BAS 320 - Assignment 7 - Multiple Regression Part 2

Jack Guptill

## Analysis of Beer Aroma

I’m using a multiple regression model to predict the review that a beer might get for its aroma. The model to predict this aroma review will use the number of spices, hops, bitterness, sourness, body of the beer, and the overall review of the taste as predictors.

I’m making this model primarily out of curiosity to see how plausible it is to predict something so subjective such as the aroma. I am also curious to the interactions of the predictors that will be taking place in the model.

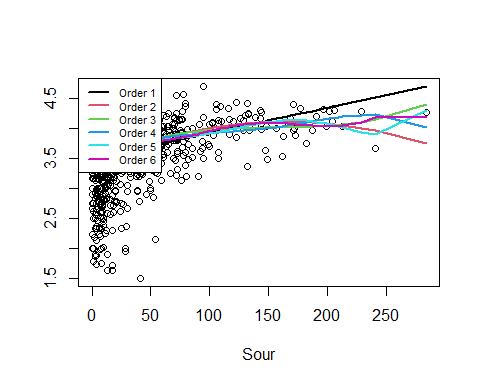
The data I’m using comes from kaggle (<https://www.kaggle.com/datasets/ruthgn/beer-profile-and-ratings-data-set/data?select=beer_profile_and_ratings.csv>) and contains 3197 rows with 25 variables. The specific subset that I will be using to build the model has 1000 rows and 7 variables.

## Investigation of the relationship between the review of the Aroma and Sourness

After running a simple linear regression, looking at the relationship between the aroma review and the sourness, it seems to me that a polynomial model would improve the R squared value due to the non-linear relationship trend that has developed. After analyzing the various orders to which the model could use, it seems that the third order would be the best to increase the r^2 value from roughly 9% to a little less than 11%. While not a large increase to the R^2 value, it is nonetheless increasing it. Order three can be seen in the visualization as the green line with the black line as the order zero polynomial model.

#Finding the most non-linear predictor variable  
M = lm(review\_aroma~Sour, data=BEER2); choose\_order(M)

#order three is the best for the model



## order R2adj AICc  
## 1 1 0.09052599 1418.487  
## 2 2 0.10599838 1404.338  
## 3 3 0.10912681 1403.845  
## 4 4 0.11058377 1405.224  
## 5 5 0.11329082 1405.194  
## 6 6 0.11384784 1407.588

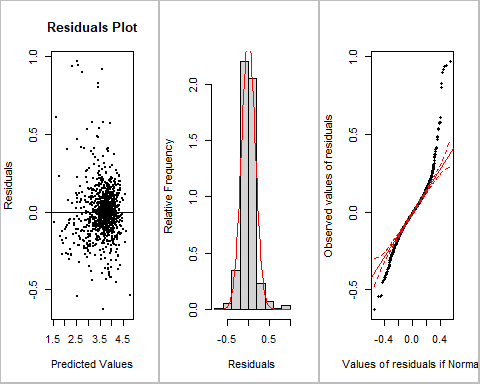
## Multiple regression model and checking of assumptions

After examining the output in testing the assumptions for a multiple regression model, the full model passes the tests for linearity when it came to the hops, taste review, bitterness, and sourness predictors. The model failed linearity for the spices and body predictor variables. It is important to note that linearity for the overall model failed due to a lack of duplicate values. If the model was more robust in overall sample size and a different ransom sample were to be taken, this condition might have passed. In the output it also states that we have failed both in equal spread and normality of the residuals. Due to our sample size being at least 25, we can further inspect the output graphs to override these decisions. Starting with the far left residual plot, it doesn’t seem like we are violating linearity which confirms most of the test output. It also seems that the equal spread condition isn’t grossly violated in any way. Data points seem to be roughly equal by eye. For the QQ plot it seems that we have minor peeling on bottom edge. We do however have skew at the top end of the graph. Based on this and my own opinion, I think the model does exhibit a reasonable reflection of reality due to the tests not being grossly violated in anyway.

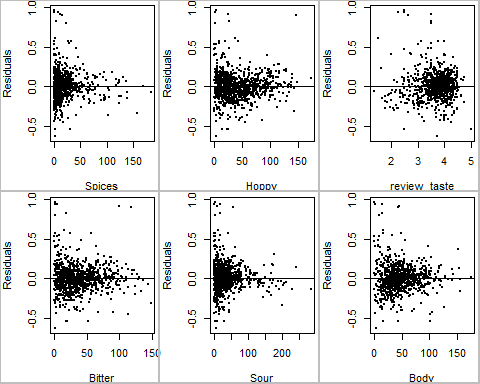
M =lm(review\_aroma ~ Spices+Hoppy+review\_taste+Bitter+Sour+Body,data=BEER2)  
summary(M)

##   
## Call:  
## lm(formula = review\_aroma ~ Spices + Hoppy + review\_taste + Bitter +   
## Sour + Body, data = BEER2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.62622 -0.08335 -0.00405 0.07331 0.96643   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.3498218 0.0389347 8.985 < 2e-16 \*\*\*  
## Spices 0.0019761 0.0002351 8.405 < 2e-16 \*\*\*  
## Hoppy -0.0013627 0.0002837 -4.804 1.80e-06 \*\*\*  
## review\_taste 0.8723513 0.0118130 73.846 < 2e-16 \*\*\*  
## Bitter 0.0021840 0.0003933 5.553 3.60e-08 \*\*\*  
## Sour 0.0011938 0.0001571 7.597 6.99e-14 \*\*\*  
## Body -0.0007872 0.0002902 -2.712 0.0068 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.1657 on 993 degrees of freedom  
## Multiple R-squared: 0.8965, Adjusted R-squared: 0.8959   
## F-statistic: 1434 on 6 and 993 DF, p-value: < 2.2e-16

check\_regression(M,extra=TRUE)



