U Wroclaw, Fall 2015
Applied Stats

DISCUSSION/LAB 6: MULTIPLE LINEAR REGRESSION ANIA

We will use data set: cheese.MTW (.xlsx).

TO DO

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:

1. Case: Sample number
2. Taste: Subjective taste test score, obtained by combining the scores of several tasters
3. Acetic: Natural log of concentration of acetic acid
4. H2S: Natural log of concentration of hydrogen sulfide
5. 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

……………………

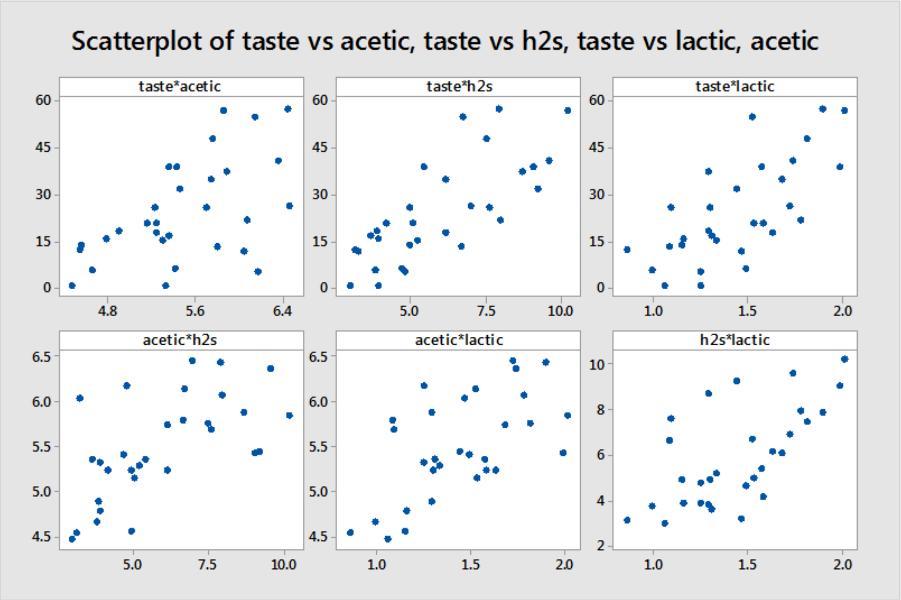
Goal: Find a model that predicts taste from the chemical variables.

For a model with all three explanatory variables find/compute/test

- Correlations between the explanatory variables and between explanatory variables and the response variable, graph scatter plots.
- Model equation,
- Discuss diagnostic plots,
- Are the slopes significantly different from zero on 5% significance level?
- Estimate of variance of the error term σ^2, that is s^2.
- What is the Pearson correlation coefficient between observed and predicted responses?
- Predict the value of Y for x1=1, and x2=2,
- Find a 99% confidence interval for the mean value of Y when x1=1, and x2=2,
- Find a 99% prediction interval for the mean value of Y when x1=1, and x2=2,
- Is the model with x1 and x2 significantly better than a model with intercept only? That is, do the explanatory variables add significant information about Y?
- Is a model with X1 and X2 significantly better than a model with X1 only?

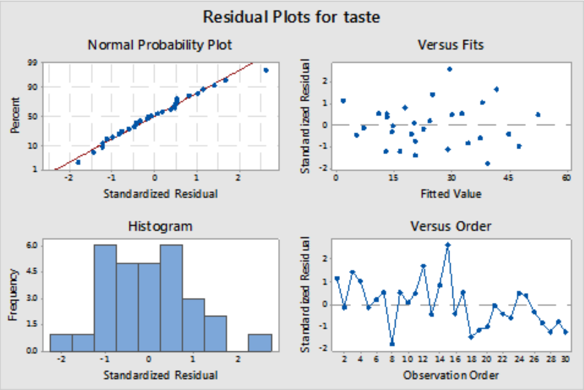
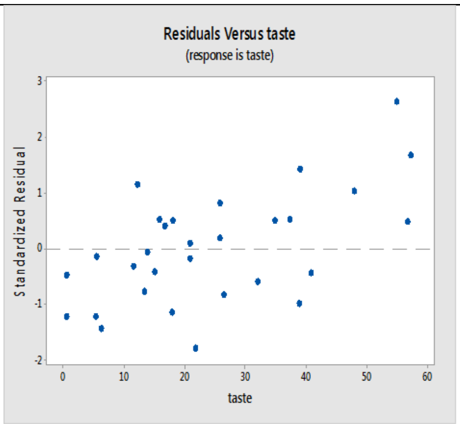
SOLUTIONS- LAB WORK

- Correlations between the explanatory variables and between explanatory variables and the response variable, graph scatter plots.



