Question 1

a) Consider first a full regression model with all the predictors used to explain the Taste response.

i) [4 marks] Write down the full statistical multiple regression model for quality explained by the other predictors. Carefully define all necessary parameters in your answer.

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \varepsilon_i \text{ where } \varepsilon_i \sim N(0, \sigma^2)$$

- Y_i : Taste Response
- X_{ji} : The value of predictor Alcohol, Esters, Lactones and PhenComp for j = 1, 2, 3, 4 respectively.
- ε_i : The unexplained variation

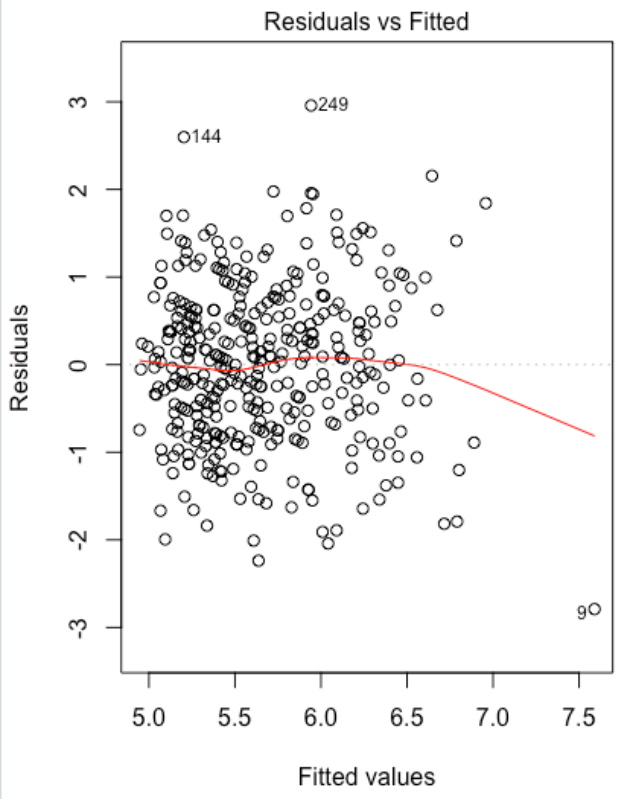
ii) [4 marks] Fit and validate the full regression model.

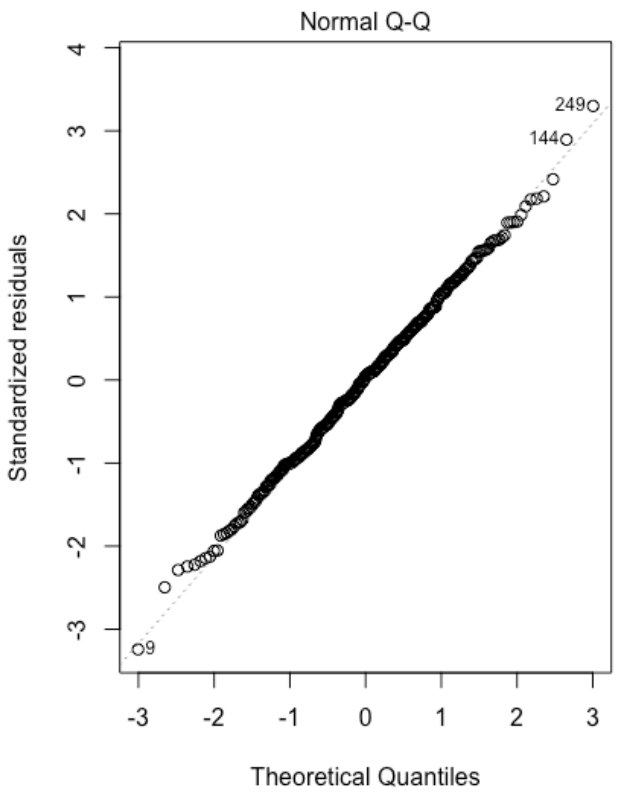
```
> whiskychem = read.csv("whiskychem.csv", header = TRUE)
> whiskychem1 = lm(Taste ~ Alcohol + Esters + Lactones + PhenComp , data = whiskychem)
> summary(whiskychem1)
```

```
Call:
lm(formula = Taste ~ Alcohol + Esters + Lactones + PhenComp,
    data = whiskychem)

Residuals:
     Min       1Q   Median       3Q      Max
-2.78994 -0.66221  0.03959  0.59085  2.95739

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.899704   1.469076   2.655  0.00828 **
Alcohol      0.120846   0.014538   8.313 1.79e-15 ***
Esters      -0.009891   0.034427  -0.287  0.77404
Lactones     0.824616   0.266191   3.098  0.00210 **
PhenComp    -0.754027   0.387226  -1.947  0.05226 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8999 on 371 degrees of freedom
Multiple R-squared:  0.1963,    Adjusted R-squared:  0.1877
F-statistic: 22.66 on 4 and 371 DF,  p-value: < 2.2e-16
```

