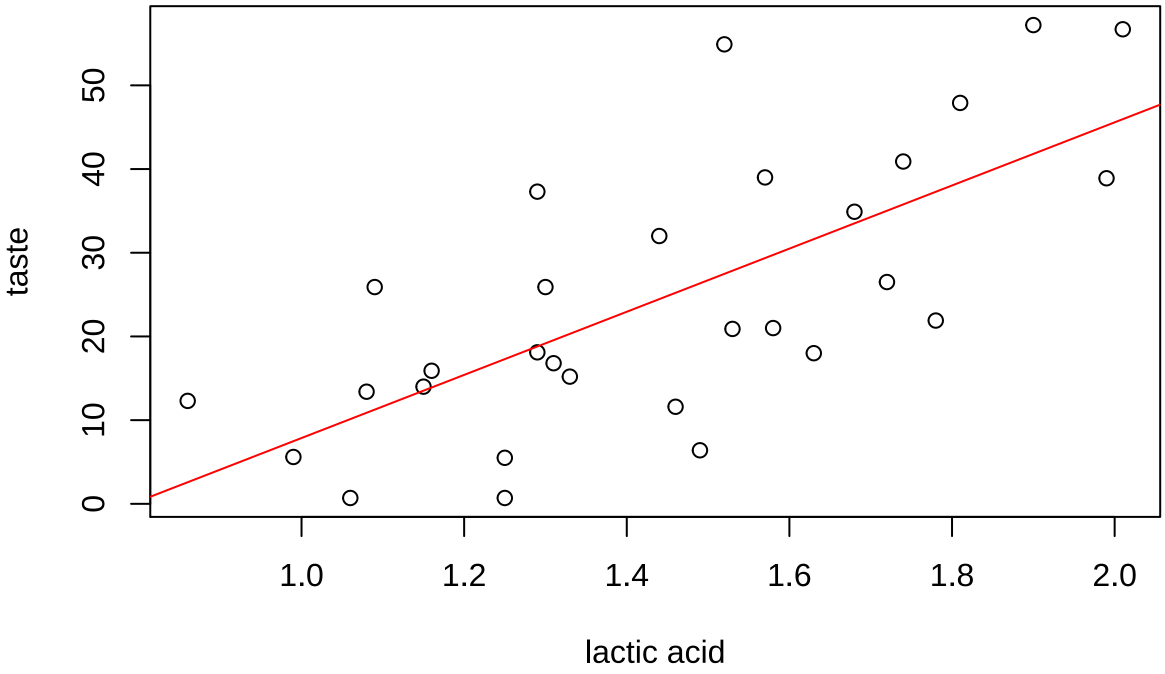
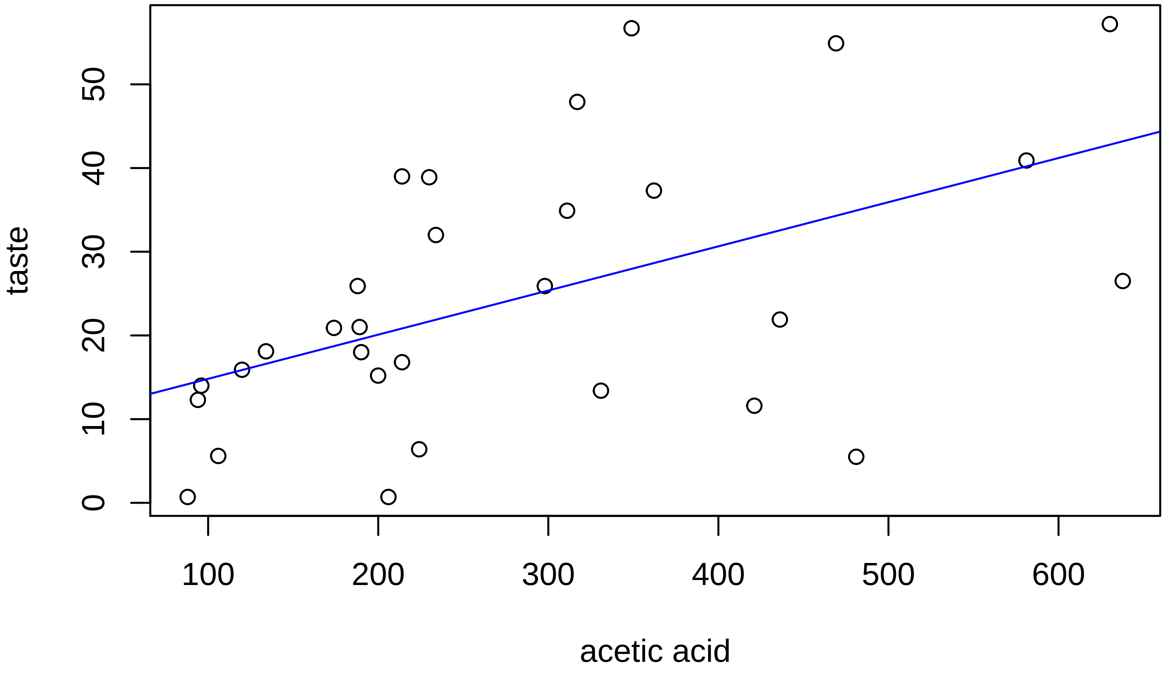
exam1_2019_Cutler_PDF

1. Why is violation of normality in the outcome variable not a problem:

The normality assumption for linear regression applies to the errors, not to the outcome variable by itself, which in this case is the taste score.