##   
## Tests of Assumptions: ( sample size n = 1000 ):  
## Linearity  
## p-value for Spices : 0   
## p-value for Hoppy : 0.9774   
## p-value for review\_taste : 1   
## p-value for Bitter : 0.9768   
## p-value for Sour : 0.8317   
## p-value for Body : 0.0033   
## p-value for overall model : NA (not enough duplicate rows)  
## Equal Spread: p-value is 0   
## Normality: p-value is 0   
##   
## Advice: if n<25 then all tests must be passed.  
## If n >= 25 and test is failed, refer to diagnostic plot to see if violation is severe  
## or is small enough to be ignored.  
##   
## Press [enter] to continue to Predictor vs. Residuals plots or q (then Return) to quit ( 6 plots to show )

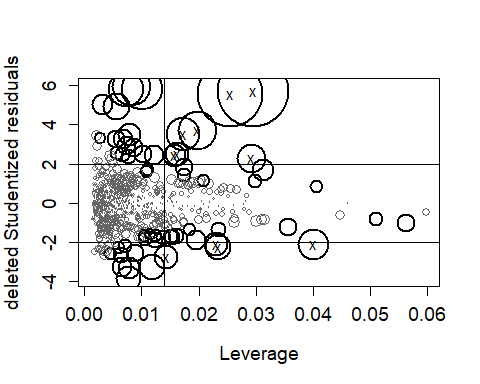


## Identification of Influential Points

For this model, we have 10 points of influence that contain both high leverage and a large deleted studentized residual. The specific row numbers can be found below. These data points are negatively impacting the overall model due to their overall influence. For this analysis I want to focus on the 387th beer. This beer is strange in the fact that has a low number of hops, spices, overall taste review, sourness, and body, but it has a high bitterness and overall aroma score making it a point of influence. Please see the data frame below that outlines the percentile to which the predictors are for this beer.

influence\_plot(M)

## $Leverage  
## [1] 30 35 153 387 411 443 644 838 924 960 981



## Spices Hoppy review\_taste Bitter Sour Body review\_aroma  
## 1 0.48 0.24 0.07 0.85 0.56 0.95 0.18  
## 2 0.15 0.92 0.75 0.41 0.19 0.19 0.91  
## 3 0.98 0.19 0.94 0.48 0.41 0.99 0.67  
## 4 0.05 0.10 0.13 0.79 0.03 0.10 0.52  
## 5 0.58 0.36 0.26 0.97 0.19 0.62 0.99  
## 6 0.15 0.56 0.00 0.23 0.28 0.10 0.03  
## 7 0.68 0.29 0.95 0.61 0.09 1.00 0.98  
## 8 0.65 1.00 0.05 0.99 0.85 0.56 0.43  
## 9 0.67 0.22 0.96 0.13 0.99 0.62 0.81  
## 10 1.00 0.12 0.38 0.21 0.22 0.44 0.33  
## 11 0.69 0.72 0.91 0.98 0.41 0.96 0.52

## Investigation of an interaction between Hoppiness and the review of the taste

I want to investigate the interaction of various levels of hoppiness for a beer and how the review taste impacts the overall aroma of the beer.

**The Overall Regression Equation:**

Review Aroma = .410 + .881*review\_taste - .006*Hoppy + .001*review\_taste*Hoppy

**Beers with the 25th percentile of Hoppy (19):**

review aroma = .296 + .9\*review\_taste

**Beers with the 75th percentile of Hoppy (59):**

review aroma = .056 + .94\*review\_taste

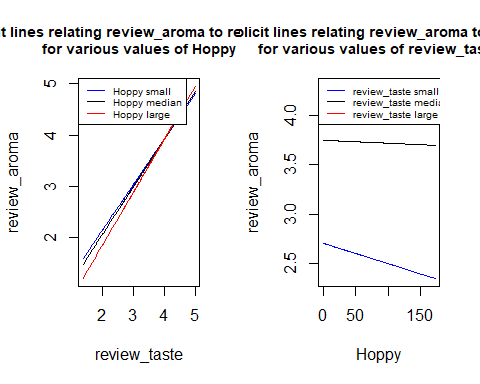
As the review of the taste increases we see an increase in the aroma review. We see this as an overall positive relationship. The rate at which it increases however, is changed with how hoppy the beer is. With a hooppier beer, the rate is much stronger than a less hoppy beer. A “medium” beer in terms of hoppiness sits right in between the two. Knowing this, we can assume that there is an interaction between how hoppy the beer is and the aroma that a beer will get from the review of the taste.

For beers that have a relatively low amount of Hoppiness there seems to be a good amount of variation in the review of the aroma. As we get towards the median value of hoppiness (33) the variation widens and it looks to be getting wider as the hoppiness of a beer increases.

M <- lm(review\_aroma ~ review\_taste\*Hoppy, data=BEER2)  
summary(M)

##   
## Call:  
## lm(formula = review\_aroma ~ review\_taste \* Hoppy, data = BEER2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.70010 -0.08729 0.00016 0.07920 1.24565   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.4103488 0.0603315 6.802 1.78e-11 \*\*\*  
## review\_taste 0.8775701 0.0160391 54.714 < 2e-16 \*\*\*  
## Hoppy -0.0059622 0.0016209 -3.678 0.000247 \*\*\*  
## review\_taste:Hoppy 0.0014893 0.0004212 3.535 0.000426 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.1745 on 996 degrees of freedom  
## Multiple R-squared: 0.8849, Adjusted R-squared: 0.8846   
## F-statistic: 2552 on 3 and 996 DF, p-value: < 2.2e-16

visualize\_model(M, cex=0.6,pos="topleft",many=TRUE)



##   
## Interaction term has p-value 0.0004259