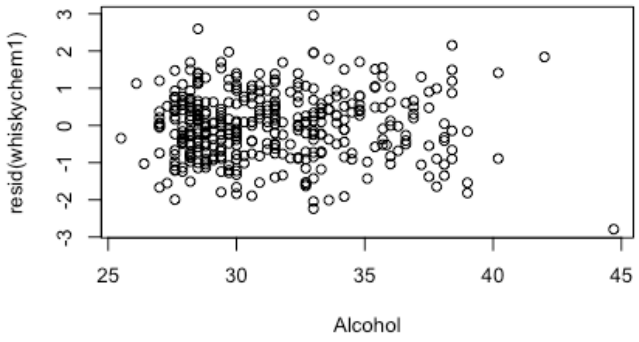
$$\widehat{Taste} = 3.899704 + (0.120846)Alcohol + (-0.009891)Esters + (0.824616)Lactones + (-0.754027)PhenComp$$

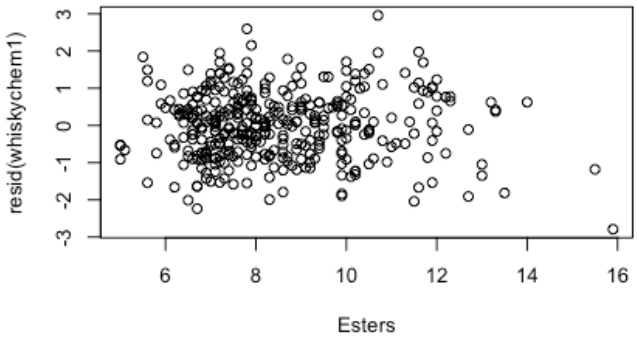
```
> par(mfrow = c(1, 2))
> plot(whiskychem1, which = 1:2)
```

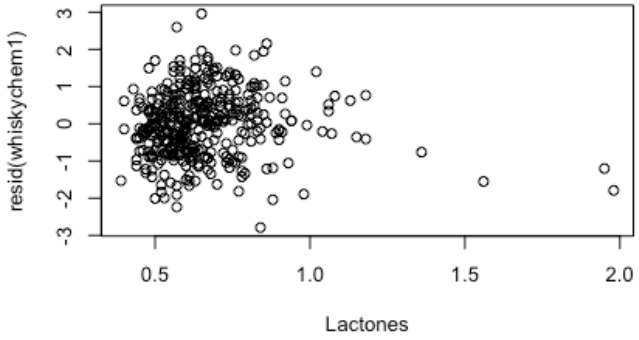


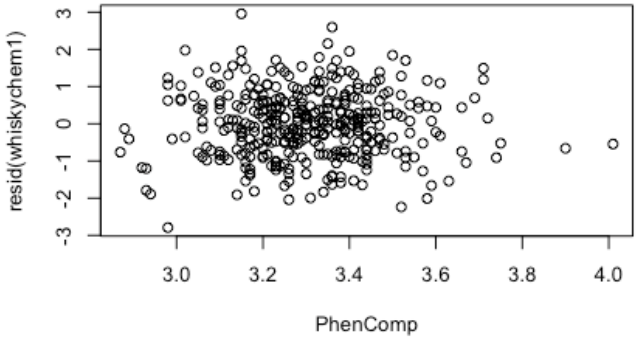



```
> par(mfrow = c(2, 2))
> plot(resid(whiskychem1) ~ Alcohol + Esters + Lactones + PhenComp, data = whiskychem)
```









