styling(latex_options = "hold_position")
```

method	statistic	p.value
Shapiro-Wilk normality test	0.9802122	0.8311695

2

In #1 you estimated both $E(taste) = \beta_0 + \beta_1$ Acetic $+ \beta_2$ H2S $+ \beta_3$ Lactic **and** $E(taste) = \beta_0 + \beta_1$ H2S $+ \beta_2$ Lactic

a

State R_a^2, C_p, AIC, SBC for both of these models

```
# coded my own functions for fun
my.SBC <- function(model) {
  require(dplyr)
  require(broom)
  rss <- anova(model) %>%
    tidy %>%
    filter(term == "Residuals") %>%
    select(sumsq) %>% as.double

  rank <- model$rank
  n <- model$fitted.values %>% length
  # return schwartz' Bayesian Criterion
  n * log(rss) - n * log(n) + log(n) * rank
}

my.AIC <- function(model) {
  require(dplyr)
  require(broom)
  rss <- anova(model) %>%
    tidy %>%
    filter(term == "Residuals") %>%
    select(sumsq) %>% as.double

  rank <- model$rank
  n <- model$fitted.values %>% length
  # return Aikaike Information Criterion
  n * log(rss) - n * log(n) + 2 * rank
}

my.Cp <- function(model, fullmodel) {
  require(dplyr)
  require(broom)

  rss <- anova(model) %>%
    tidy %>%
    filter(term == "Residuals") %>%
    select(sumsq) %>% as.double

  mse.full <- anova(fullmodel) %>%
    tidy %>%
    filter(term == "Residuals") %>%
    select(meansq) %>% as.double

  rank <- model$rank
```

```
  n <- model$fitted.values %>% length
  # return Aikaike Information Criterion
  (rss / mse.full) - (n - 2 * rank)
}

# test that my functions match the output of existing libraries
assertthat::are_equal(
  my.AIC(cheese.model1),
  ols_aic(cheese.model1, method = "SAS")
)
```

```
## [1] TRUE
```

```
assertthat::are_equal(
  my.SBC(cheese.model1),
  ols_sbc(cheese.model1, method = "SAS")
)
```

```
## [1] TRUE
```

```
assertthat::are_equal(
  my.Cp(cheese.model.no_acetic, cheese.model1),
  ols_mallows_cp(cheese.model.no_acetic, cheese.model1)
)
```

```
## [1] TRUE
```

```
# creates the model words using the model
get_terms <- function(model) paste(c("(Intercept)", format(model$term[[3]])), collapse = " + ")

# get adj r-squared
glance(cheese.model1) %>%
  select(adj.r.squared) %>%
  # add the other columns for this
  mutate(
    Cp = my.Cp(cheese.model1, cheese.model1),
    AIC = my.AIC(cheese.model1),
    SBC = my.SBC(cheese.model1),
    model_words = get_terms(cheese.model1)
  ) %>%
  # create and union a row for the no-acetic acid model
  bind_rows(
    glance(cheese.model.no_acetic) %>%
    select(adj.r.squared) %>%
    mutate(
      Cp = my.Cp(cheese.model.no_acetic, cheese.model1),
      AIC = my.AIC(cheese.model.no_acetic),
      SBC = my.SBC(cheese.model.no_acetic),
      model_words = get_terms(cheese.model.no_acetic)
    )
  ) %>%
  # output
  select(model_words, adj.r.squared, Cp, AIC, SBC) %>%
  kable %>% kable_styling(latex_options = "hold_position")
```

model_words	adj.r.squared	Cp	AIC	SBC
(Intercept) + Acetic + H2S + Lactic	0.6115948	4.000000	142.6412	148.2460
(Intercept) + H2S + Lactic	0.6259025	2.005401	140.6475	144.8511

b

Which model is better using the criterion in (a)? Explain.

The Non-acetic model has a higher Adjusted R^2, lower AIC, lower SBC, and lower C_p. This makes the non-acetic model a better fit compared to the full model.

c

Discuss how your results in (b) coincide with the t-test for acetic acid result from 1(d)

The t-test for the Acetic Acid predictor concluded that there was no evidence that Acetic Acid has an associated impact on taste score after H2S and Lactic Acid are already in the model. By selecting the model based on AIC, Adjusted R^2, SBC, and Mallows' C_p, we ruled out the model with Acetic Acid in it meaning The conclusions derived from the model selection statistics confirm the conclusions derived from the t-statistics.

3

Consider a model with hydrogen sulfide and lactic acid within i.

a

Perform a hypothesis test to determine if the model should include an interaction term between the variables. Use $\alpha = 0.05$

```
cheese.model.hypolactic <- lm(taste ~ H2S + Lactic, data = cheese)
cheese.model.hypolactic.x <- lm(taste ~ H2S * Lactic, data = cheese)

cheese.model.hypolactic.x %>%
  tidy %>%
  kable %>% kable_styling(latex_options = "hold_position")
```

term	estimate	std.error	statistic	p.value
(Intercept)	-23.1866518	27.749040	-0.8355839	0.4110023
H2S	3.2355550	4.382070	0.7383623	0.4669074
Lactic	16.7248904	20.479127	0.8166798	0.4215317
H2S:Lactic	0.4880266	2.902326	0.1681502	0.8677663

