

Stat 538 - Biostatistics I - Fall 2005

Lab 11

Multiple Linear Regression

Linear regression relating a response variable to more than one predictor variable.

For this lab we use a dataset from DASL, the Data and Story Library from Carnegie Mellon University. We will regress the taste of mature cheddar cheese on acetic acid, hydrogen sulfide, and lactic acid.

The data is found here and is reproduced below for convenience:

<http://lib.stat.cmu.edu/DASL/Datafiles/Cheese.html>

Reference:

Moore, David S., and George P. McCabe (1989). *Introduction to the Practice of Statistics*.

Description:

As cheese ages, various chemical processes take place that determine the taste of the final product. This dataset contains concentrations of various chemicals in 30 samples of mature cheddar cheese, and a subjective measure of taste for each sample. The variables "Acetic" and "H2S" are the natural logarithm of the concentration of acetic acid and hydrogen sulfide respectively. The variable "Lactic" has not been transformed.

Number of cases:

30

Variable Names:

Case: Sample number

Taste: Subjective taste test score, obtained by combining the scores of several tasters

Acetic: Natural log of concentration of acetic acid

H2S: Natural log of concentration of hydrogen sulfide

Lactic: Concentration of lactic acid

The Data:

Case	taste	Acetic	H2S	Lactic
1	12.3	4.543	3.135	0.86
2	20.9	5.159	5.043	1.53
3	39	5.366	5.438	1.57
4	47.9	5.759	7.496	1.81
5	5.6	4.663	3.807	0.99
6	25.9	5.697	7.601	1.09
7	37.3	5.892	8.726	1.29
8	21.9	6.078	7.966	1.78
9	18.1	4.898	3.85	1.29
10	21	5.242	4.174	1.58
11	34.9	5.74	6.142	1.68
12	57.2	6.446	7.908	1.9
13	0.7	4.477	2.996	1.06
14	25.9	5.236	4.942	1.3
15	54.9	6.151	6.752	1.52
16	40.9	6.365	9.588	1.74
17	15.9	4.787	3.912	1.16
18	6.4	5.412	4.7	1.49
19	18	5.247	6.174	1.63
20	38.9	5.438	9.064	1.99
21	14	4.564	4.949	1.15
22	15.2	5.298	5.22	1.33
23	32	5.455	9.242	1.44
24	56.7	5.855	10.199	2.01
25	16.8	5.366	3.664	1.31
26	11.6	6.043	3.219	1.46
27	26.5	6.458	6.962	1.72
28	0.7	5.328	3.912	1.25
29	13.4	5.802	6.685	1.08
30	5.5	6.176	4.787	1.25

This lab's plan:

1. look at correlations and scatterplots of the variables.
2. look at all possible linear regression models of the original variables.
3. choose a model, evaluate the assumptions of normality, equal variance, and independence of the residuals versus everything we can think of.
4. comment on how well we expect the model to predict new observations.
5. interpret the regression coefficients (the numbers multiplied by each of our predictor variables).
6. discuss any unusual observations.

