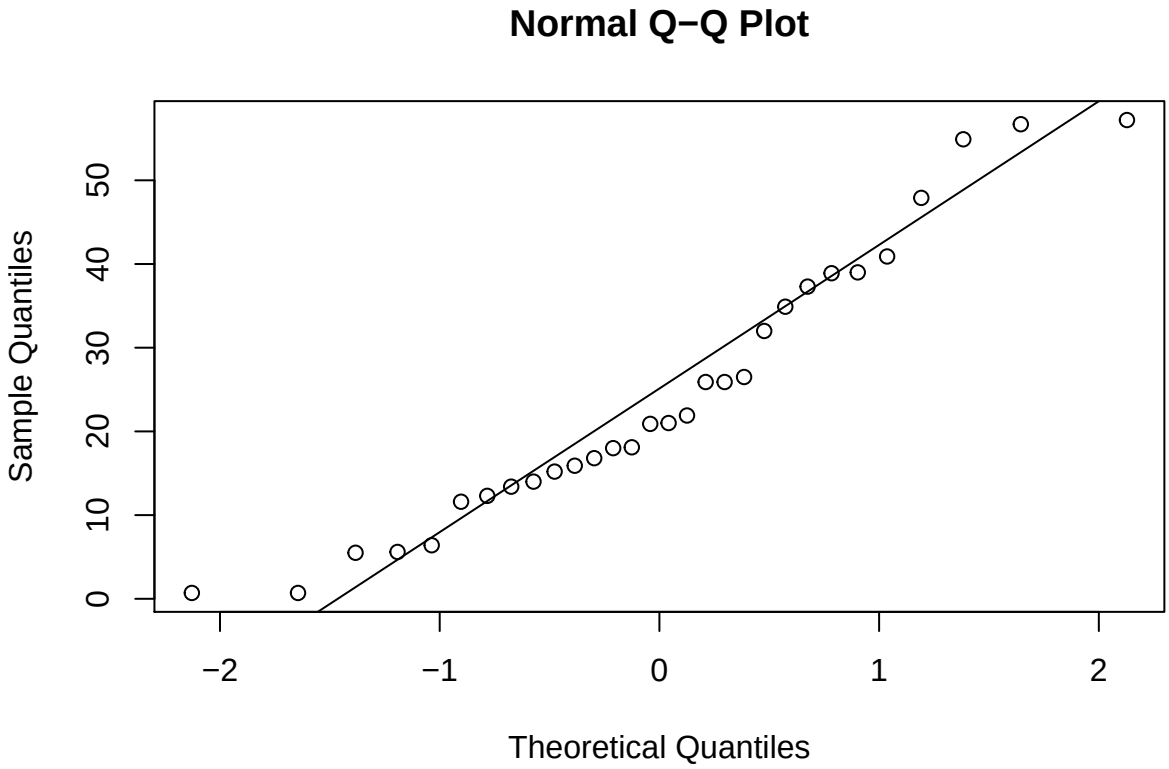
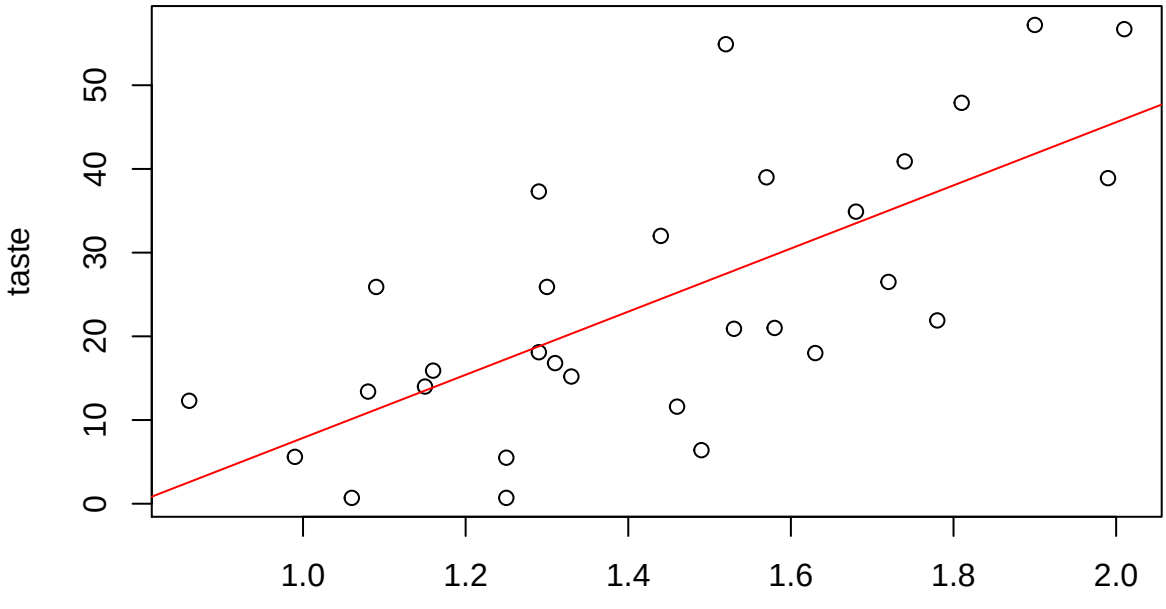
Zhang_exam1_PDF

