

## ST512 – Lab 03

These exercises, which are meant to be completed during the lab session, should be helpful for completing the homework assignment on the same material. Feel free to discuss the problems with your classmates or with the lab assistant.

Consider a study to investigate how people's cheese taste preference depends on the concentration level of certain compounds. The data, available in the file `cheese.txt` on the course website, contains five variables. The response/dependent variable, `taste`, is a subjective taste score (obtained by combining the scores of several tasters) and there are three predictor/explanatory/independent variables:

$$\begin{aligned}\text{acetic} &= \log(\text{acetic acid concentration}) \\ \text{h2s} &= \log(\text{hydrogen sulfide concentration}) \\ \text{lactic} &= \text{lactic acid concentration.}\end{aligned}$$

The variable `case` is just a meaningless label that will not be used in the analysis. This lab walks you through the fitting of a multiple linear regression model.

1. (a) Find pairwise correlations and pairwise scatterplots for all four variables. Is `taste` linearly related to the predictor variables individually?  
(b) Justify this claim: among these three predictor variables, `h2s` has the strongest relationship with the response.  
(c) Fit a simple linear regression model of `taste` versus `h2s`. Find the estimated slope parameter and, based on the plots and output produced, comment on the quality of the model fit and on appropriateness of the model assumptions.
2. Since we have three predictor variables available, we investigate their joint effect on the response. Fit a multiple linear regression model with all three variables.
  - (a) Find the equation of the fitted least-squares regression line. Compare the new estimated slope for `h2s` to that from Problem 1(c) above. Should the two estimates be the same? Explain.
  - (b) Find the F test p-value given in the output. State the null and alternative hypotheses being tested with that p-value, and your conclusion.
  - (c) Give a 95% confidence interval for the slope attached to `h2s` and use this to carry out a test of the null hypothesis that `h2s` and `taste` are not related.
  - (d) Look at the plot of the residuals versus the three predictor variables. Any indication that the assumption of constant error variance is violated?
  - (e) Create a QQ plot of the residual. What does this plot tell you?
3. Get a 90% prediction interval for the actual `taste` value for a new cheese with the following predictor variable values: `acetic` = 10, `h2s` = 5.5, and `lactic` = 4.