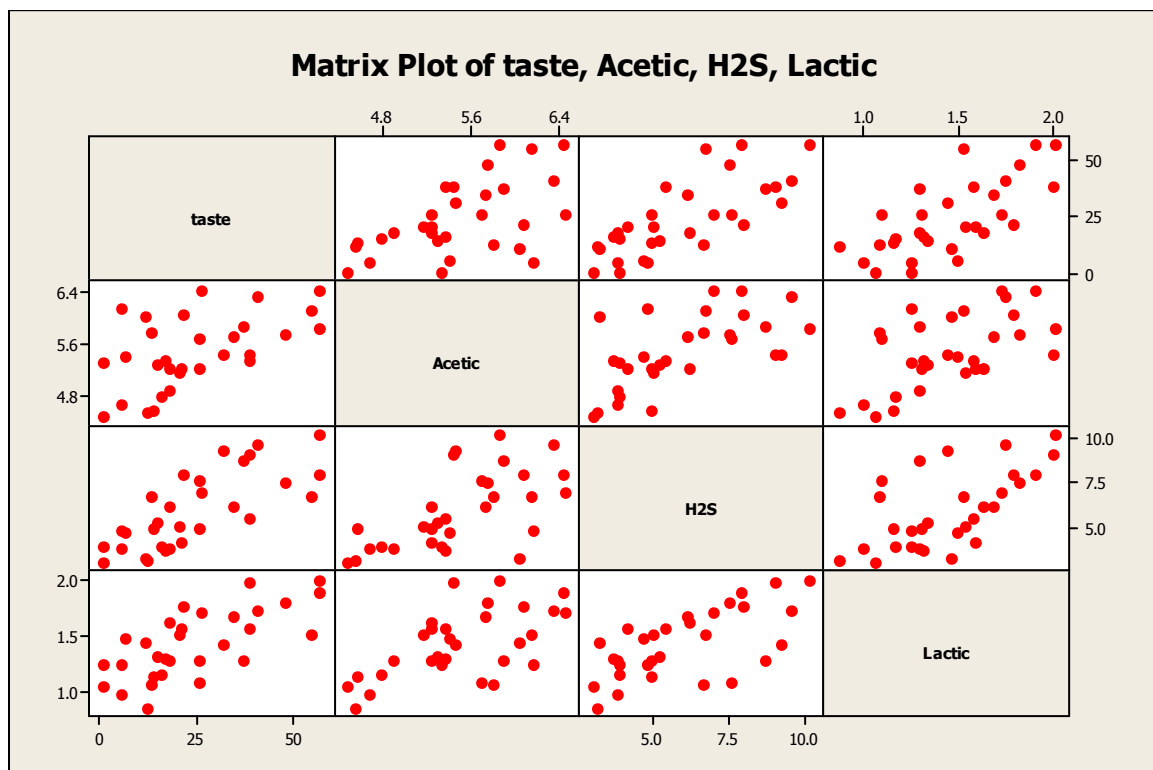
Note that this order is slightly different from the order in homework 10. Please follow the order in the homework when writing up the homework. Also, I cover much more here than is asked for in the homework.

Before fitting the regression, let's first have a look at the relationships between all the variables. *Always plot the data first.* First, we notice that all the correlations are significant (all p-values are less than 0.01), and the correlations are moderate from 0.550 to 0.756. In the scatterplot matrix, we see the positive correlations between all the variables.

Correlations: taste, Acetic, H2S, Lactic

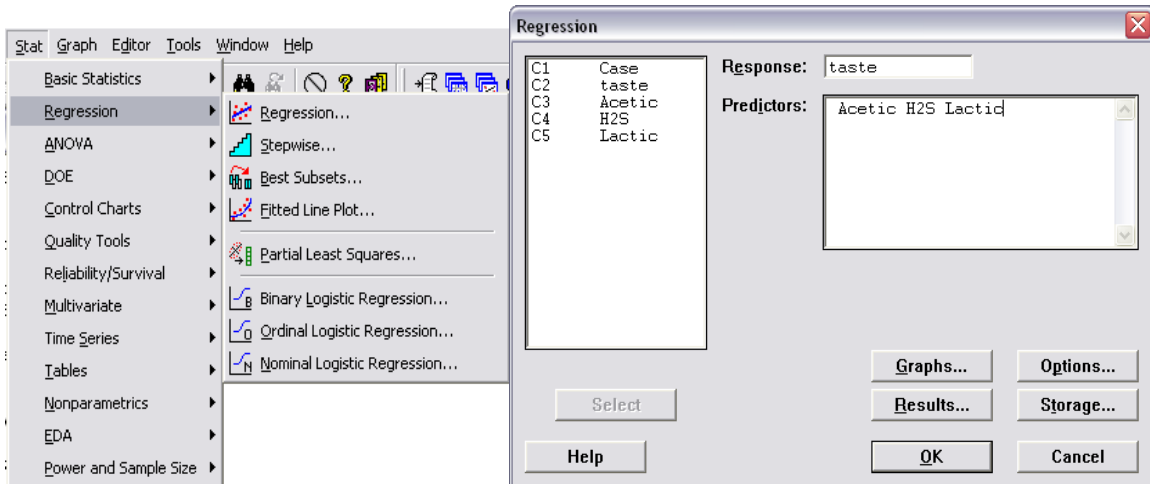
	taste	Acetic	H2S
Acetic	0.550 0.002		
H2S	0.756 0.000	0.618 0.000	
Lactic	0.704 0.000	0.604 0.000	0.645 0.000