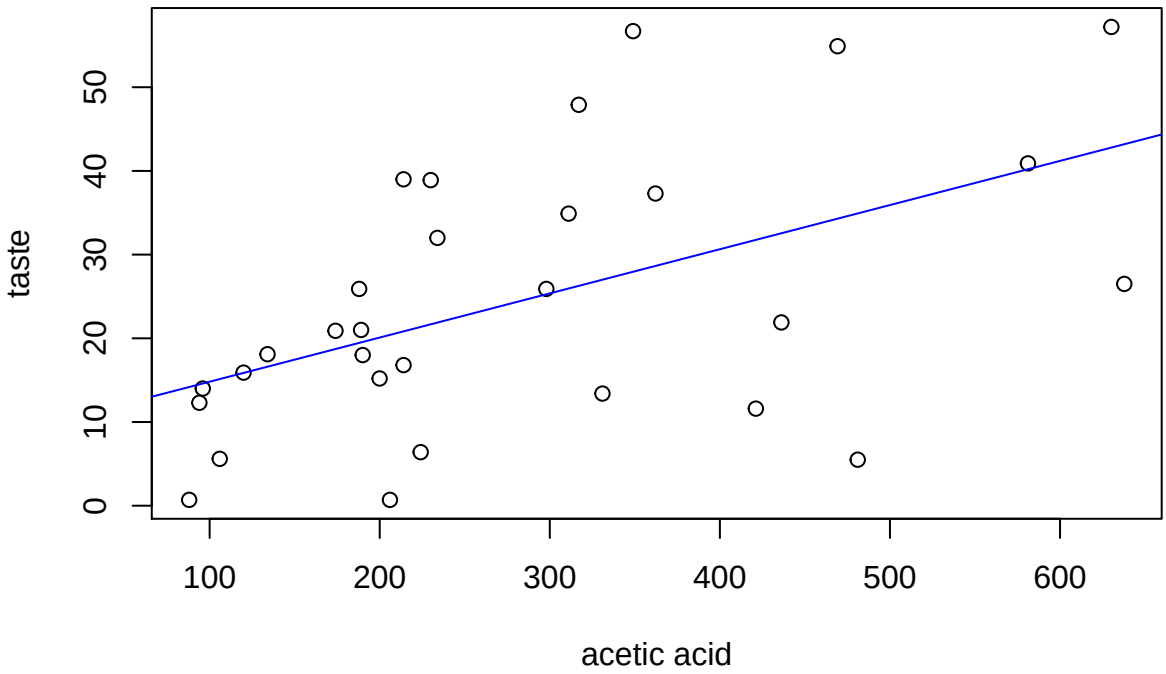
1. Why is violation of normality in the outcome variable not a problem:

Because in this case it's close enough to being normal:



2. SLR and LOESS plots; SLR models (Tables 1-3, with CIs following):





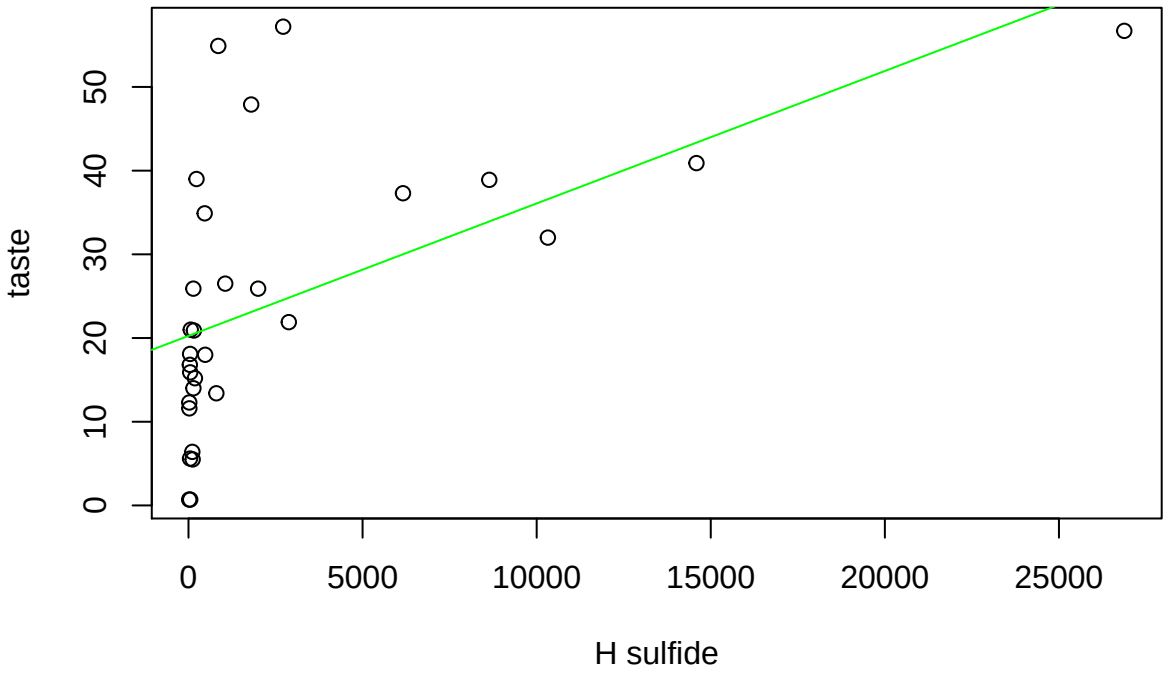


Table 1: SLR model for lactic acid ('Lacid')