2. Simple Linear Regression and LOESS:





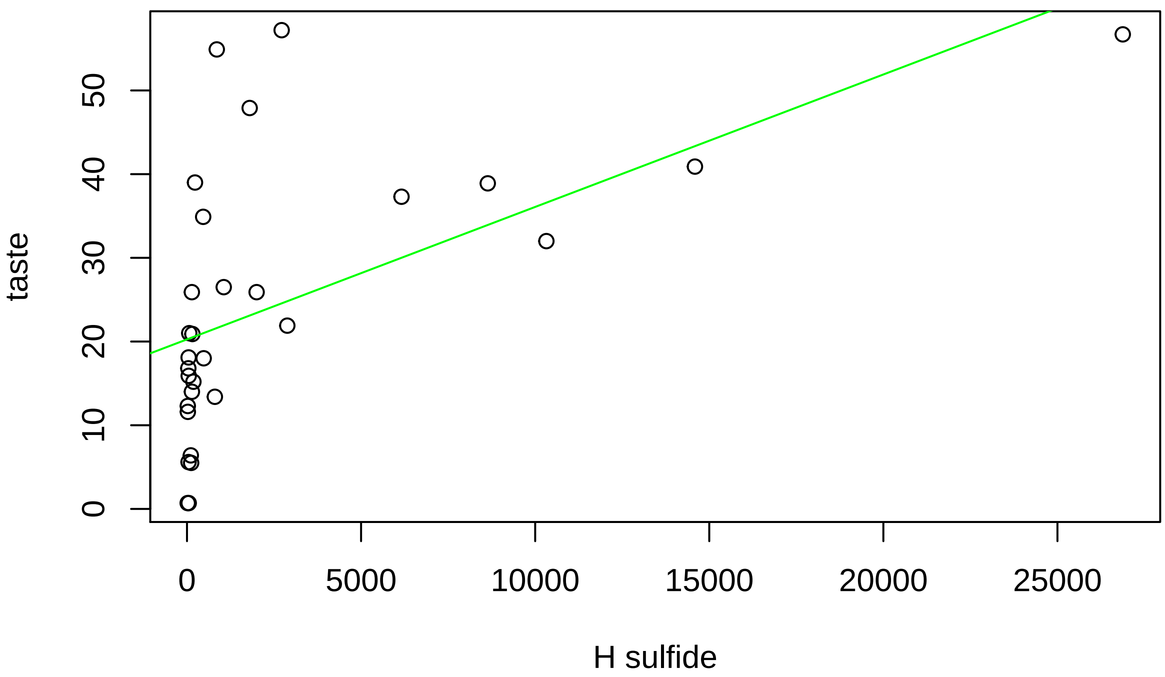


Table 1: SLR model for lactic acid ('Lacid')

	Dependent variable:
	taste
Lacid	37.720*** (7.186)
Constant	-29.859^{***} (10.582)
Observations	30
R^2	0.496
Adjusted R^2	0.478
Residual Std. Error	11.745 (df = 28)
F Statistic	27.550*** (df = 1; 28)
Note:	*p<0.1; **p<0.05; ***p<0.01

2.5 % 97.5 % (Intercept) -51.53573 -8.181935 Lacid 22.99928 52.440613 2.5 % 97.5 % (Intercept) -1.52478617 20.60252313 Aacid 0.01860862 0.08691771 2.5 % 97.5 % (Intercept) 14.57774505 25.933841789 Hsulf 0.00067731 0.002487749