Cell Contents: Pearson correlation
P-Value



There are many ways building regression models from a number of predictor variables. There are both forward and backward stepwise methods, best subset selection, and others. At the end, I'll show briefly a backward stepwise selection. Because we have only three variables, let's look at all possible models. There will be 7 of them, 3 with one variable, 3 with two variables, and 1 with all three.

We are going to regress taste on Acetic, H2S, and Lactic. Choose Stat/Regression/Regression. Select taste for the response variable, and we will select combinations of Acetic, H2S, and Lactic for the predictor variables.



Simple linear regression models (one predictor variable)

Let's first look at each predictor variable individually. Below is abbreviated output for the three regression models with a single predictor variable.

Regression Analysis: taste versus Acetic

The regression equation is
taste = - 61.5 + 15.6 Acetic

Predictor	Coef	SE Coef	T	P
Constant	-61.50	24.85	-2.48	0.020
Acetic	15.648	4.496	3.48	0.002

S = 13.8212 **R-Sq = 30.2%** R-Sq(adj) = 27.7%

Regression Analysis: taste versus H2S

The regression equation is
taste = - 9.79 + 5.78 H2S

Predictor	Coef	SE Coef	T	P
Constant	-9.787	5.958	-1.64	0.112
H2S	5.7761	0.9458	6.11	0.000

S = 10.8334 **R-Sq = 57.1%** R-Sq(adj) = 55.6%

Regression Analysis: taste versus Lactic

The regression equation is
taste = - 29.9 + 37.7 Lactic

Predictor	Coef	SE Coef	T	P
Constant	-29.86	10.58	-2.82	0.009
Lactic	37.720	7.186	5.25	0.000

S = 11.7450 **R-Sq = 49.6%** R-Sq(adj) = 47.8%

In each case, the predictor variable explains a significant amount of the variation of taste, since the p-value for each slope coefficient is less than 0.01. Notice that H2S has the largest R-Sq=0.571, so it makes the best choice for a one predictor variable model.

Multiple linear regression models (two predictor variables)

Next we look at each pair of predictor variables. Below is abbreviated output for the three regression models with pairs of predictor variables.

Regression Analysis: taste versus Acetic, H2S

The regression equation is

$$\text{taste} = -26.9 + 3.80 \text{ Acetic} + 5.15 \text{ H2S}$$

Predictor	Coef	SE Coef	T	P
Constant	-26.94	21.19	-1.27	0.215
Acetic	3.801	4.505	0.84	0.406
H2S	5.146	1.209	4.26	0.000

S = 10.8896 **R-Sq = 58.2%** R-Sq(adj) = 55.1%

Regression Analysis: taste versus Acetic, Lactic

The regression equation is

$$\text{taste} = -51.4 + 5.57 \text{ Acetic} + 31.4 \text{ Lactic}$$

Predictor	Coef	SE Coef	T	P
Constant	-51.37	21.17	-2.43	0.022
Acetic	5.571	4.761	1.17	0.252
Lactic	31.392	8.956	3.51	0.002

S = 11.6684 **R-Sq = 52.0%** R-Sq(adj) = 48.5%

Regression Analysis: taste versus H2S, Lactic

The regression equation is

$$\text{taste} = -27.6 + 3.95 \text{ H2S} + 19.9 \text{ Lactic}$$

Predictor	Coef	SE Coef	T	P
Constant	-27.592	8.982	-3.07	0.005
H2S	3.946	1.136	3.47	0.002
Lactic	19.887	7.959	2.50	0.019

S = 9.94236 **R-Sq = 65.2%** R-Sq(adj) = 62.6%

The first thing I noticed was that when Acetic is in the model with either H2S or Lactic, Acetic is no longer significant (p-value larger than 0.25). Why would this be? Recall, these variables are correlated, which means that they in part describe the same information. H2S and Lactic both describe information contained in Acetic, which makes Acetic no longer needed in the model. However, in the model including both H2S and Lactic, both predictor variables are significant. So, while they are correlated, their contributed information to the model is somewhat different. What can I say, the relationships between variables can be complex. In the homework we see the opposite situation, where an individual variable is not significant until another variable is in the model with it.

