STATS 500 HOMEWORK 5 YUAN YIN

Problem 1

After setting up our model, we want to check as follows assumptions.

- First we want to check whether the constant variance assumption holds. We plot the studentized residuals vs fitted values as follows:

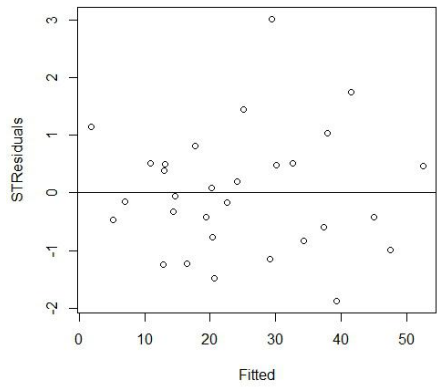

Fig 1 studentized residuals vs fitted model

Seeing that there is no obvious non - linearity or heteroscedasticty relations between the two variables. We check residuals against each variables “Acetic”, “H2S” and “Lactic” again to see if they have some non - linearity relations. The plots are as follows:

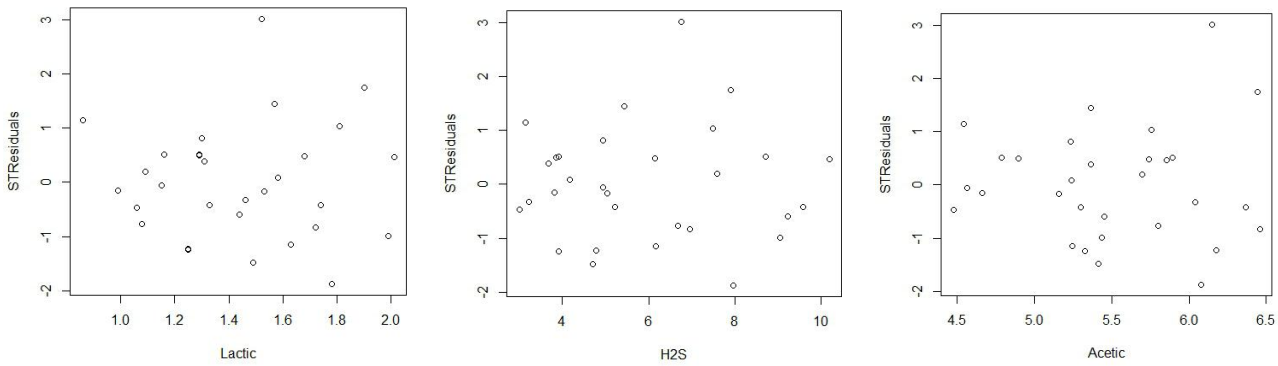

Fig 2 studentized residuals vs predictors individually

Still, there aren’t any predictors have an obvious non - linearity relations with residuals. In this way, we can assume that the variance is constant.

- Next we want to check if the normality assumption holds. Find that QQ - plot shows as follows:

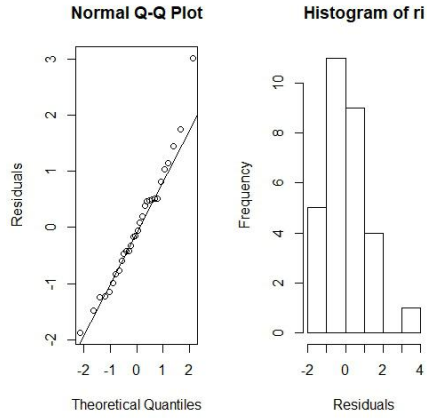

Fig 3 QQ - plot of studentized residuals

We can tell from the left plot that almost all the points are on the quantiles line of normal ditribution, except the last few points. Also we can confirm this from the right plot, which shows there is a small peak after 2 - 3 of residuals. However, we can still think normality assumption holds. This is because the last 2 - 3 points is a small amount against the whole data, and as the probability of the extreme points is low, they can be too random to happenly be on the line above. What's more, the data we have is only 30, which is also a very small amount. In conclusion, we can treat this model as satisfying normality assumption.

- The third step is to find large leverage points. Use half - norm plot as follows:

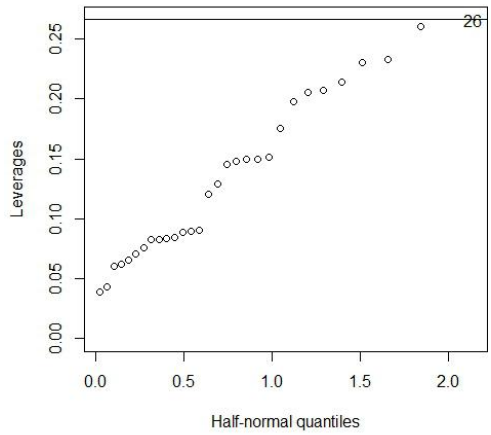

Fig 4 half - norm plot of h (leverage)

According to the definition of high leverage, h should be larger than: 2*(p+1)/30 = 0.266667 (where p is 3), and it's obvious that the largest h of the points above is no larger than 0.3. So there is no high leverage points for our model.

