STAT 425 Assignment 1

Due Monday, February 8, 11:59pm. Submit through Moodle.

Name: (insert your name here)

Netid: (insert)

Submit your computational work both as an R markdown (*.Rmd) document and as a pdf, along with any files needed to run the code. Embed your answers to each problem in the document below after the question statement. If you have hand-written work, please scan or take pictures of it and include in a pdf file, ideally combined with your pdf output file from R Markdown.

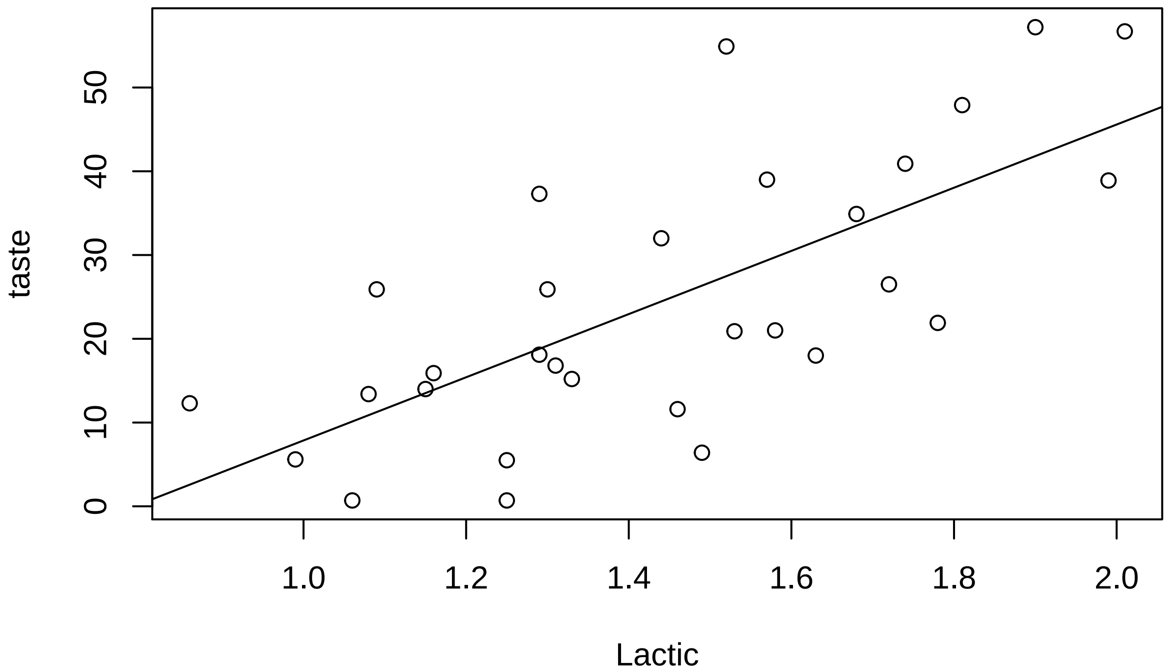
Problem 1

Thirty samples of cheddar cheese were analyzed for their content of acetic acid, hydrogen sulfide and lactic acid. A panel of judges tasted each sample and scored them, and the average taste score for each sample was recorded. The data are available as the data frame 'cheddar' in the **faraway** library. After loading the library enter 'help(cheddar)' for more information.

a) Make a scatter plot of 'taste' versus 'Lactic' and include the least squares regression line on the graph. Comment on whether the graph appears consistent with data that follow a linear model.

Answer:

```
library(faraway)
attach(cheddar)
plot(Lactic, taste) +
  abline(lm(taste ~ Lactic))
```




```
## integer(0)
```

```r
#detach(cheddar)
```

The data points follow a fairly strong linear model.

b) Obtain and display the summary of the least square fitted model, including coefficient estimates, standard errors, t-values and p-values. Is there is a statistically significant association between lactic acid content and the average taste score, using a significance level of $\alpha = 0.05$? Explain based on your results, making clear what information from the results you are using.

```r
fit = lm(taste ~ Lactic)

#summary
summary = summary(fit)

#coefficient estimates, standard errors, t-values, and p-values.
summary$coefficients
```

```
##               Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) -29.85883  10.582319 -2.821577 8.690703e-03
## Lactic       37.71995   7.186396  5.248799 1.405117e-05
```

With a p-value of 1.405e-05, we have enough evidence to conclude at the .05 significance level that there is a statistically significant association between lactic acid content and average taste score. We can reject H0 and conclude the linear coefficient is not 0.

Answer:

c) In R, the 'cor' function can compute the sample correlation coefficient between two variables in a data set. Compute the **squared** correlation between 'taste' and 'Lactic'. Verify that this is numerically equal to R^2 for the model. (Note: to refer to a variable within a data frame use the dataframe$variable syntax.)

```
cor(taste, Lactic)^2
```

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

```
summary$r.squared
```

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

```
cor(taste, Lactic)^2 - summary$r.squared
```

```
## [1] 5.551115e-17
```

Answer:

d) Compute a 95% confidence interval for the coefficient of 'Lactic' in the model.

```
confint(fit, 'Lactic', level = .95)
```

```
##            2.5 %   97.5 %
## Lactic 22.99928 52.44061
```

Answer:

e) Compute a 95% confidence interval for the mean taste value expected for a cheddar cheese sample with lactic acid concentration of 2.0.