Note the model with H2S and Lactic has the largest R-Sq=0.652.

Multiple linear regression models (Full model, all three predictor variables)

Next we look at the full model. The full model is simply the model including all predictor variables available.

Regression Analysis: taste versus Acetic, H2S, Lactic

The regression equation is

$$\text{taste} = -28.9 + 0.33 \text{ Acetic} + 3.91 \text{ H2S} + 19.7 \text{ Lactic}$$

Predictor	Coef	SE Coef	T	P
Constant	-28.88	19.74	-1.46	0.155
Acetic	0.328	4.460	0.07	0.942
H2S	3.912	1.248	3.13	0.004
Lactic	19.671	8.629	2.28	0.031

S = 10.1307 **R-Sq = 65.2%** R-Sq(adj) = 61.2%

Unusual Observations

Obs	Acetic	taste	Fit	SE Fit	Residual	St Resid
15	6.15	54.90	29.45	3.04	25.45	2.63R

R denotes an observation with a large standardized residual.

Notice that Acetic has a p-value of 0.942, which is as close to 1 as you'll usually get. Predictor variable Acetic is not required in a model with both H2S and Lactic. We have a large R-Sq=0.652.

The R-Sq value will never get smaller as you add variables, in practice always increases. Notice that this R-Sq=0.652 is the same as the two variable model with just H2S and Lactic. The addition of Acetic has not improved our ability to predict new observations at all (or at least no improvement to 3 decimal places).

The method of backwards stepwise regression starts with the full model, then drops the least significant variable *one at a time* until all of the variables are significant. Let's removed Acetic from our model and continue the remaining analysis on the reduced model including on ly H2S and Lactic.

Reduced model

When I fit this model of taste on H2S and Lactic, I will output the four-in-one plot of residuals, as well as storage of residuals (to check normality).

Regression Analysis: taste versus H2S, Lactic

The regression equation is

$$\text{taste} = -27.6 + 3.95 \text{ H2S} + 19.9 \text{ Lactic}$$

Predictor	Coef	SE Coef	T	P
Constant	-27.592	8.982	-3.07	0.005
H2S	3.946	1.136	3.47	0.002
Lactic	19.887	7.959	2.50	0.019

S = 9.94236 **R-Sq = 65.2%** R-Sq(adj) = 62.6%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	4993.9	2497.0	25.26	0.000
Residual Error	27	2669.0	98.9		
Total	29	7662.9			

Unusual Observations

Obs	H2S	taste	Fit	SE Fit	Residual	St Resid
15	6.8	54.90	29.28	1.95	25.62	2.63R

R denotes an observation with a large standardized residual.

Above we see the regression equation.

$$\text{taste} = -27.6 + 3.95 \text{ H2S} + 19.9 \text{ Lactic}$$

This shows how taste is linearly related to H2S and Lactic given our data. If you wanted to predict a new taste value for a set of H2S and Lactic values, you would plug into the equation your value of H2S and value of Lactic, then just multiply and add to get the predicted taste.

Next, we see that both our predictor variables H2S and Lactic are significant. The three p-values in this table are part of different hypothesis tests. The model we are looking at is

$$\text{Taste} = \beta_0 + \beta_1 \text{ H2S} + \beta_2 \text{ Lactic}$$

Here are our three hypotheses:

For the intercept term, β_0

$$H_0: \beta_0 = 0.$$

$$H_A: \beta_0 \neq 0.$$

For the slope term for H2S, β_1

$$H_0: \beta_1 = 0.$$

$$H_A: \beta_1 \neq 0.$$

For the slope term for Lactic, β_2

$$H_0: \beta_2 = 0.$$

$$H_A: \beta_2 \neq 0.$$

Because the p-values are less than 0.05 for all three coefficients in the table, we reject H_0 in each of these hypotheses, in favor of their alternatives H_A that the coefficients are not 0. Each of these tests are based on the other coefficients being what they are estimated to

be in the model. Therefore, they are conditional tests. For example, the test for β_1 is given that the values of β_0 and β_2 are what they are in the model.