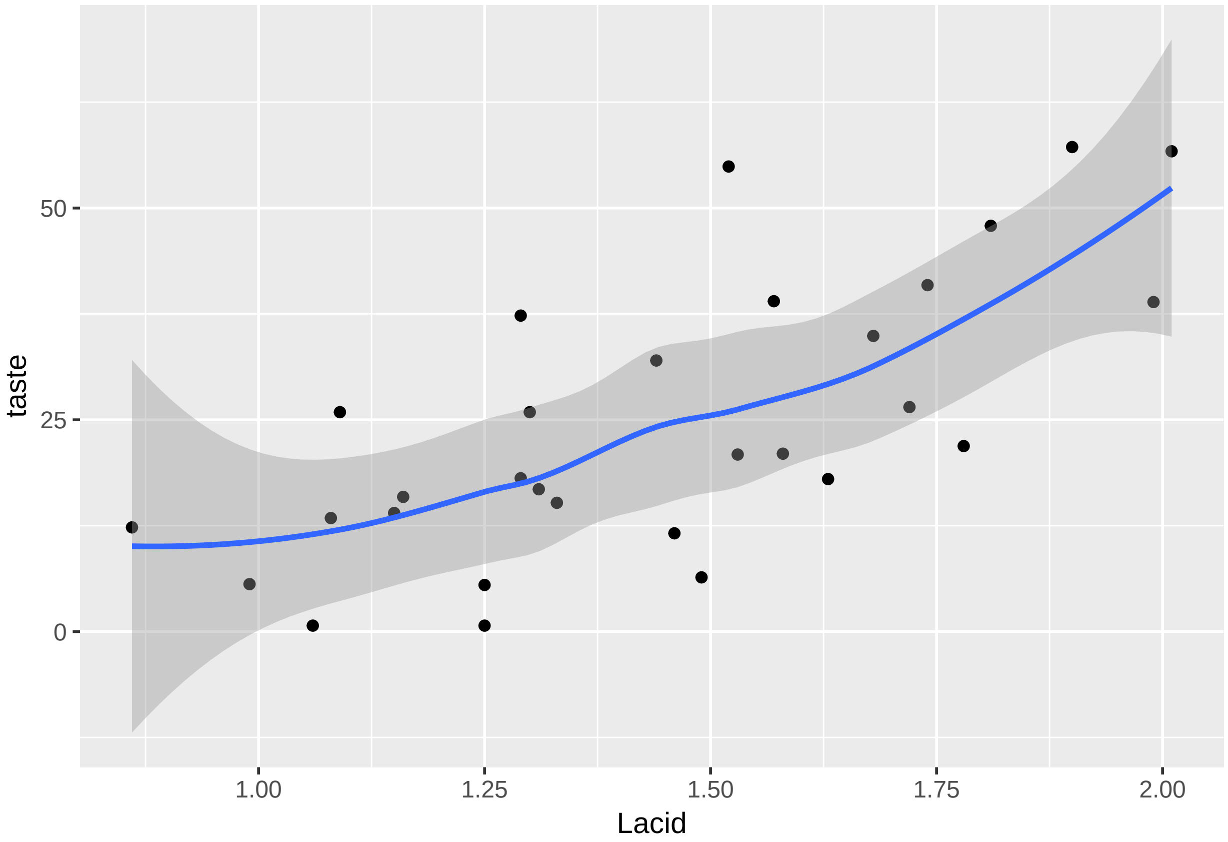
```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

Table 2: SLR model for acetic acid ('Aacid')

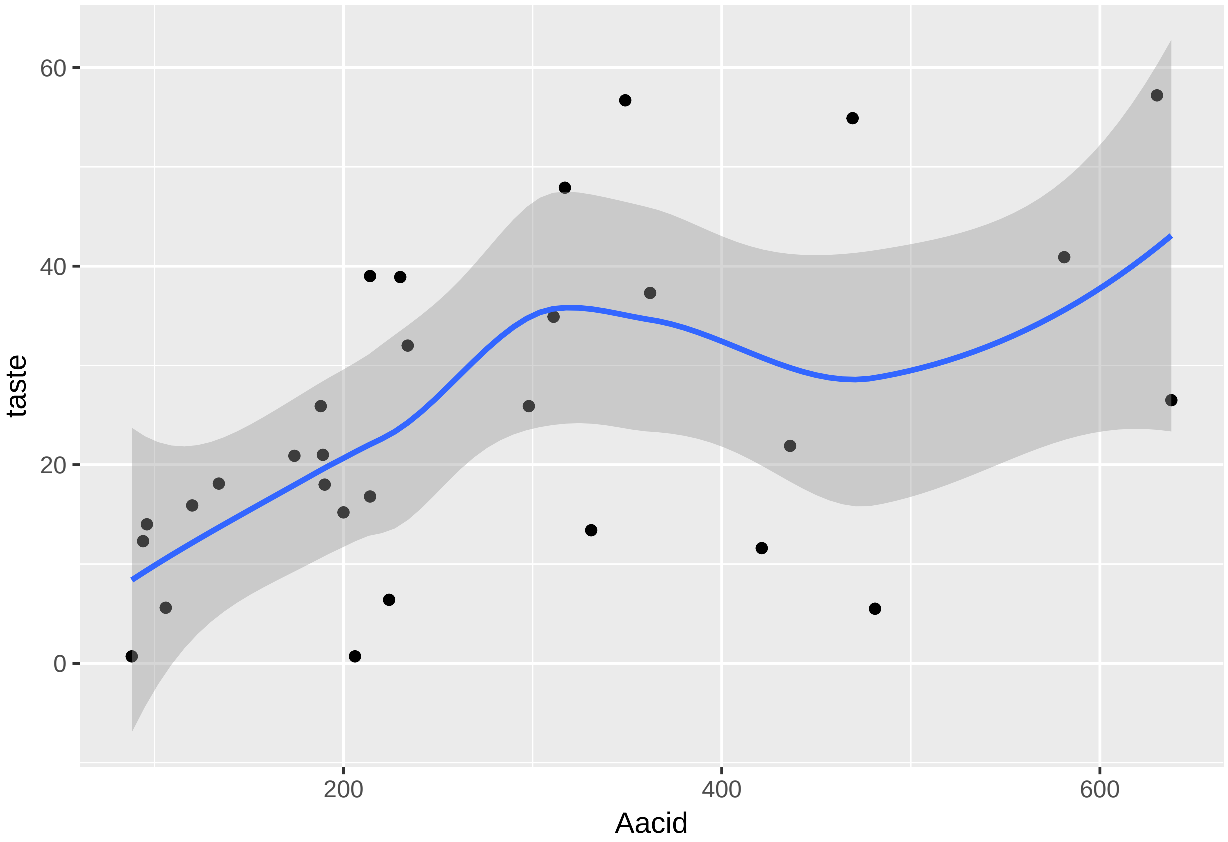
	Dependent variable:
	taste
Aacid	0.053^{***}
	(0.017)
Constant	9.539^{*}
	(5.401)
Observations	30
R^2	0.263
Adjusted R^2	0.237
Residual Std. Error	14.198 (df = 28)
F Statistic	10.014^{***} (df = 1; 28)
Note:	$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

Table 3: SLR model for hydrogen sulfide ('Hsulf')

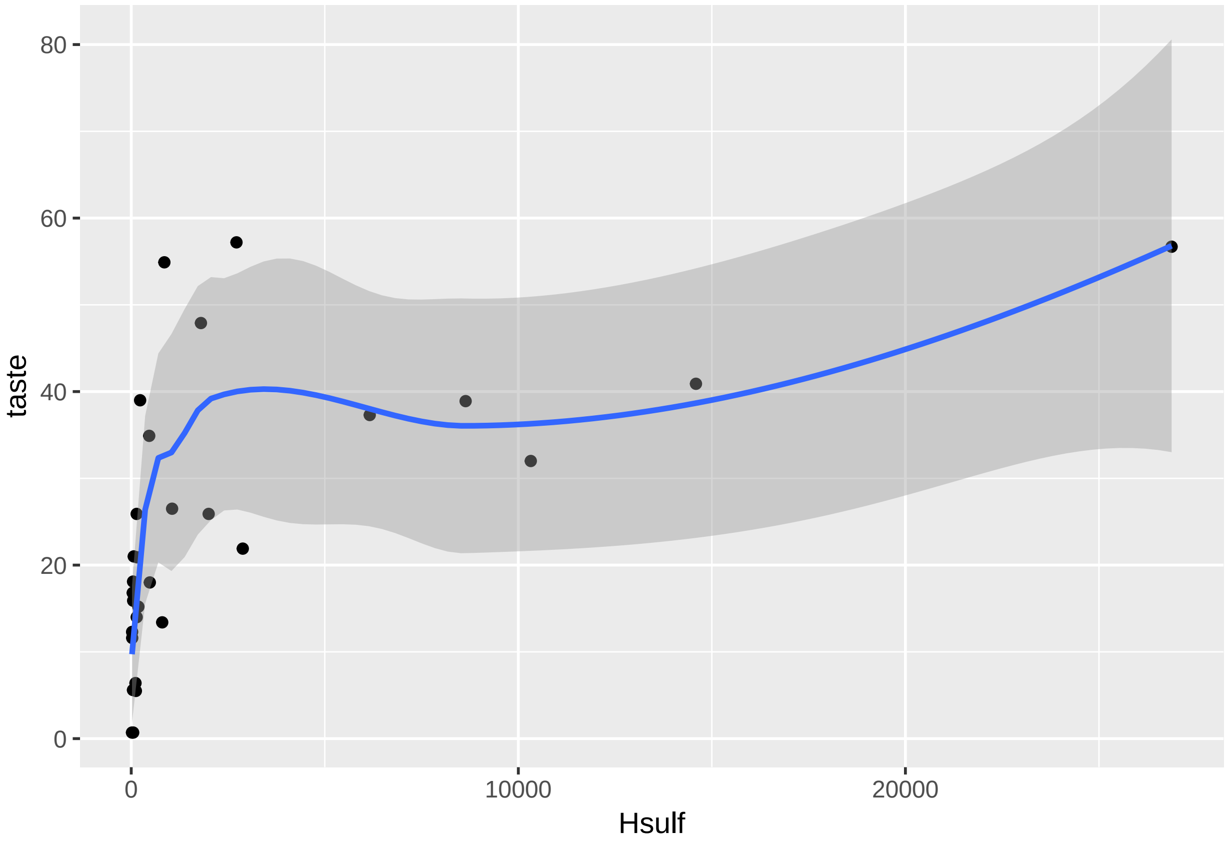
	Dependent variable:
	taste
Hsulf	0.002^{***}
	(0.0004)
Constant	20.256^{***}
	(2.772)
Observations	30
R^2	0.314
Adjusted R^2	0.290
Residual Std. Error	13.701 (df = 28)
F Statistic	12.824^{***} (df = 1; 28)
Note:	$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$



```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



3. Equations of models:

Association between taset and acetic acid:

Y ~ 9.54 taste score points + .05 taste score points per mg/L of acetic acid + error

Taste and lactic acid:

Y ~ -29.86 taste score points + 37.72 taste score points per mg/L of lactic acid + error

Taste and H sulfide:

Y ~ 20.26 taste score points + .0016 taste score points per mg/L of hydrogen sulfide + error

4. Sentences quantifying the associations:

Taste and acetic acid:

Taste score increases .05 points (give or take some error) with each added mg/L of acetic acid, from a baseline of 9.54 points at 0 mg/L of acetic acid.

Taste and lactic acid:

Taste score increases 37.72 points (give or take some error) with each added mg/L of lactic acid, from a baseline of -29.86 points at 0 mg/L of lactic acid.

Taste and H sulfide:

Taste score increases .0016 points with each added mg/L of hydrogen sulfide, from a basline of 20.26 points at 0 mg/L of hydrogen sulfide.

5. Because unless we use CIs, our point estimate of the population parameter in question will always be guaranteed to be wrong (not equal to the true parameter), at least when dealing with continuous data. This is because the probability of a random variable taking on a specific value is essentially zero, due to the laws of probability. A CI will at least give us a good chance of including the true parameter within the sweep of the CI.