Answer:

```
fit2 = lm(taste ~ Lactic)
predict(fit, data.frame(Lactic = 2.0), interval = "confidence",level = .95)
```

```
##        fit      lwr      upr
## 1 45.58106 36.26624 54.89588
```

Problem 2

The simple regression through the origin model has the form

$$y_i = \beta_1 x_i + e_i, \quad i = 1, 2, \ldots, n,$$

where the standard assumptions are that $E(e_i) = 0$, $\text{var}(e_i) = \sigma^2$, and $\text{cov}(e_i, e_j) = 0$ if $i \neq j$. The least squares estimate of β_1 minimizes the residual sum of squares,

$$RSS(\beta_1) = \sum_{i=1}^{n} (y_i - \beta_1 x_i)^2$$

as a function of β_1.

a) Take the derivative of $RSS(\beta_1)$ with respect to β_1, set the derivative to zero. Solve the resulting equation algebraically to obtain the formula for the estimate, $\hat{\beta}_1$.

Answer: Taking the partial derivative of $RSS(\beta_1)$ with respect to β_1 yields

$$\frac{\partial}{\partial \beta_1} RSS(\beta_1) = \frac{\partial}{\partial \beta_1} \sum_{i=1}^{n} (y_i - \beta_1 x_i)^2 = 2 \sum_{i=1}^{n} x_i (y_i - \beta_1 x_i).$$

Setting the derivative to zero yields the equation

$$\sum_{i=1}^{n} x_i y_i = \beta_1 \sum_{i=1}^{n} x_i^2,$$

which implies that

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}.$$

b) Use your formula form Part a) to show that $\hat{\beta}_1$ is an unbiased estimator of β_1 under the standard assumptions.

Answer: Note that only y_i is random and x_i is fixed for $i = 1, \ldots, n$. Taking expectation of $\hat{\beta}_1$ results in

$$\begin{aligned} E(\hat{\beta}_1) = \frac{\sum_{i=1}^{n} x_i E(y_i)}{\sum_{i=1}^{n} x_i} &= \frac{\sum_{i=1}^{n} x_i E(\beta_1 x_i + e_i)}{\sum_{i=1}^{n} x_i^2} \\ &= \frac{\beta_1 \cdot \sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n} x_i^2} \qquad (E(e_i) = 0) \\ &= \beta_1, \end{aligned}$$

which shows that $\hat{\beta}_1$ is an unbiased estimator of β_1.

c) Show that

$$\mathrm{var}\left(\hat{\beta}_1\right) = \frac{\sigma^2}{\sum_{i=1}^{n} x_i^2}$$

Answer: Through using ordinary variance operator, we can obtain

$$\mathrm{var}(\hat{\beta}_1) = \mathrm{var}\left(\frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}\right) = \frac{\sum_{i=1}^{n} x_i^2 \cdot \mathrm{var}(y_i)}{\left(\sum_{i=1}^{n} x_i^2\right)^2},$$

where we used the fact that $\mathrm{cov}(e_i, e_j) = 0$ for $i \neq j$ and $\mathrm{cov}(y_i, y_j) = \mathrm{cov}(e_i, e_j)$ since x_i is fixed. Because $\mathrm{var}(y_i) = \mathrm{var}(\beta_1 x_i + e_i) = \mathrm{var}(e_i) = \sigma^2$, we can obtain

$$\mathrm{var}(\hat{\beta}_1) = \frac{\sigma^2 \sum_{i=1}^{n} x_i^2}{\left(\sum_{i=1}^{n} x_i^2\right)^2} = \frac{\sigma^2}{\sum_{i=1}^{n} x_i^2},$$

which finishes the proof.

d) Under the standard assumptions find $E(y_1)$ and $\text{var}(y_1)$.

Answer: Since $E(e_1) = 0$ and $E(e_1) = \sigma^2$,

$$E(y_1) = E(\beta_1 x_1 + e_1) = \beta_1 x_1 + E(e_1) = \beta_1 x_1$$
$$\text{var}(y_1) = \text{var}(\beta_1 x_1 + e_1) = \text{var}(e_1) = \sigma^2.$$

e) Under the standard assumptions find expressions for $E(\hat{y}_1)$ and $\text{var}(\hat{y}_1)$.

Answer: Since $\hat{y}_1 = \hat{\beta}_1 x_1$,

$$E(\hat{y}_1) = E(\hat{\beta}_1 x_1) = x_1 E(\hat{\beta}_1) = \beta_1 x_1,$$

where we used the fact that $\hat{\beta}_1$ is unbiased from b). Next,

$$\text{var}(\hat{y}_1) = \text{var}(\hat{\beta}_1 x_1) = x_1^2 \cdot \text{var}(\hat{\beta}_1) = \frac{\sigma^2 x_1^2}{\sum_{i=1}^n x_i^2},$$

where we use the expression for $\text{var}(\hat{\beta}_1)$ from c).

