02441 Applied Statistics and Statistical Software

Exercise 2D - Cheese

As cheddar cheese matures, a variety of chemical processes take place. The taste of matured cheese is related to the concentration of several chemicals in the final product. In a study of cheddar cheese from the LaTrobe Valley of Victoria, Australia, samples of cheese were analyzed for their chemical composition and were subjected to taste tests. Overall taste scores were obtained by combining the scores from several tasters.

Variable name	Description
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

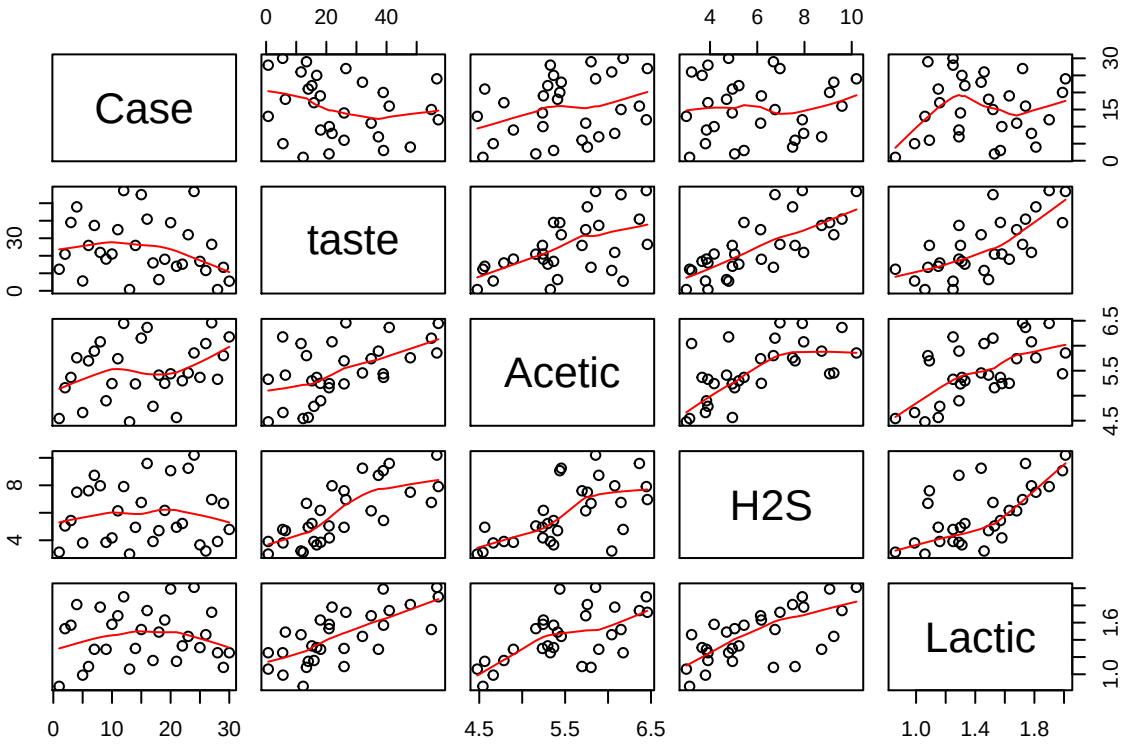
1. Use scatterplots, correlation, and simple regression to examine the relationships among the individual variables.

Start by loading data

```
ch <- read.table("cheese.txt", header = TRUE)
```

Visually investigate the relation between the variables:

```
pairs(ch, panel = panel.smooth)
```




```
library(kableExtra)
kable(cor(ch))
```

	Case	taste	Acetic	H2S	Lactic
Case	1.0000000	-0.2148926	0.2838356	0.0438323	0.0575628
taste	-0.2148926	1.0000000	0.5495393	0.7557523	0.7042362
Acetic	0.2838356	0.5495393	1.0000000	0.6179559	0.6037826
H2S	0.0438323	0.7557523	0.6179559	1.0000000	0.6448123
Lactic	0.0575628	0.7042362	0.6037826	0.6448123	1.0000000

2. Why do you think acetic and h2s has been transformed?

To remove heteroscedasticity regarding the residuals (variance was inhomogeneous). In other words, the transformation corrects for a linear relationship between the independent and the dependent variable.

3. What happens when you run a regression model with all the independent variables in the model?

```
lm1a <- lm(taste~., data = ch)
summary(lm1a)
```

```
##
## Call:
## lm(formula = taste ~ ., data = ch)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -22.3523  -4.9735  -0.5089   4.8531  23.1311
##
```