When doing regression, save residuals, and fitted values.

- Model equation: `taste = -28.9 + 0.33 acetic + 3.91 h2s + 19.67 lactic`
- Discuss diagnostic plots,

Regression Analysis: taste versus acetic, h2s, lactic

```
Analysis of Variance

Source       DF   Adj SS   Adj MS  F-Value  P-Value
Regression    3  4994.48  1664.83    16.22    0.000
  acetic      1     0.55     0.55     0.01    0.942
  h2s         1  1007.66  1007.66     9.82    0.004
  lactic      1   533.32   533.32     5.20    0.031
Error        26  2668.41   102.63
Total        29  7662.89


Model Summary

      S    R-sq  R-sq(adj)  R-sq(pred)
10.1307  65.18%     61.16%      55.60%
```

```
Coefficients

Term       Coef  SE Coef  T-Value  P-Value   VIF
Constant  -28.9     19.7    -1.46    0.155
acetic     0.33     4.46     0.07    0.942  1.83
h2s        3.91     1.25     3.13    0.004  1.99
lactic    19.67     8.63     2.28    0.031  1.94


Regression Equation

taste = -28.9 + 0.33 acetic + 3.91 h2s + 19.67 lactic


Fits and Diagnostics for Unusual Observations

                            Std
Obs  taste    Fit  Resid  Resid
 15  54.90  29.45  25.45   2.63  R

R  Large residual
```

- Are the slopes significantly different from zero on 5% significance level?

- Estimate of variance of the error term σ^2, that is s^2.

- What is the Pearson correlation coefficient between observed and predicted responses?

- Since it makes sense to estimate the model with two explanatory variables h2s and lactic acid only, do so and check the slopes of the explanatory variables, find the estimate of the variance of the error, and correlation between the predicted and observed y's. Comment on the diagnostic plots.

Regression Analysis: taste versus h2s, lactic

```
Analysis of Variance

Source      DF  Adj SS   Adj MS  F-Value  P-Value
Regression   2  4993.9  2496.96    25.26    0.000
  h2s        1  1193.5  1193.52    12.07    0.002
  lactic     1   617.2   617.18     6.24    0.019
Error       27  2669.0    98.85
Total       29  7662.9


Model Summary

      S    R-sq  R-sq(adj)  R-sq(pred)
9.94236  65.17%     62.59%      59.08%


Coefficients

Term        Coef  SE Coef  T-Value  P-Value   VIF
Constant  -27.59     8.98    -3.07    0.005
h2s         3.95     1.14     3.47    0.002  1.71
lactic     19.89     7.96     2.50    0.019  1.71


Regression Equation

taste = -27.59 + 3.95 h2s + 19.89 lactic


Fits and Diagnostics for Unusual Observations

                            Std
Obs  taste    Fit  Resid  Resid
 15  54.90  29.28  25.62   2.63  R

R  Large residual
```

- Predict the value of Y for h2s=6 and lactic=1.
- Find a 98% confidence interval for the mean value of Y when h2s=6 and lactic=1,
- Find a 98% prediction interval for the mean value of Y when h2s=6 and lactic=1.

Prediction for taste

```
Regression Equation

taste = -27.59 + 3.95 h2s + 19.89 lactic


Variable  Setting
h2s             6
lactic          1


    Fit   SE Fit         98% CI               98% PI
15.9730  3.99686  (6.09011, 25.8559)  (-10.5232, 42.4692)
```

- Is the model with x1 and x2 significantly better than a model with intercept only? That is, do the explanatory variables add significant information about Y?

- Is a model with X1 and X2 significantly better than a model with X1 only?

Regression Analysis: taste versus h2s, lactic

```
Analysis of Variance

Source      DF  Seq SS   Seq MS  F-Value  P-Value
Regression   2  4993.9  2496.96    25.26    0.000
  h2s        1  4376.7  4376.75    44.28    0.000
  lactic     1   617.2   617.18     6.24    0.019
Error       27  2669.0    98.85
Total       29  7662.9
```