- To find all the outliers, we first choose the largest residual to compute it's p - value.

As it is the 15th data in "cheddar" and the value of its residual is 3.01547, assume null hypothsis is "the data with the largest residual isn't an outlier", and we compute the p - value is 0.00581769. Comparing it to the adjusted alpha which is $\alpha/n = 0.05/30 = 0.001666667$, we found the p - value is larger than adjusted alpha, which means we fail to reject null hypothsis. That is to say, we are not confident to say the 15th data is an outlier, let alone the other points. In conclusion, we aren't able to say there are outliers for our model.

- Cook's distance is used to find all the influential points, the plot is as follows:

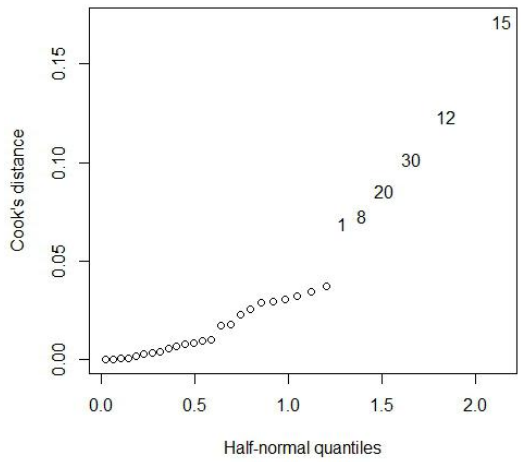

Fig 5 cook's distance of all data

From the plot above, we found that there are six points with high cook's distance, to check their influential degree to the model, we verify them by plot the change of each coefficient:

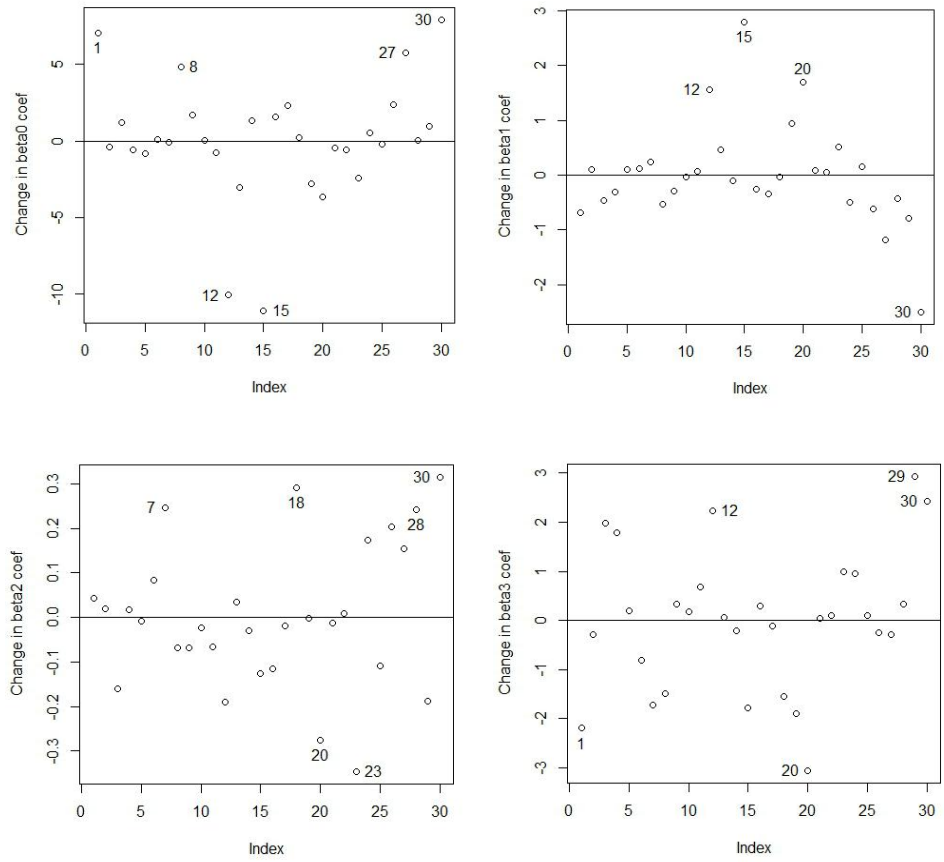

Fig 6 changes of each coeffcient

Identify the points with large changes, we can see that except the 8th data, all the other five influential points we found before have shown many times in the plots above. For all the points showed above, we compare them with the origin coeffcient:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -28.8768    19.7354  -1.463  0.15540
Acetic        0.3277     4.4598   0.073  0.94198
H2S           3.9118     1.2484   3.133  0.00425 **
Lactic       19.6705     8.6291   2.280  0.03108 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Fig 7 summary of fitted model