	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

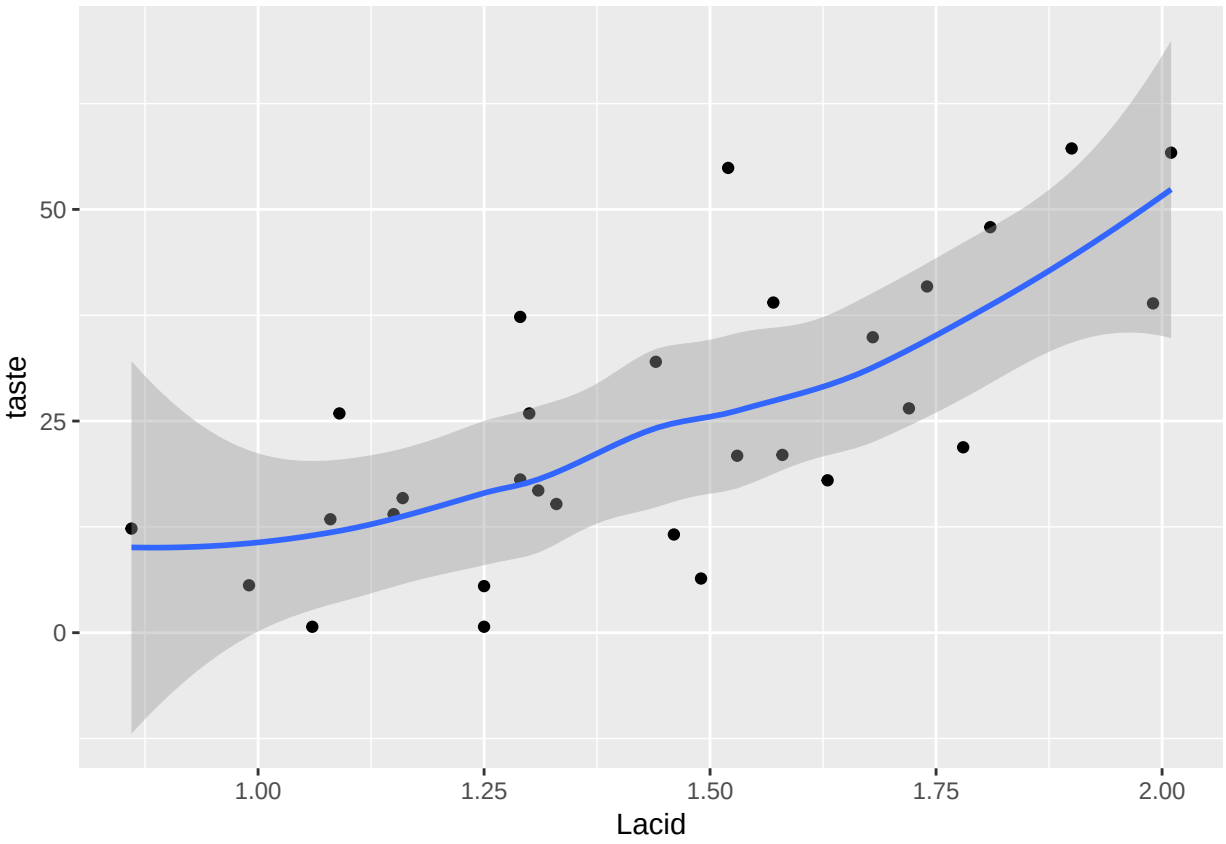
```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

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

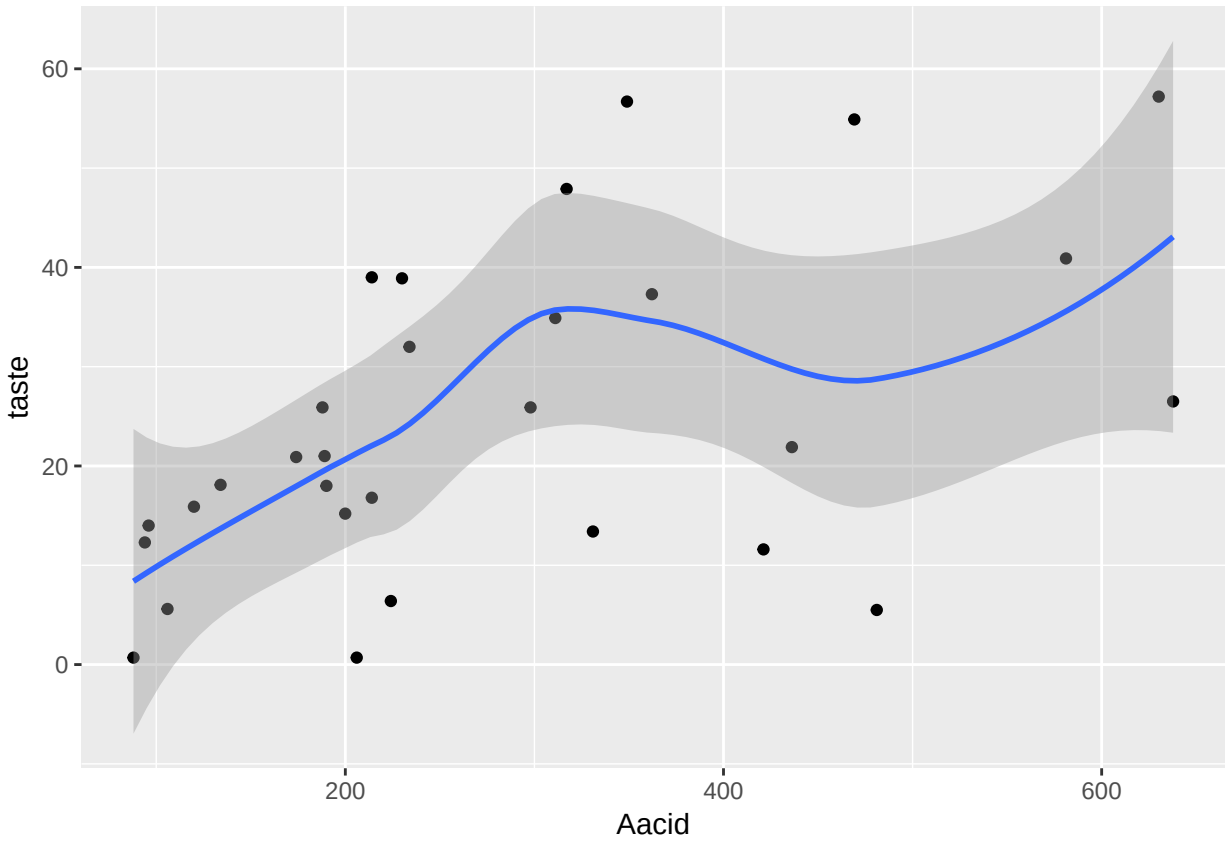
	Dependent variable:
	taste
Aacid	0.053^{***}
	(0.017)
Constant	9.539^{*}
	(5.401)
Observations	30
R^2	0.263
Adjusted R^2	0.237
Residual Std. Error	14.198 (df = 28)
F Statistic	10.014^{***} (df = 1; 28)
Note:	*p<0.1; **p<0.05; ***p<0.01

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

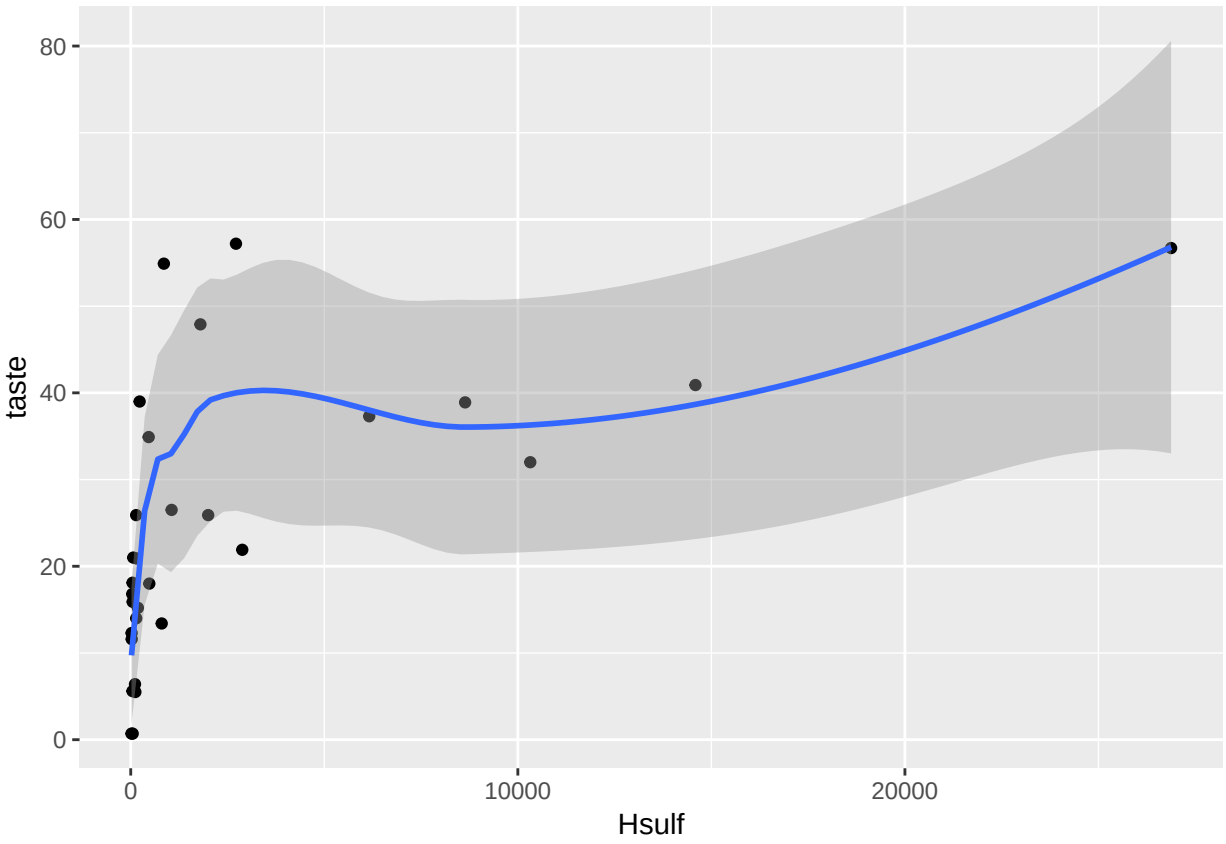
	Dependent variable:
	taste
Hsulf	0.002^{***}
	(0.0004)
Constant	20.256^{***}
	(2.772)
Observations	30
R^2	0.314
Adjusted R^2	0.290
Residual Std. Error	13.701 (df = 28)
F Statistic	12.824^{***} (df = 1; 28)
Note:	*p<0.1; **p<0.05; ***p<0.01




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




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


80
60
40
20
0
taste
0
10000
20000
Hsulf

3. Equations of models:

Association between taset and acetic acid:

Taste and lactic acid:

Taste and H sulfide:

4. Sentences quantifying the associations:

Taste and acetic acid:

Taste and lactic acid:

Taste and H 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. '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 estimates can change depending on what methods are used to calculate them, while the sample statistics shouldn't vary for the same data. An example illustrating this will come up below, in question 24.

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

8. Comparing linear and quadratic models for lactic acid (quadratic: Table 4):

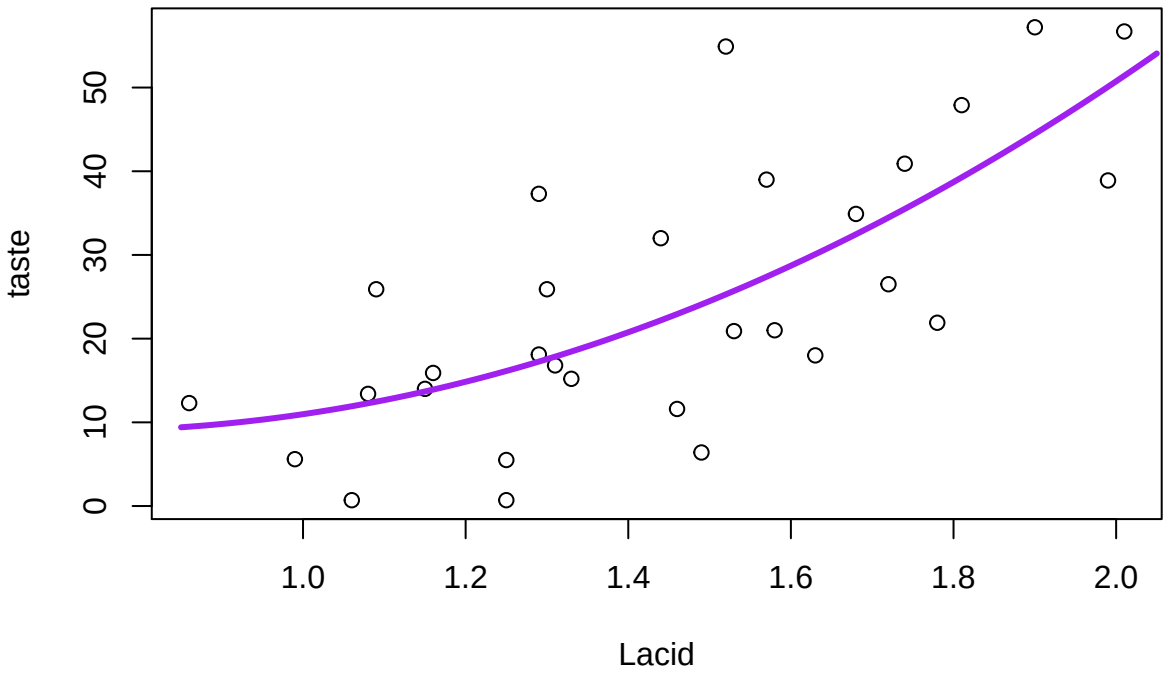


Table 4: Quadratic model for lactic acid

	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:

```
## [1] 128.7496
```

```
## [1] 0.4959486
```

MSE and R^2 for lactic acid quadratic model:

```
mse.Lq = mean(Lquad$residuals^2); mse.Lq
```

```
## [1] 122.7613
```

```
# Get r^2 from Lquad?
```

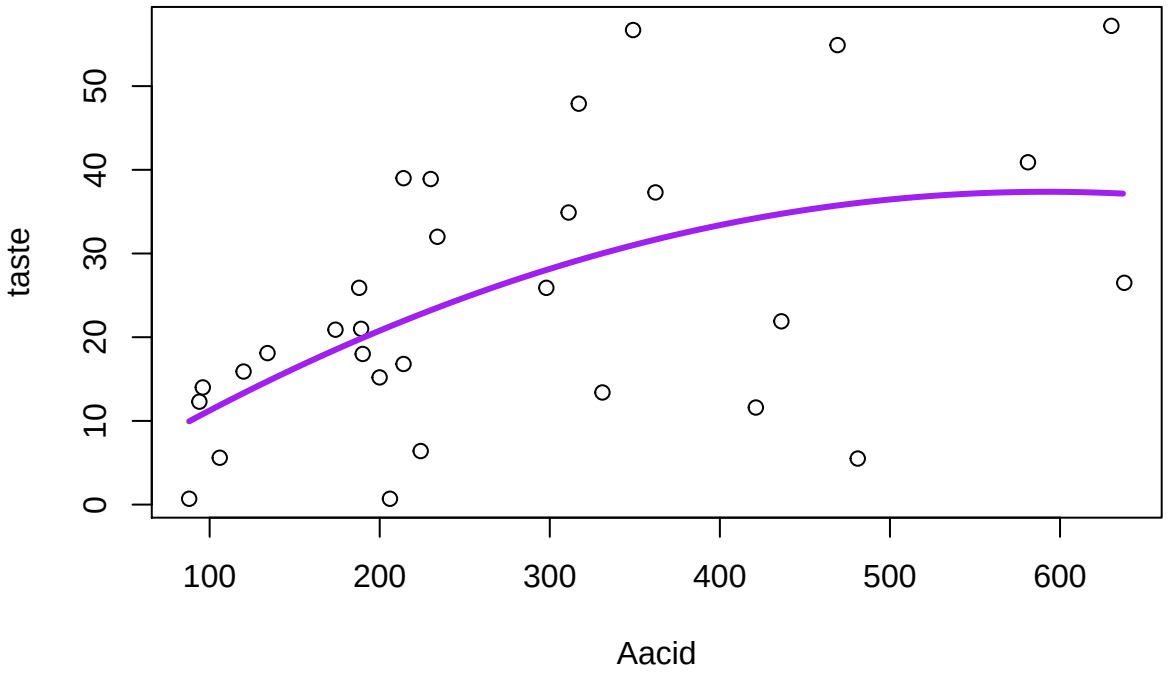
Table 5: Quadratic model for acetic acid

	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 R^2 value and the lower MSE.

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

10. Comparing linear and quadratic models for acetic acid (quadratic: Table 5):



MSE and R^2 for acetic acid linear model:

```
## [1] 188.1431
```

```
## [1] 0.2634246
```

MSE and R^2 for acetic acid quadratic model:

```
## [1] 180.7652
```

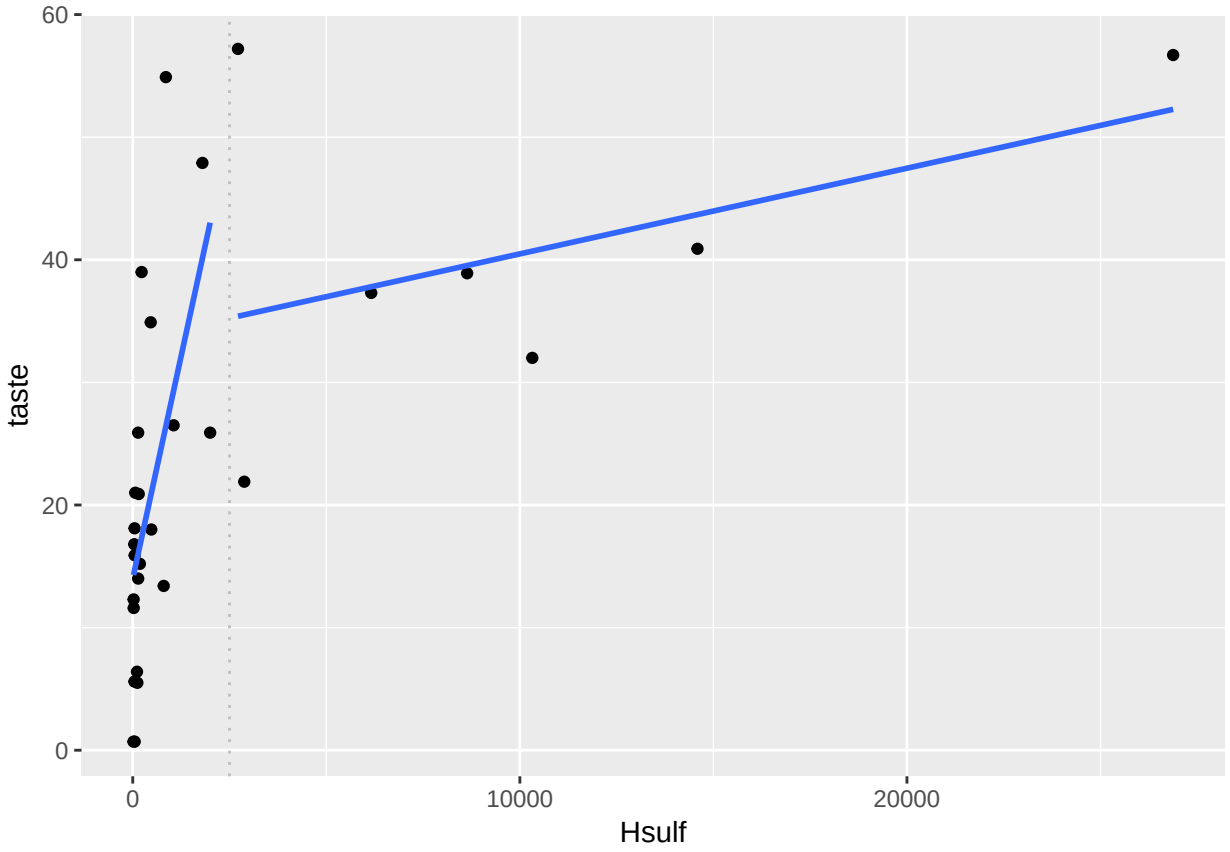
11. Define the type III SS for the acetic acid quadratic model:

```
## 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 acetic acid, and 221.3 for the squared acetic acid, with a residual SS of 5,423.

12. Run a segmented regression model on hydrogen sulfide with a knot at concentration of 2500 mg/L:

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



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

14. MSE and R^2 for the hydrogen sulfide segmented regression model:

15. Simple linear regression of log-transformed hydrogen sulfide (Table 6):

MSE and R^2 for the logHS SLR model and untransformed SLR model:

logHS:

```
## [1] 109.5392
```

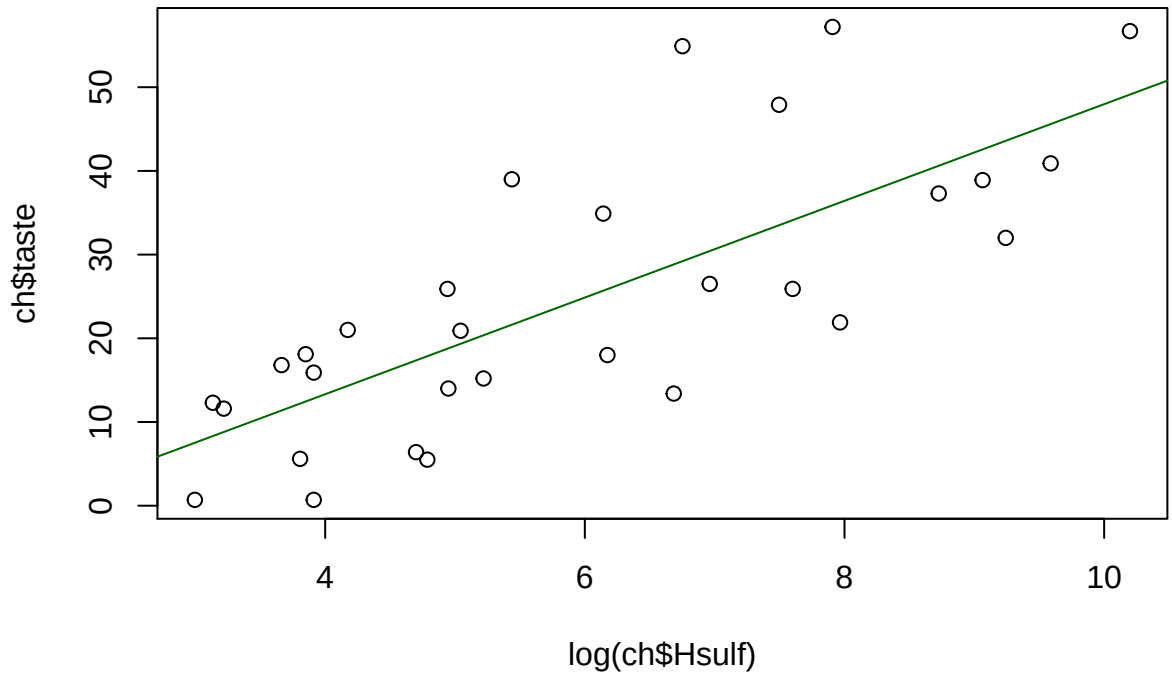
Untransformed:

```
## [1] 175.1911
```

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):

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

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

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

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

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?

22. Which multivariable regression model best fits the observations on taste? How did I arrive at my final model? (Full model is Table 7, and Reduced model is Table 8)

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

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.

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
```

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

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 . . .