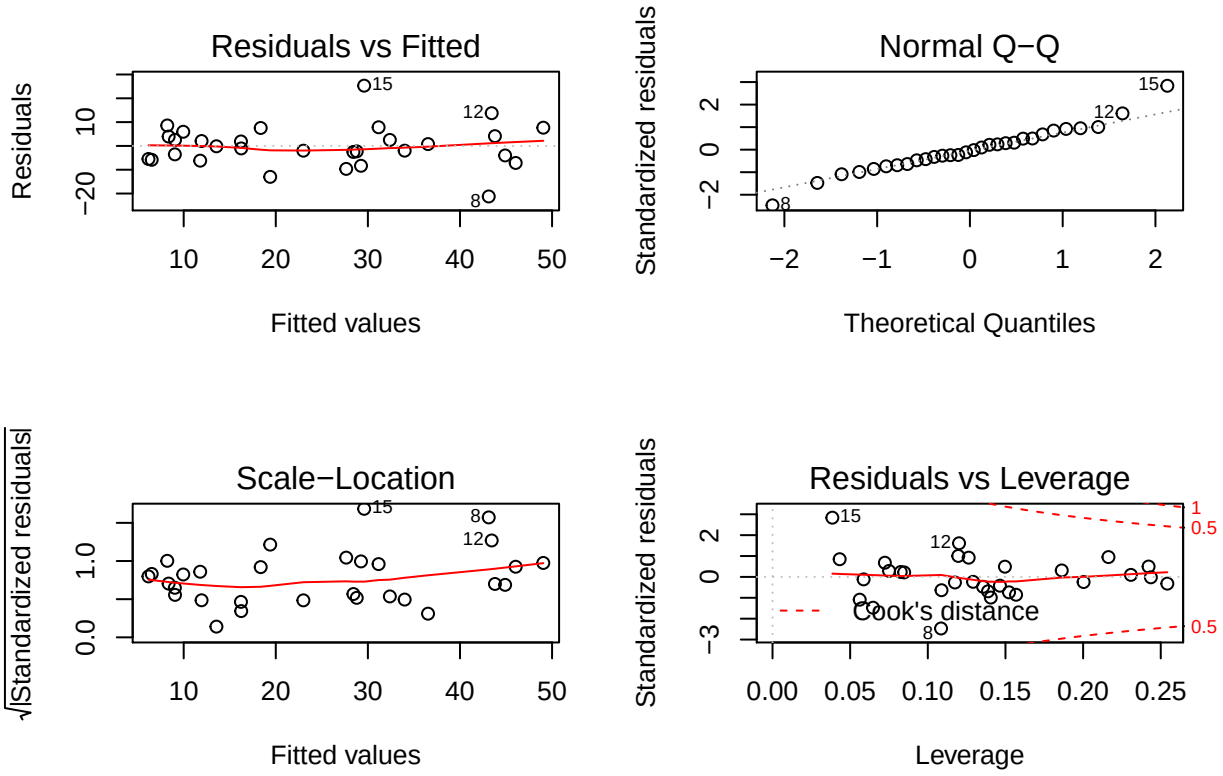
```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -36.6127    17.9845  -2.036  0.05250 .
## Case         -0.5459     0.2043  -2.672  0.01306 *
## Acetic        4.1275     4.2556   0.970  0.34139
## H2S           3.5387     1.1315   3.127  0.00444 **
## Lactic       17.9527     7.7875   2.305  0.02973 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.112 on 25 degrees of freedom
## Multiple R-squared:  0.7291, Adjusted R-squared:  0.6858
## F-statistic: 16.83 on 4 and 25 DF,  p-value: 8.205e-07
```

4. What model would you prefer for prediction?

```
lm1b <- update(lm1a, .~. - Acetic)
summary(lm1b)
```

```
##
## Call:
## lm(formula = taste ~ Case + H2S + Lactic, data = ch)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -21.2230  -5.1078  -0.6005   4.0627  25.3053
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -21.2987     8.6002  -2.477 0.020095 *
## Case         -0.4797     0.1923  -2.494 0.019300 *
## H2S           3.9691     1.0396   3.818 0.000751 ***
## Lactic       20.5850     7.2910   2.823 0.008996 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.101 on 26 degrees of freedom
## Multiple R-squared:  0.719,  Adjusted R-squared:  0.6865
## F-statistic: 22.17 on 3 and 26 DF,  p-value: 2.454e-07
```

```
par(mfrow=c(2,2))
plot(lm1b)
```



The reduced model (without Acetic) would be preferred for prediction modeling. When Acetic is removed the p-values for the remaining x variables get even smaller.

It is surprising that case remains significant, meaning that case number affects taste. This could be due to the fact that some cheeses were produced later in time, which in turn could have affected the taste itself.

5. Predict the 'taste' of a cheese where (log) acetic is 5.3, (log) h2s is 8.0 and lactic is 3.0

```
lm1c <- update(lm1b, .~.-Case)
summary(lm1c)
```

```
##
## Call:
## lm(formula = taste ~ H2S + Lactic, data = ch)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.343  -6.530  -1.164   4.844  25.618
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -27.592      8.982  -3.072  0.00481 **
## H2S            3.946      1.136   3.475  0.00174 **
## Lactic        19.887      7.959   2.499  0.01885 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 9.942 on 27 degrees of freedom
## Multiple R-squared:  0.6517, Adjusted R-squared:  0.6259
## F-statistic: 25.26 on 2 and 27 DF,  p-value: 6.551e-07
```

```
predict(lm1c, newdata = data.frame(Acetic=5.3, H2S=8.0, Lactic=3.0), interval = "prediction")
```

```
##        fit      lwr      upr
## 1 63.63994 32.93147 94.34841
```

6. Could there be some problems with the above prediction?

```
summary(ch)
```

```
##       Case           taste           Acetic           H2S
##  Min.   : 1.00   Min.   : 0.70   Min.   :4.477   Min.   : 2.996
##  1st Qu.: 8.25   1st Qu.:13.55   1st Qu.:5.237   1st Qu.: 3.978
##  Median :15.50   Median :20.95   Median :5.425   Median : 5.329
##  Mean   :15.50   Mean   :24.53   Mean   :5.498   Mean   : 5.942
##  3rd Qu.:22.75   3rd Qu.:36.70   3rd Qu.:5.883   3rd Qu.: 7.575
##  Max.   :30.00   Max.   :57.20   Max.   :6.458   Max.   :10.199
##      Lactic
##  Min.   :0.860
##  1st Qu.:1.250
##  Median :1.450
##  Mean   :1.442
##  3rd Qu.:1.667
##  Max.   :2.010
```

The Lactic acid value for the predicted samples is much higher than for the collected data points yielding a large prediction interval.