We can find that beta2 is the coeffcient of H2S, which is 3.9118, besides, the plot above of changes in beta2 are no more than 0.3, which means their changes are no more than 10% of beta2, considering those points also don't have a large effect on the other three coeffcients, we can exclude them as influential points. It's the same reason with data in beta3 plot (the changes are no more than 20%) and data in beta0 plot (the changes are no more than 20%). But in beta1 plot, we can see the four data have more than 100% changes of beta0, so they should be treated as influential points. In conclusion, the influential points we found are 12,15,20,30.

To confirm this, we compute the updated models without one of these points.

(Intercept)	Acetic	H2S	Lactic
-36.815015	2.832364	3.596473	17.248022

Fig 8 updated coef after delete 30th data

(Intercept)	Acetic	H2S	Lactic
-25.242061	-1.364014	4.187601	22.732932

Fig 9 updated coef after delete 20th data

```
(Intercept)      Acetic          H2S      Lactic
 -17.756822   -2.470399    4.038725   21.458484
```

Fig 10 updated coef after delete 15th data

```
(Intercept)      Acetic          H2S      Lactic
 -18.792267   -1.236676    4.101001   17.437080
```

Fig 11 updated coef after delete 12th data

We can see that the coeffcient of Acetic changes a lot when delete these data, in this way, 12, 15, 20, 30 are influential points.

- At last, we check the structure of the relationship between the predictors and response

We plot partial regression plots and partial residual plots for each predictor as follows:

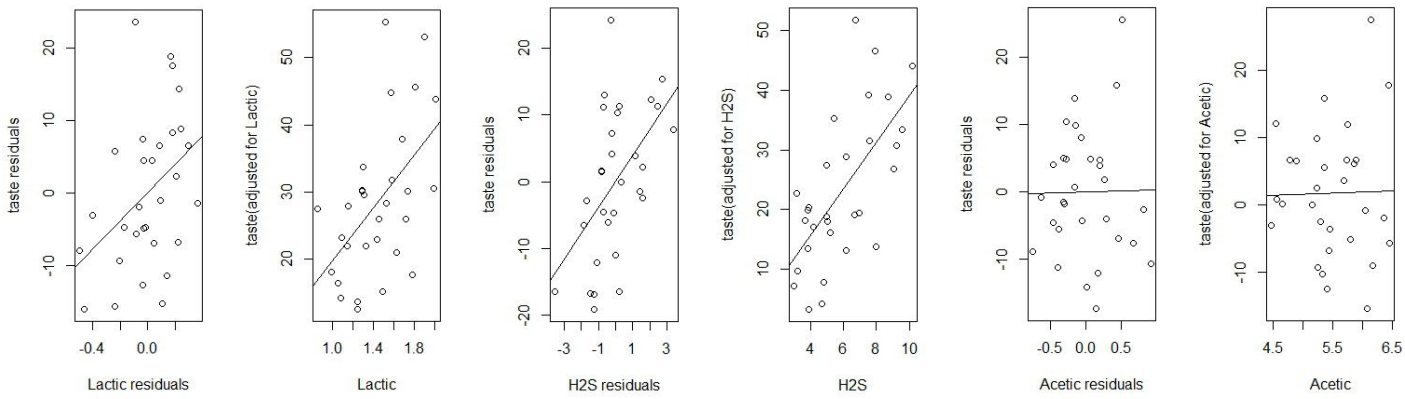

Fig 12 partial regression plots & partial residual plots

From the figure above, we find that for each predictor, the plots look basically fine, there is no significant nonlinearity or outliers or influential points that affect the structure of our model. However, we find that the slope for “Acetic” are almost 0, we wonder if this term is necessary for setting up our model. We set up a new model without “Acetic” again and here is the result:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -27.592      8.982  -3.072  0.00481 **
H2S            3.946      1.136   3.475  0.00174 **
Lactic        19.887      7.959   2.499  0.01885 *
```

Fig 13 summary of new model

We can find that in the new model without “Acetic”, every p-value fits very well. So in conclusion, if we want to improve our model, we can take more analysis on the need of exitence of “Acetic”. It also explains why there are so many influential points in the original model, as they all mainly affect the coeffcient of beta1 a lot.

Appendix