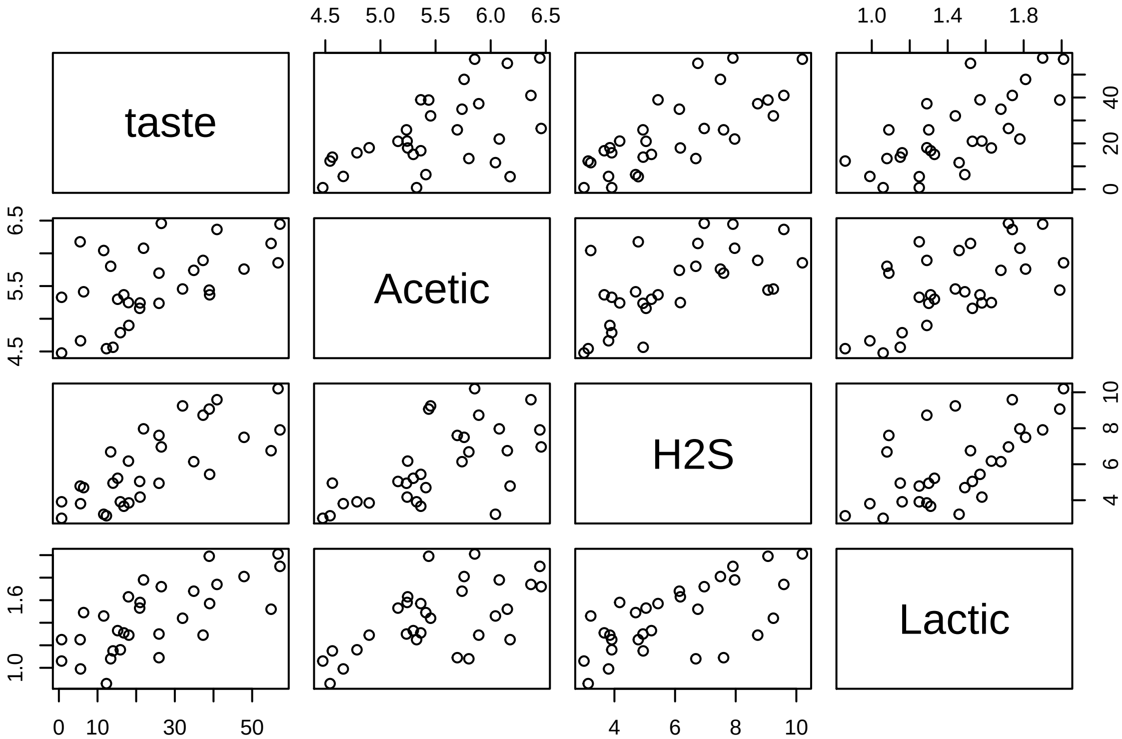
Problem 3:

This problem refers again to the 'cheddar' data described in Problem 1.

a) Make a 'pairs' plot of the data, i.e., a matrix of all the pairwise scatter plots between variables.

Answer:

```
library("faraway")
plot(cheddar)
```



b) Fit a multiple linear regression model with 'taste' as the response and the three chemical constituent concentrations as the predictors. Display a summary of your fitted model. Note: the 'lm' function can fit a multiple linear regression model using a formula of the form 'y ~ x1 + x2 + ... + xp'.

Answer:

```
model = lm(taste~Acetic + H2S + Lactic, data = cheddar)
summary(model)
```

```
##
## Call:
## lm(formula = taste ~ Acetic + H2S + Lactic, data = cheddar)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.390  -6.612  -1.009   4.908  25.449
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -28.8768    19.7354  -1.463  0.15540
## Acetic        0.3277     4.4598   0.073  0.94198
## H2S           3.9118     1.2484   3.133  0.00425 **
## Lactic       19.6705     8.6291   2.280  0.03108 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 10.13 on 26 degrees of freedom
## Multiple R-squared:  0.6518, Adjusted R-squared:  0.6116
## F-statistic: 16.22 on 3 and 26 DF,  p-value: 3.81e-06
```

c) Report the values of the regression coefficients for associated with the predictors.

Answer:

```
coef(model)
```

```
## (Intercept)      Acetic         H2S      Lactic
## -28.8767696   0.3277413   3.9118411  19.6705434
```

d) Which of the predictor variables have statistically significant coefficients, rejecting the null hypothesis that the coefficient is zero, at the 5% level of significance? Explain.

Answer: H2S and Lactic. By checking the summary of fitted model, we can see these two are the variable with p value smaller then 5%.

e) Compute an estimate of the average taste score for a cheddar sample with Acetic= 5.5, H2S=5.0, Lactic=1.5.

Answer:

```
predict(model, newdata= data.frame(Acetic= 5.5, H2S=5.0, Lactic=1.5))
```

```
##        1
## 21.99083
```