6. 'Sample statistic' refers to a function of a sample (e.g. X-bar is defined as the sum of the Xi's divided by the sample size), and is meant to approximate a true population parameter. 'Parameter estimate' refers to a predicted change in the value of the dependent variable in response to a 1-unit change in a predictor variable. Because methods for predicting (choice of model) the response variable differ, the parameter estimates calculated from those models can differ. Sample statistics can relate to parameter estimates in that sample statistics do ideally approximate population parameters, but this is based on a different usage of the term 'parameter estimate'.

7. A least squares estimate is an estimate that minimizes the amount of squared error, or distance between Y-hat (predicted Y values of the response variable) and the observed Y (dependent/response variable) values.

8. Comparing linear and quadratic models for lactic acid:

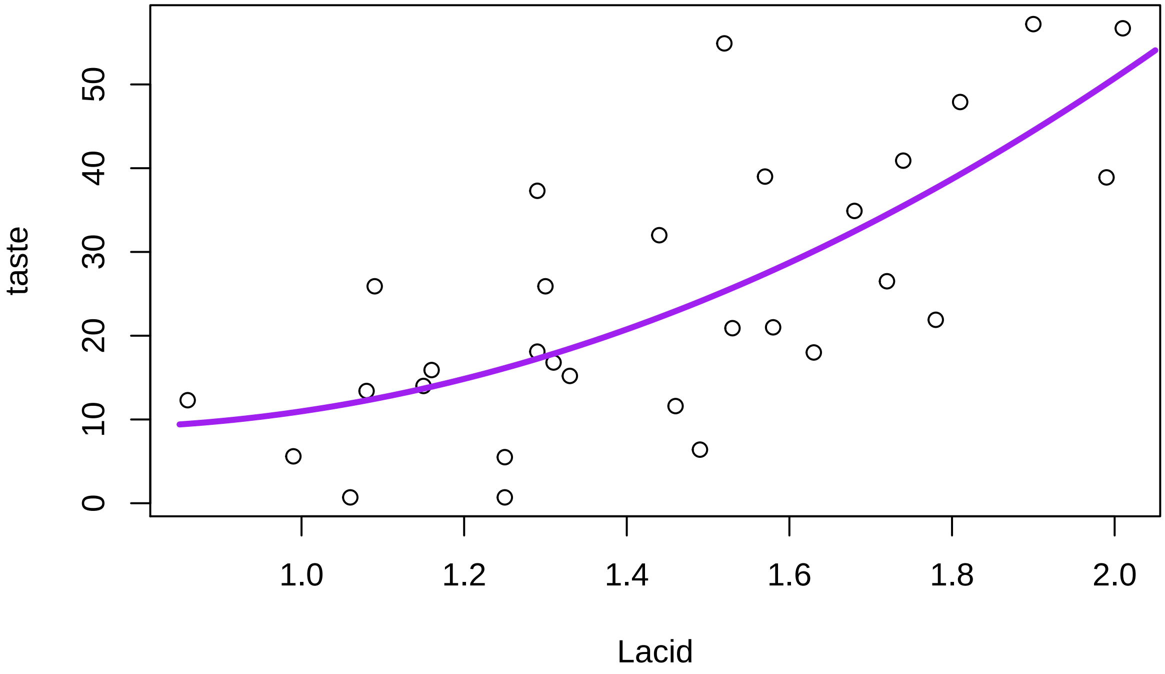


Table 4: Quadratic model for lactic acid

	Dependent variable:
	taste
Lacid	-36.761 (65.292)
Lacid2	25.511 (22.229)
Constant	22.225 (46.588)
Observations	30
R^2	0.519
Adjusted R^2	0.484
Residual Std. Error	11.679 (df = 27)
F Statistic	14.589^{***} (df = 2; 27)
Note:	*p<0.1; **p<0.05; ***p<0.01

MSE and R^2 for lactic acid linear model:

MSE: 128.7496232

R^2: .4959

MSE and R^2 for lactic acid quadratic model:

MSE: 122.7613418

R^2: .5194

Table 5: Quadratic model for acetic acid

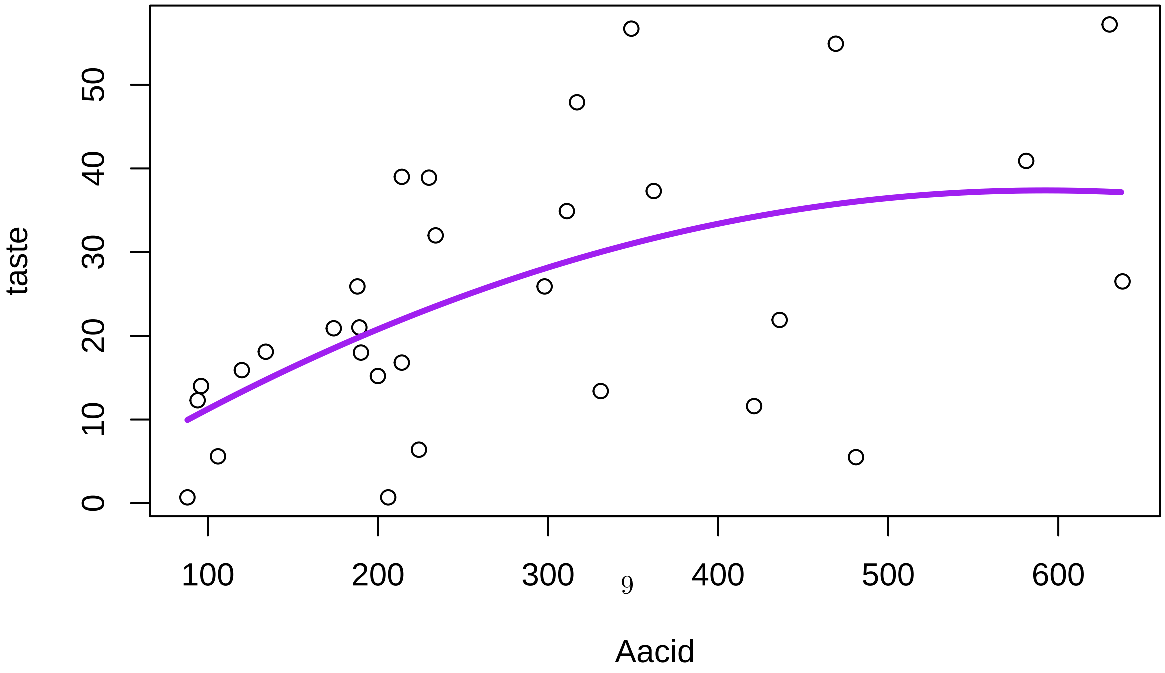
	Dependent variable:
	taste
Aacid	0.128*
	(0.073)
Aacid2	−0.0001
	(0.0001)
Constant	−0.456
	(10.942)
Observations	30
R^2	0.292
Adjusted R^2	0.240
Residual Std. Error	14.172 (df = 27)
F Statistic	5.576*** (df = 2; 27)
Note:	*p<0.1; **p<0.05; ***p<0.01

The quadratic model looks better than the linear, based on the higher Rˆ2 value and the lower MSE. That means it is more explanatory, and it has smaller error.