They explain $R\text{-Sq}=0.652=65.2\%$ of the variation in taste. That's a moderately high $R\text{-Sq}$, so it will make decent predictions.

By the way, the ANOVA table tests the hypothesis that all the regression coefficients are equal to 0 or at least one is different from 0.

For the intercept term, β_0

$$H_0: \beta_0 = \beta_1 = \beta_2 = 0.$$

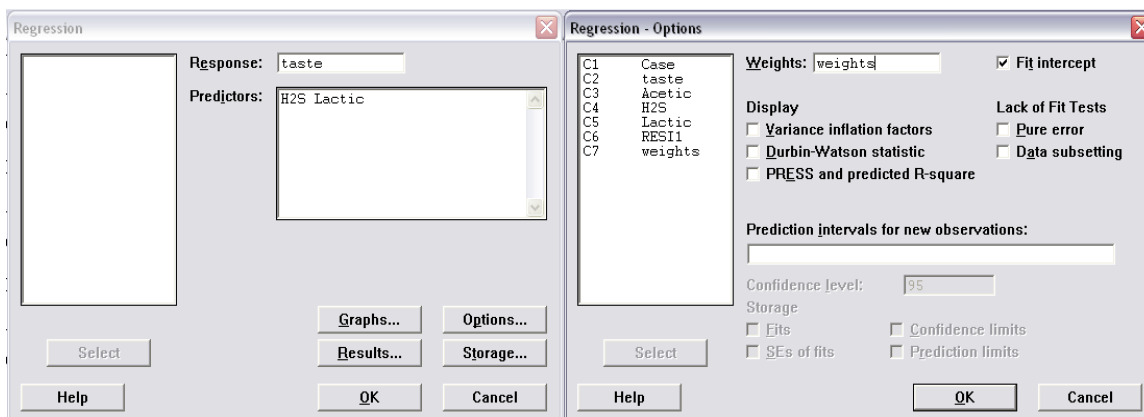
H_A : at least one β is different from 0.

This p-value is almost always very small. But if it is large, forget it, your predictor variables do not describe the variation in the response variable (at least not in the combination you have specified, maybe with more or less variables).

Lastly, we have the unusual observations. This points out that observation 15 (row 15 in the data worksheet) has a large standardized residual (R). If you hover your mouse over the points in the residual plots in Minitab, it will tell you the observation numbers. This observation is the largest in the normal probability plot, the right outlier in the histogram, and the largest value in the center of the fitted values scatterplot. It is 2.63 standard deviations from the mean, which is large, but not wildly unexpected in 30 observations. Without any good reason for removing it, it will stay.

Removing an observation from our Final model

For illustration, let's fit the model leaving out this observation. Create a column of weights that are 1 for observations we want to retain (all but 15) and 0 for observations we wish to omit (only row 15). Fit the regression again, but under Options, choose the weight variable for Weights.



Without observation 15, the regression equation is practically the same. The R-Sq increased a little (from 65.2% to 70.4%) because that observation was large in the y direction, which is far from a line. But now we have another unusual observation (12). Our main conclusions here should be that removing this observation did not change the significance of any of our predictor variables, the regression equation did not change much (the coefficients are basically still -27.592, 3.946, and 19.887), and the R-Sq is not very different. Thus, this observation is not severely effecting our model.