```
anova(cheese.model.hypolactic.x, cheese.model.hypolactic) %>%
  kable %>% kable_styling(latex_options = "hold_position")
```

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
26	2666.066	NA	NA	NA	NA
27	2668.965	-1	-2.899294	0.0282745	0.8677663

Given that H2S and Lactic are in the model, there is no evidence that their interaction significantly affects taste score (two-tailed t-test. p-value = 0.868). This is confirmed when compared against the additive counterpart (Sum of Squares F-Test. p-value = 0.8678).

b

Perform one hypothesis test to determine if the model should include the interaction term along with the two quadratic terms between the variables. Use $\alpha = 0.05$

```
cheese.model.hypolactic.xtreme <- lm(taste ~ poly(H2S, 2) + poly(Lactic, 2) + H2S:Lactic, data = cheese)
anova(cheese.model.hypolactic.xtreme, cheese.model.hypolactic) %>%
  kable %>% kable_styling(latex_options = "hold_position")
```

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
24	2614.958	NA	NA	NA	NA
27	2668.965	-3	-54.00741	0.1652261	0.9187466

There is no evidence that the squared terms or the interaction term have any significant effect when compared to the reduced additive model (Sum of Squares F-Test. p-value = 0.9187).

c

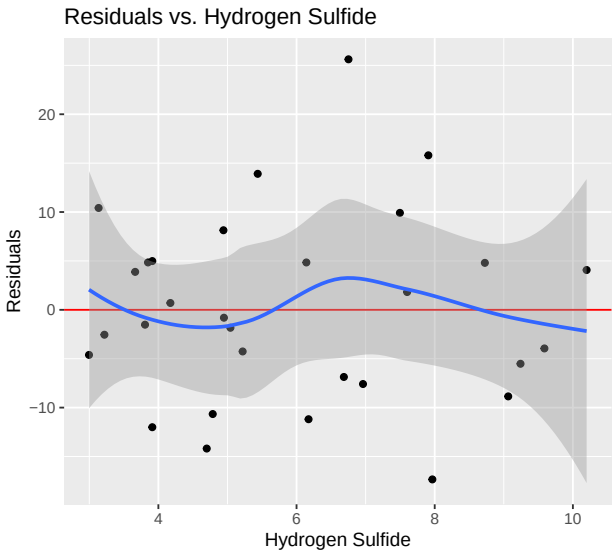
For the model with only the hydrogen sulfide and lactic acid linear terms within it, examine plots of the residuals versus hydrogen sulfide and residual versus lactic acid. Is there evidence of quadratic term(s) needed in the model?

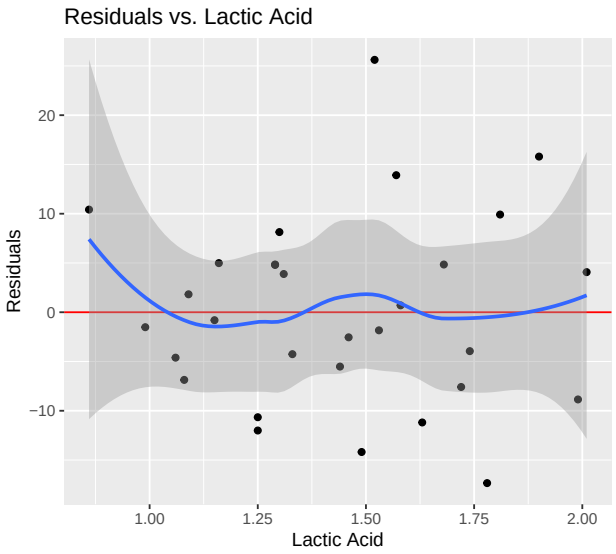
```
plot.resid.base <-
  cheese.model.hypolactic %>%
    ggplot(aes(y = .resid)) +
    ylab("Residuals") +
    geom_hline(yintercept = 0, color = "red")

plot.h2s.resid <-
  plot.resid.base +
  geom_point(aes(x = H2S)) +
  geom_smooth(aes(x = H2S), method = loess) +
  xlab("Hydrogen Sulfide") +
  ggtitle("Residuals vs. Hydrogen Sulfide")

plot.lactic.resid <-
  plot.resid.base +
  geom_point(aes(x = Lactic)) +
  geom_smooth(aes(x = Lactic), method = loess) +
  xlab("Lactic Acid") +
  ggtitle("Residuals vs. Lactic Acid")

grid.arrange(plot.h2s.resid, plot.lactic.resid, ncol = 2)
```






```
shapiro.test(cheese.model.hypolactic$residuals) %>%
  tidy %>%
  select(method, statistic, p.value) %>%
  kable %>% kable_styling(latex_options = "hold_position")
```

method	statistic	p.value
Shapiro-Wilk normality test	0.9794474	0.8106746

$Y = 0$ falls within the confidence bands for the smoother line for both residual plots which indicate that there is not a significant pattern to be detected. Therefore, there is no evidence for adding quadratic terms to the model.

The assumptions described in (1g) still hold true for each residual plot. The linear model is still appropriate, constant variance is met, the same outlier is present, and there is no evidence that the data is non-normal (Shapiro-Wilk Test. p-valu = 0.8107).