```
>> library(faraway) >> data(cheddar) >> temp = lm(taste ~ Acetic + H2S + Lactic, data = cheddar)
>> summary(temp) >> ri = rstudent(temp) >> plot(temp$fitted, ri, xlab = "Fitted", ylab = "STResiduals")
abline(h = 0) >> plot(cheddar$Acetic, ri, xlab = "Acetic", ylab = "STResiduals")
>> plot(cheddar$H2S, ri, xlab = "H2S", ylab = "STResiduals") >> plot(cheddar$Lactic, ri, xlab = "Lactic",
ylab = "STResiduals") >> par(mfrow = c(1,2)) >> qqnorm(ri, ylab = "Residuals") >> qqline(ri)
>> hist(ri, xlab = "Residuals") >> par(mfrow = c(1,1))
>> halfnorm(lm.influence(temp)$hat, nlab = 1, ylab = "Leverages") >> abline(h = 2*(3+1)/30)
>> max(abs(ri)) >> which(ri == max(abs(ri))) >> 2*(1 - pt(max(abs(ri)), df = 30-4-1)) >> 0.05/30
>> cook = cooks.distance(temp) >> plot(dfbeta(temp)[,1], ylab = "Change in beta0 coef")
>> abline(h=0) >> identify(dfbeta(temp)[,1]) >> halfnorm(cook, nlab = 6, ylab = "Cook's distance")
```

```
>> cheddar[c(1,8,20,30,12,15),] >> plot(dfbeta(temp)[,2], ylab = "Change in beta1 coef") >> abline(h=0)
>> identify(dfbeta(temp)[,2]) >> plot(dfbeta(temp)[,3], ylab = "Change in beta2 coef") >> abline(h=0)
>> identify(dfbeta(temp)[,3]) >> plot(dfbeta(temp)[,4], ylab = "Change in beta3 coef") >> abline(h=0)
>> identify(dfbeta(temp)[,4]) >> summary(temp) >> cheddar1 = cheddar[-c(30),] >> y = lm(taste ~ Acetic
+ H2S + Lactic, data = cheddar1) >> coef(y) >> cheddar1 = cheddar[-c(20),]
>> y = lm(taste ~ Acetic + H2S + Lactic, data = cheddar1) >> coef(y) >> cheddar1 = cheddar[-c(15),]
>> y = lm(taste ~ Acetic + H2S + Lactic, data = cheddar1) >> coef(y) >> cheddar1 = cheddar[-c(12),]
>> y = lm(taste ~ Acetic + H2S + Lactic, data = cheddar1) >> coef(y) >> par(mfrow = c(1,2))
>> delta = residuals(lm(taste ~ H2S + Lactic, data = cheddar))
>> gamma = residuals(lm(Acetic ~ H2S + Lactic, data = cheddar))
>> plot(gamma, delta, xlab = "Acetic residuals", ylab = "taste residuals") >> tempi = lm(delta ~ gamma)
>> abline(reg = tempi) >> plot(cheddar$Acetic, temp$residuals + coef(temp)['Acetic']*cheddar$Acetic,
xlab = "Acetic", ylab = "taste(adjusted for Acetic)") >> abline(a = 0, b = coef(temp)['Acetic'])
>> delta = residuals(lm(taste ~ Acetic + Lactic, data = cheddar))
>> gamma = residuals(lm(H2S ~ Acetic + Lactic, data = cheddar))
>> plot(gamma, delta, xlab = "H2S residuals", ylab = "taste residuals") >> tempi = lm(delta ~ gamma)
>> abline(reg = tempi) >> plot(cheddar$H2S, temp$residuals + coef(temp)['H2S']*cheddar$H2S, xlab = "H2S",
ylab = "taste(adjusted for H2S)") >> abline(a = 0, b = coef(temp)['H2S'])
>> delta = residuals(lm(taste ~ Acetic + H2S, data = cheddar))
>> gamma = residuals(lm(Lactic ~ Acetic + H2S, data = cheddar))
>> plot(gamma, delta, xlab = "Lactic residuals", ylab = "taste residuals") >> tempi = lm(delta ~ gamma)
>> abline(reg = tempi) >> plot(cheddar$Lactic, temp$residuals + coef(temp)['Lactic']*cheddar$Lactic,
xlab = "Lactic", ylab = "taste(adjusted for Lactic)") >> abline(a = 0, b = coef(temp)['Lactic'])
>> result = lm(taste ~ H2S + Lactic, data = cheddar)
>> summary(result)
```