Regression Analysis: taste versus H2S, Lactic

Weighted analysis using weights in weights

The regression equation is

$$\text{taste} = -27.4 + 3.79 \text{ H2S} + 19.8 \text{ Lactic}$$

29 cases used, 1 cases contain missing values
 or had zero weight

Predictor	Coef	SE Coef	T	P
Constant	-27.449	7.897	-3.48	0.002
H2S	3.7872	0.9999	3.79	0.001
Lactic	19.828	6.997	2.83	0.009

S = 8.74095 R-Sq = 70.4% R-Sq(adj) = 68.1%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	4722.4	2361.2	30.90	0.000
Residual Error	26	1986.5	76.4		
Total	28	6709.0			

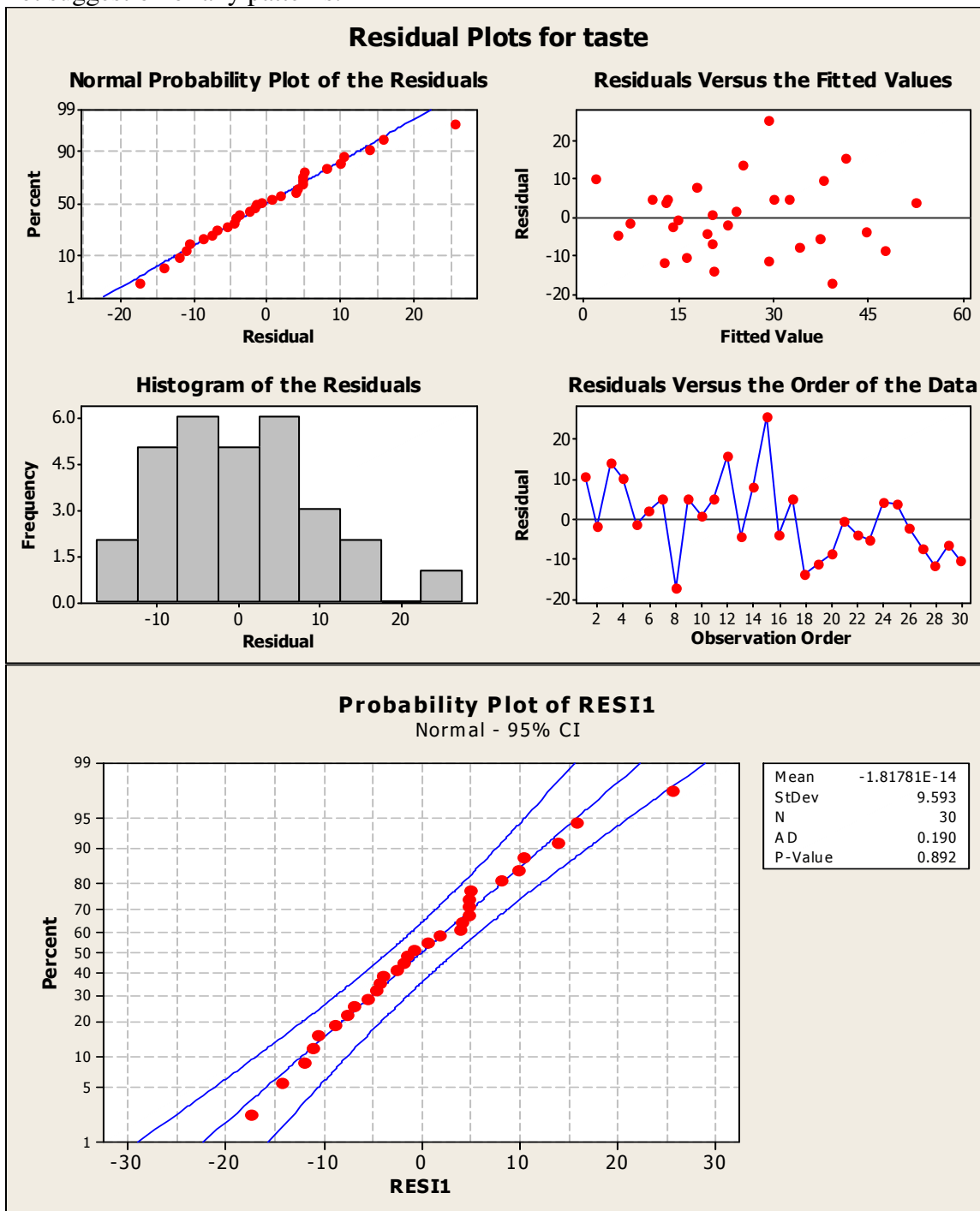
Unusual Observations

Obs	H2S	taste	Fit	SE Fit	Residual	St Resid
12	7.9	57.20	40.17	2.95	17.03	2.07R

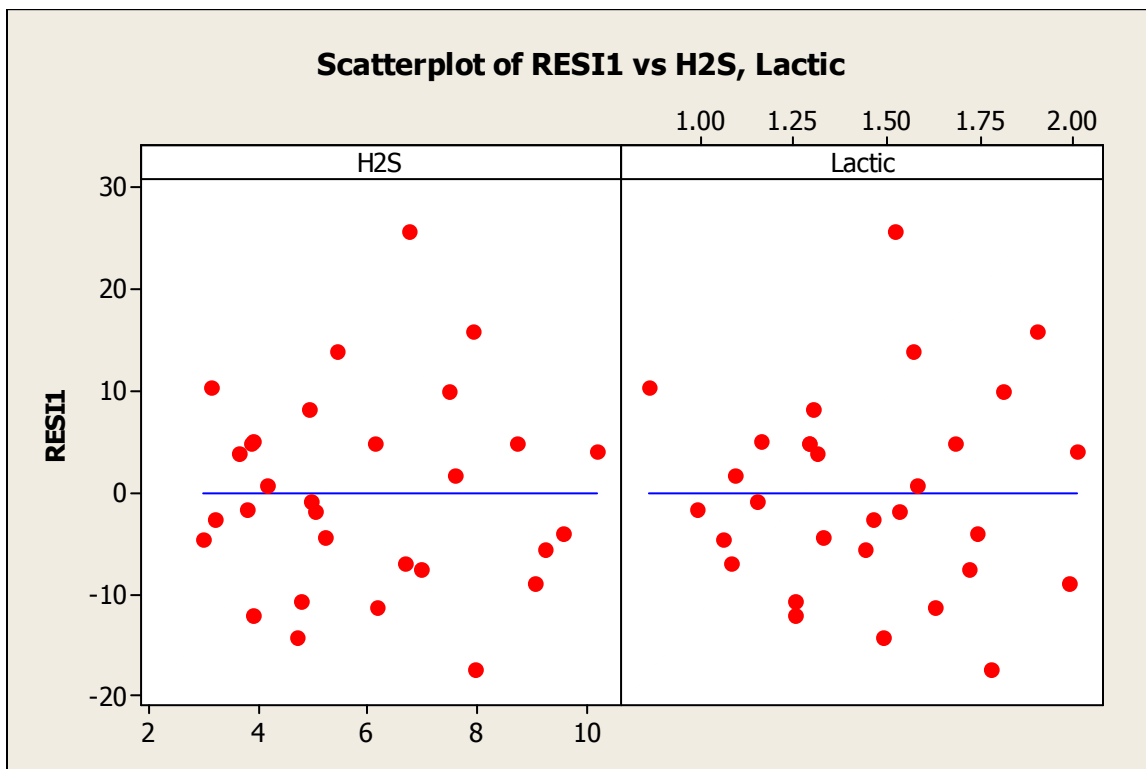
R denotes an observation with a large standardized residual.

Checking model assumptions with residual plots of the Final reduced model with all observations.

Below are the residual plots. On the left, it appears that normality is satisfied by both the normal probability plot and histogram. Additionally, the normal probability plot at the bottom looks very good, and the Anderson-Darling normality test gives a p-value of 0.892. The scatterplot to the right does not deviate from constant variance and there is not suggestion of any patterns.



Additionally, this scatterplot of the residuals versus our two predictor variables, indicates that the residuals appear not to be related to either of our predictor variables. In fact, the residuals should be completely random, and independent of everything.



You can get this plot by choosing Graph/Scatterplot, with regression, choose RESI1 as the Y variable, and the predictor variables as the X variables, then multiple graphs.