9. Why must the quadratic model also include an unsquared term?

In the hierarchical approach, if one includes a 2nd power term in their model, then a first power term must also be included. If the model includes 3rd power term, then all powers below that are also included, and so on with higher powered terms. The lower power terms must be included, because they provide more basic information about the shape of the response function, meaning the shape of the regression curve fitted to the data. The higher power terms (in the quadratic case, the squared term) only provide refinements to the shape of the response function.

10. Comparing linear and quadratic models for acetic acid:



MSE and R^2 for acetic acid linear model:

MSE: 188.1431155

R^2: .2634

MSE and R^2 for acetic acid quadratic model:

MSE: 180.7651966

R^2: .2923

The MSE is lower and R^2 higher for the quadratic model. On this basis I prefer the quadratic. Samse reasoning as with lactic acid.

11. Define the type III SS for the acetic acid quadratic model:

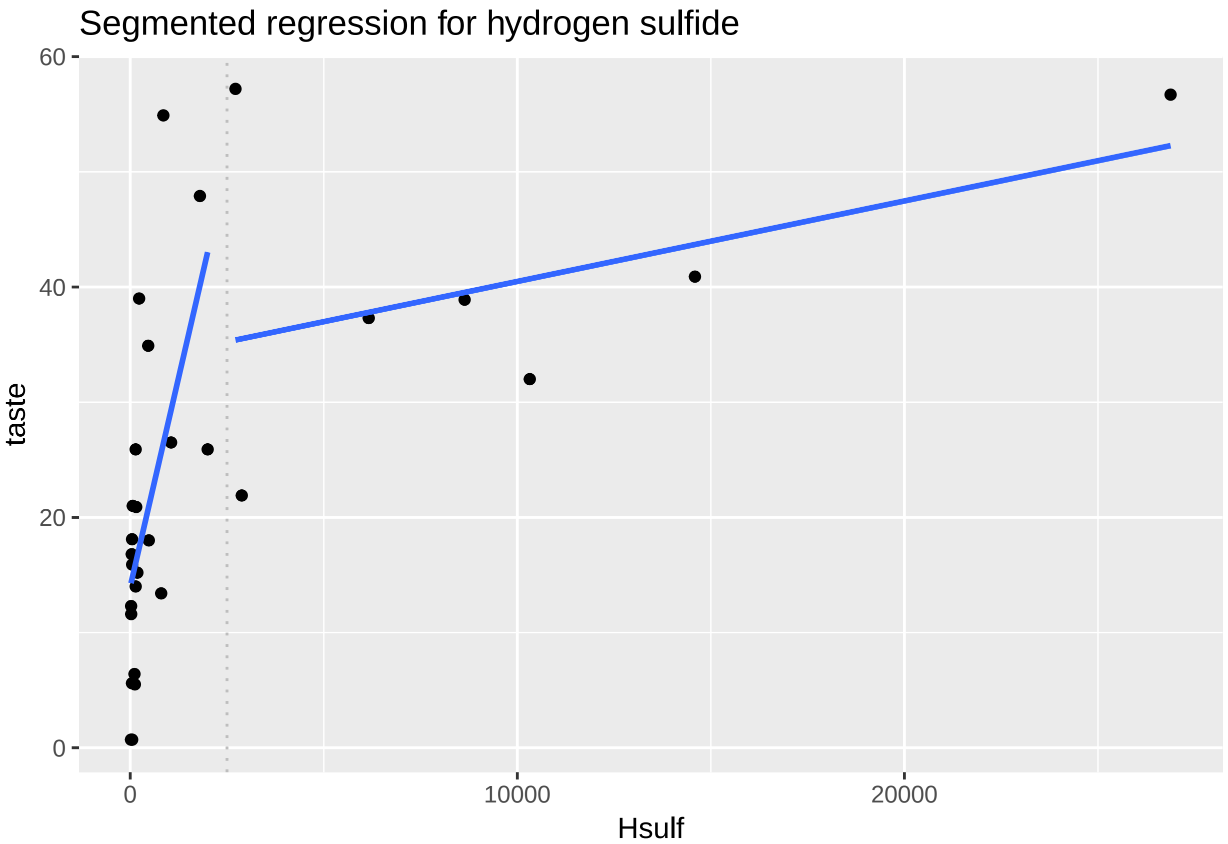
```
## Anova Table (Type III tests)
##
## Response: taste
##             Sum Sq Df F value  Pr(>F)
## (Intercept)    0.3  1  0.0017 0.96705
## Aacid        609.2  1  3.0331 0.09296 .
## Aacid2       221.3  1  1.1020 0.30313
## Residuals   5423.0 27
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The type III sum of squares for the acetic acid quadratic model is 609.2 for the unsquared predictor, and 221.3 for the squared predictor. 5,423 for the residuals.

12. Run a segmented regression model on hydrogen sulfide with a knot at concentration of 2500 mg/L (the U1.Hsulf estimate is the difference between it and the estimate for the previous segment; the segment segment's slope is .00083):

% Error: Unrecognized object type.

```
## Warning: Ignoring unknown parameters: method, se
```



$Hsulf Est. St.Err. t value CI(95%).l CI(95%).u slope1 0.089414 0.04329300 2.0653 4.2452e-04 0.1784000 slope2 0.000824 0.00040343 2.0425 -5.2674e-06 0.0016533

[1] 0.0424414 [1] 0.04948791

13. Hydrogen sulfide segmented regression model's coefficients and hypothesis tests:

first segment: coeff=.09239, and p=.0424

second segment: coeff=.00083, and p=.0495

H sulfide's associtionwith taste under 2500 mg/L is that taste points go up .0924 with every mg/L increase of H sulfide. It is a significant association (p=.0424).

The slope of the line above 2500 mg/L is .00083 points per mg/L increase in H suflide.

The association at lower values is not equal to the association at higher values, based on the significant p-values for each segment, reported above. Though oddly, the p-value for the test for existence of a break point is not significant:

```
##
##  Score test for one change in the slope
##
## data:  formula = taste ~ Hsulf ,   method = lm
## model = gaussian , link = identity
## segmented variable = Hsulf
## observed value = -1.243, n.points = 10, p-value = 0.2242
## alternative hypothesis: two.sided
```

A p-value tells you the probability of a result under assumptions that the null hypothesis is true. Alpha is an arbitrary bar set for determining statistical significance. If the p-value clears that bar, then the result is deemed significant.

14. MSE and Rˆ2 for the hydrogen sulfide segmented regression model:

MSE: 102.54

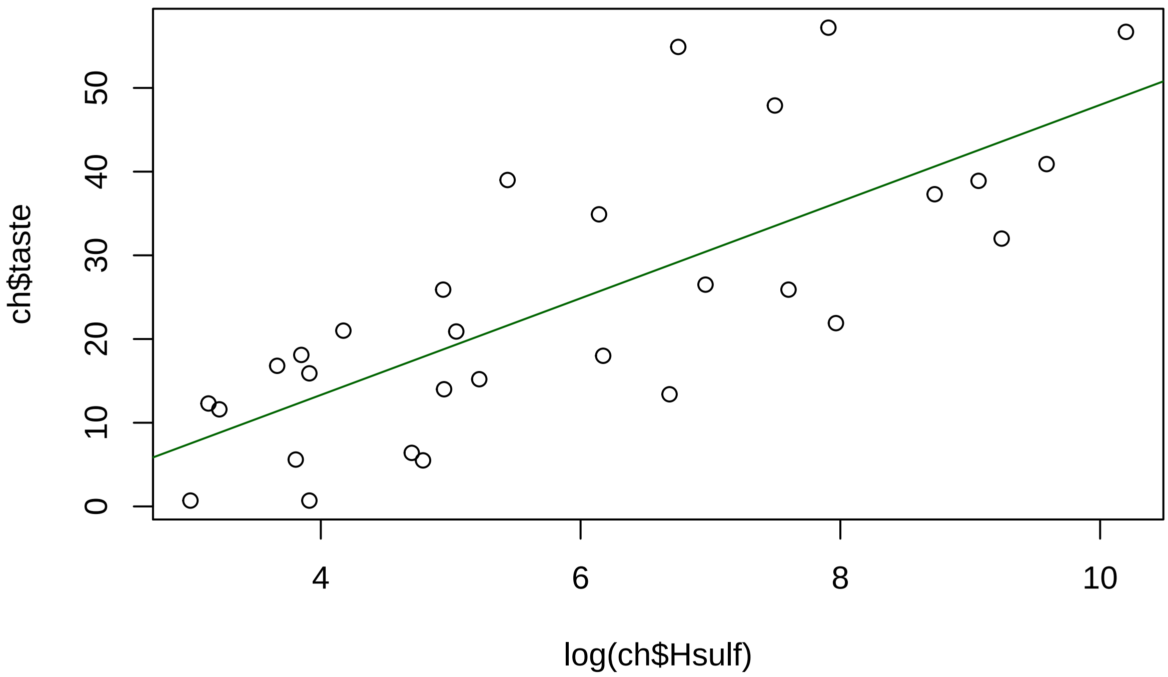
Rˆ2: .5985

15. Simple linear regression of log-transformed hydrogen sulfide:

Table 6: SLR model for log-transformed hydrogen sulfide

	Dependent variable:
	taste
logHS	5.776***
	(0.946)
Constant	−9.787
	(5.958)
Observations	30
R^2	0.571
Adjusted R^2	0.556
Residual Std. Error	10.833 (df = 28)
F Statistic	37.292*** (df = 1; 28)
Note:	*p<0.1; **p<0.05; ***p<0.01

Plot of the log-transformed data:



16. Comparing the hydrogen sulfide log-transformed SLR model to the segmented model (including definitions of the MSE and R^2 statistics):

Linear log-transformed model MSE and R^2: 109.53 and .5712

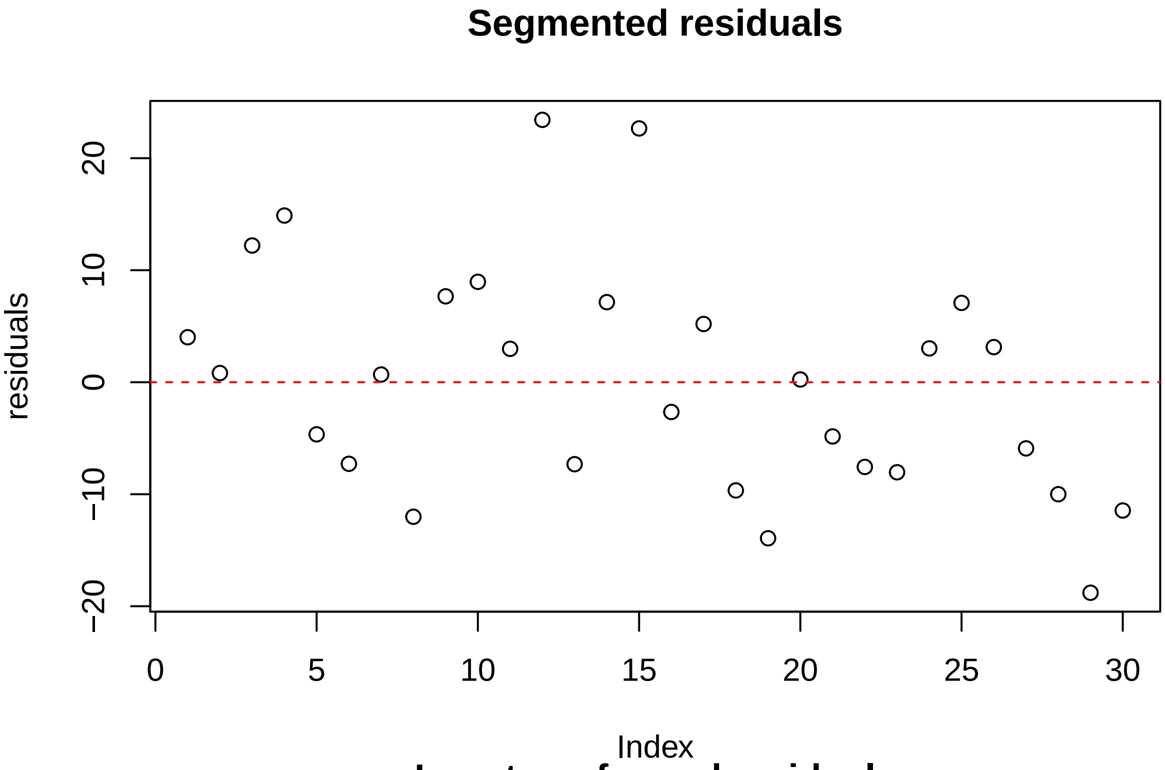
Segmented model MSE and R^2: 102.54 and .5985

Based on the foregoing MSE and R^2 values, the segmented model appears slightly preferrable.

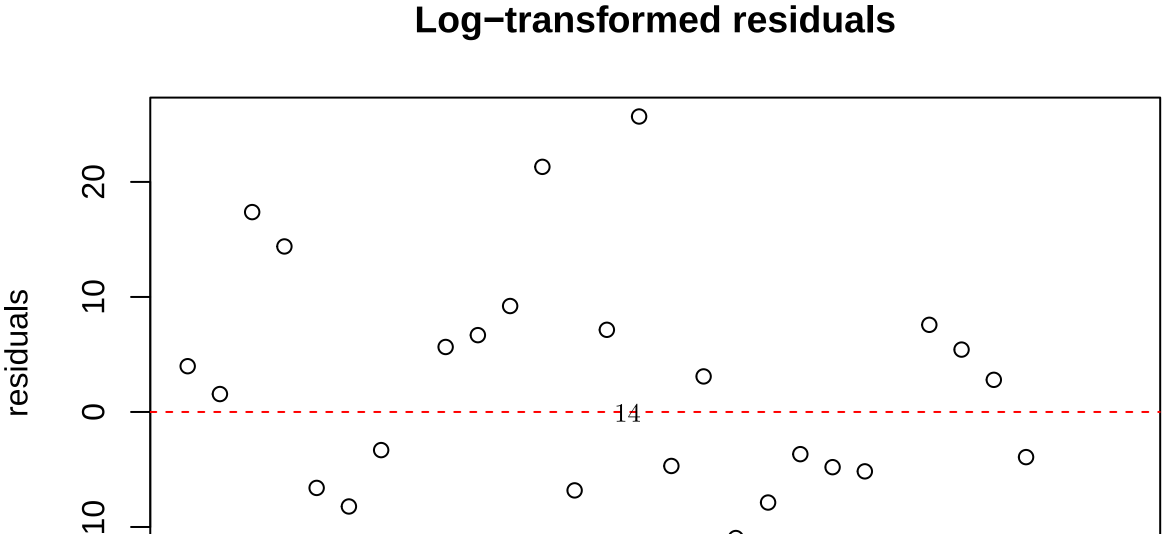
MSE is defined as the mean squared error, in other words, the average amount by which the observations deviate from the model, squared.

R^2 is the coefficient of determination, which tells the percentage of the change in the Y variable that our model can account for. An R^2 of .5985, for example, means that our model accounts for 59.85% of the variation in the dependent variable.

17. Residual plots for the hydrogen sulfide log-transformed SLR model and segmented model (which is preferrable?):



Segmented residuals



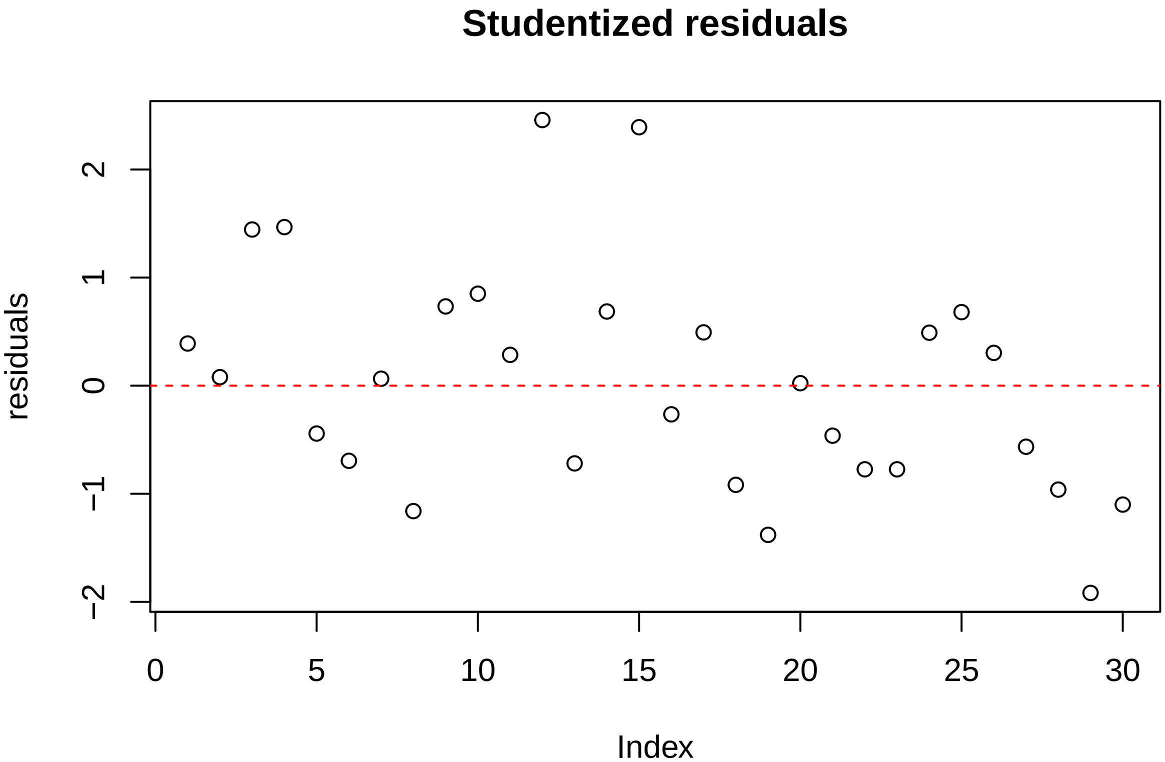
Log−transformed residuals

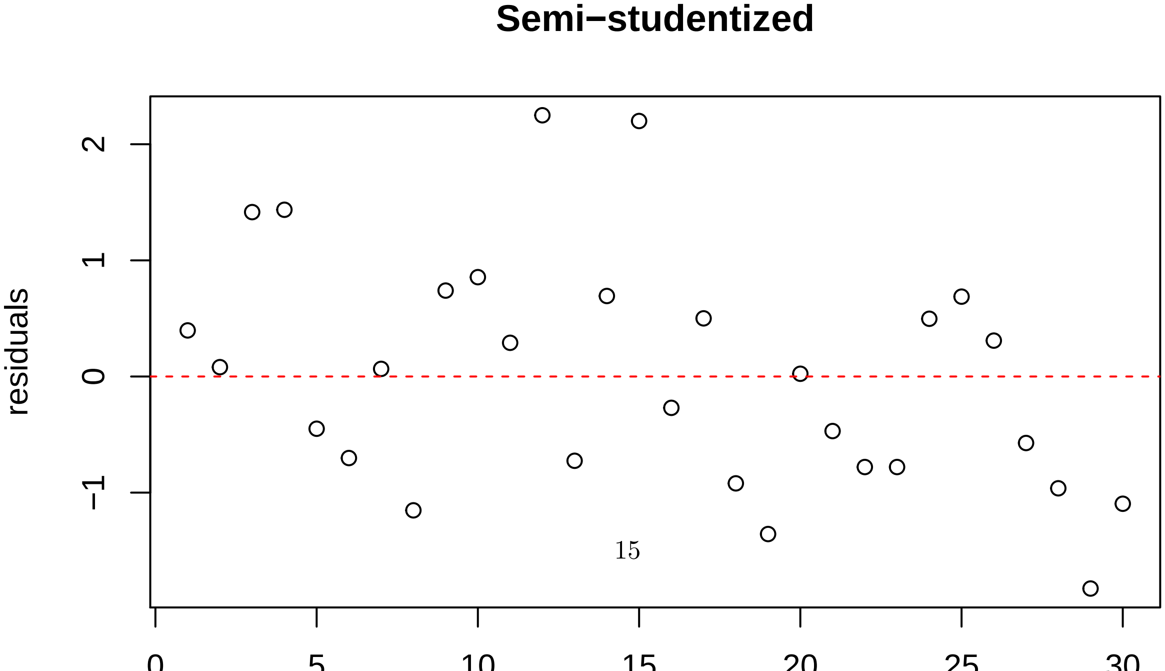
The residual plot for the segmented model looks a little more balanced to me, whereas the log-transformed SLR residuals are more top-heavy. I would say that based on these residual plots that the segmented model is preferrable.

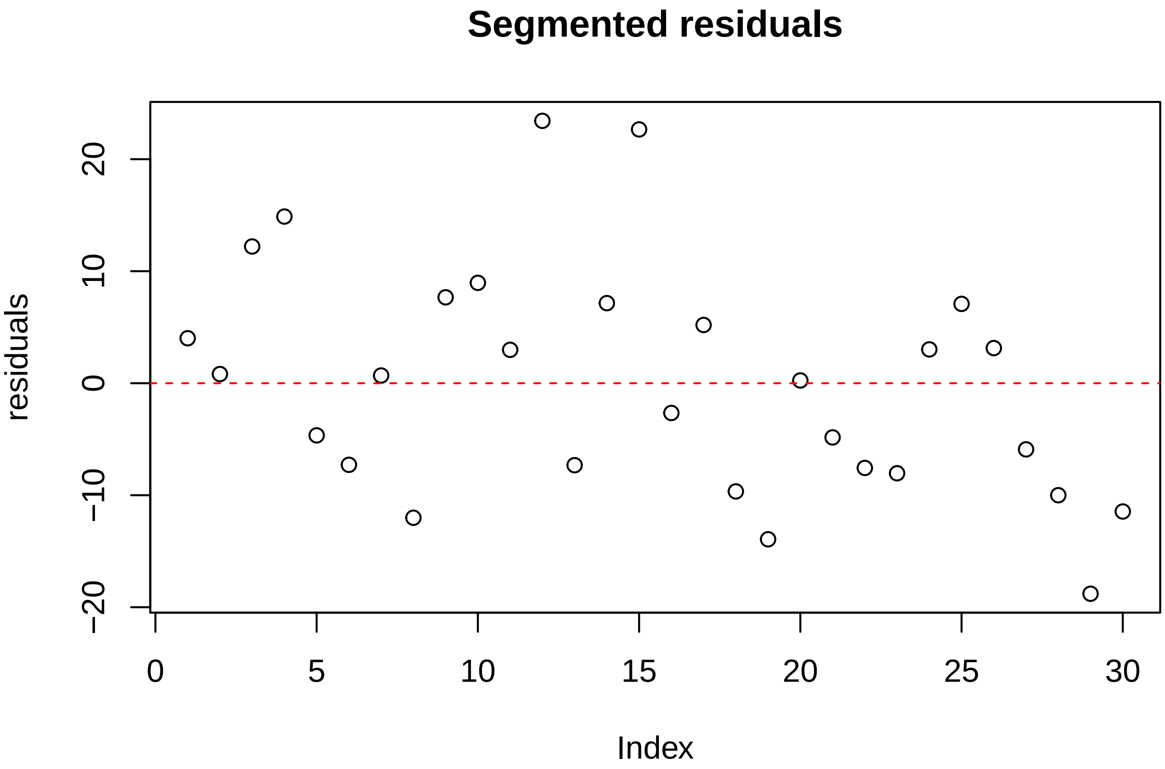
18. What kind of residuals did I use to perform the regression diagnostics, and why?

I used residuals that were not standardized (semi-studentized) or studentized, and they look exactly the same as the semi-studentized and studentized residuals. I would say that I should use studentized residuals in order to address the issue of outliers, since it does look like there are maybe some outliers, at least in the segmented model, but studentizing didn't change anything, as you can see below:

Segmented residuals (studentized, then semi-studentized, then un-studentized):






Segmented residuals
residuals
20
10
0
−10
−20
0
5
10
15
20
25
30
Index

I would plot all three versions of residuals for the log-transformed as well, but they all look identical to one another as well, and that pattern of similarity is illustrated above for the segmented model.

19. What are the consequences for a statistical model whose underlying assumptions are violated? Why do we care?

When the underlying assumptions of a statistical model are violated, the results from that model are no longer valid. We should care because from invalid results we will draw an invalid conclusion about our data and the potential relationships we're investigating.

20. What is the meaning of the intercept in a regression model?

The intercept represents the value of the dependent variable when the independent variable is zero. Usually it is not of scientific interest.

21. We do not always interpret the intercept when we report results from a regression model. Under what circumstances would we be able to interpret the intercept?

We would be able to interpret the intercept when there are predictor variable values that are close to, or span, zero. When it would not make sense for the predictor variable to have a value of zero, then I would think it would be inappropriate to interpret the intercept. For example, if BMI was our predictor variable for some continuous dependent variable, it would not make sense to report what the dependent variable's value would be predicted to be when BMI is zero, because BMI can never be zero.

22. Which multivariable regression model best fits the observations on taste? How did I arrive at my final model?

The full model above shows no significant interactions. This indicates that the reduced or main effects model is worth examining:

```
2.5 % 97.5 % (Intercept) -46.21346358 -8.07066654 Lacid 1.81641427 36.58636137 Aacid -0.02645859 0.03484837
logHS 1.32904794 6.34370461
```

Table 7: Full MLR model

	Dependent variable:
	taste
Lacid	-5.388 (48.827)
Aacid	-0.286 (0.325)
logHS	0.005 (15.118)
Lacid:Aacid	0.185 (0.217)
Lacid:logHS	2.263 (9.159)
Aacid:logHS	0.033 (0.058)
Lacid:Aacid:logHS	-0.020 (0.035)
Constant	11.935 (68.856)
Observations	30
R^2	0.680
Adjusted R^2	0.578
Residual Std. Error	10.558 (df = 22)
F Statistic	6.678^{***} (df = 7; 22)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 8: Reduced (main effects) MLR model

	Dependent variable:
	taste
Lacid	19.201**
	(8.458)
Aacid	0.004
	(0.015)
logHS	3.836***
	(1.220)
Constant	$-$27.142***
	(9.278)
Observations	30
R^2	0.653
Adjusted R^2	0.613
Residual Std. Error	10.116 (df = 26)
F Statistic	16.292*** (df = 3; 26)
Note:	*p<0.1; **p<0.05; ***p<0.01

The reduced model is my choice for the final model, given that no significant interactions exist.

23. Describe the association of taste with the chemicals using the coefficients and CIs from the final model. Recall in your description that the model's coefficients represent adjusted estimates.

Taste score increases 19 points (95% CI: 1.8 to 26.6) with every added mg/L concentration of lactic acid (p=.007). Taste score MAY increase .004 points with every added mg/L concentration of acetic acid, but the association is not significant (p=.781). Taste score increases 3.8 points (95% CI: 1.3 to 6.3) with every log-transformed added mg/L concentration of hydrogen sulfide (p=.0004). The estimated increases in taste score associated with increases in concentration of chemical are adjusted estimates.

24. Comparison between the coefficient and CI for the log-transformed hydrogen sulfide SLR model and the coefficient and CI for the log-transformed hydrogen sulfide in the MLR model:

SLR:

```
## (Intercept)       logHS
##   -9.786909    5.776095
```

```
##                 2.5 %   97.5 %
## (Intercept) -21.991270 2.417453
## logHS         3.838588 7.713602
```

As can be seen, the coefficients from the SLR model for log-transformed hydrogen sulfide and those of the same log-transformed chemical in the MLR model are different. Adjustment affects parameter estimates in MLR models.