- The Normal Q-Q plot of residuals shows a reasonable linear relationship with very little deviation, suggesting errors normally distributed.
- The Residuals vs. Fitted plot does not show a discernible pattern
- The Residuals vs. Predictor plots do not show an obvious pattern. Therefore, linear model seems adequate.

iii) [3 marks] Compute a 95% confidence interval for the regression coefficient (slope) for the Esters variable. Explain what the confidence interval represents in the context of the data.

$$CI: -0.009891 \pm 1.648971 \times 0.034427 = (-0.57758225, 0.04687812)$$

- $estimate = \widehat{\beta_{Esters}} = -0.009891$
- $s.e.(estimate) = 0.034427$
- $quantile = t_{(371,0.05)} = 1.648971$
- 95% confident that for each unit of increase in the level of esters when alcohol, lactones and phenolic compounds are fixed, the mean change in taste will increase between -0.57758225 and 0.04687812 units.

b) Consider now a reduced model that can explain the taste response with a reduced set of predictors.

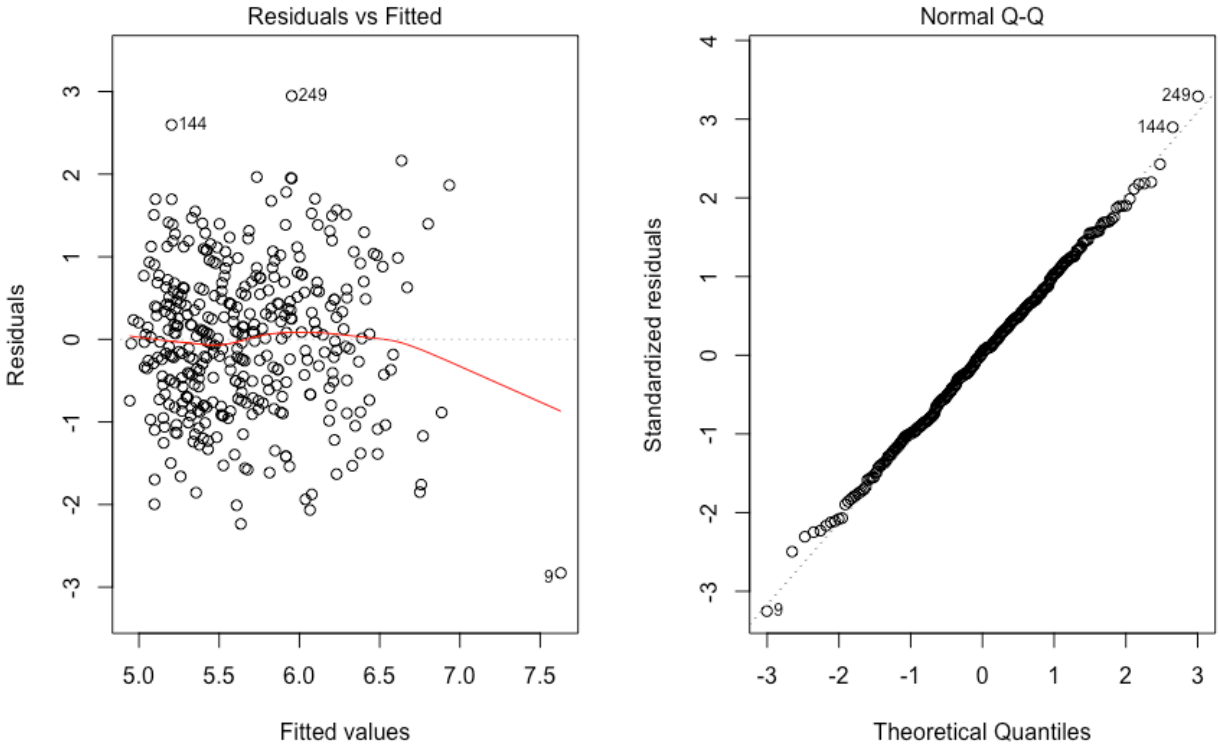
i) [2 marks] Using the appropriate backward model selection method discussed in the course, determine the best regression model for the data.

Remove Esters from the model, as it is not significant and has the largest P-value.

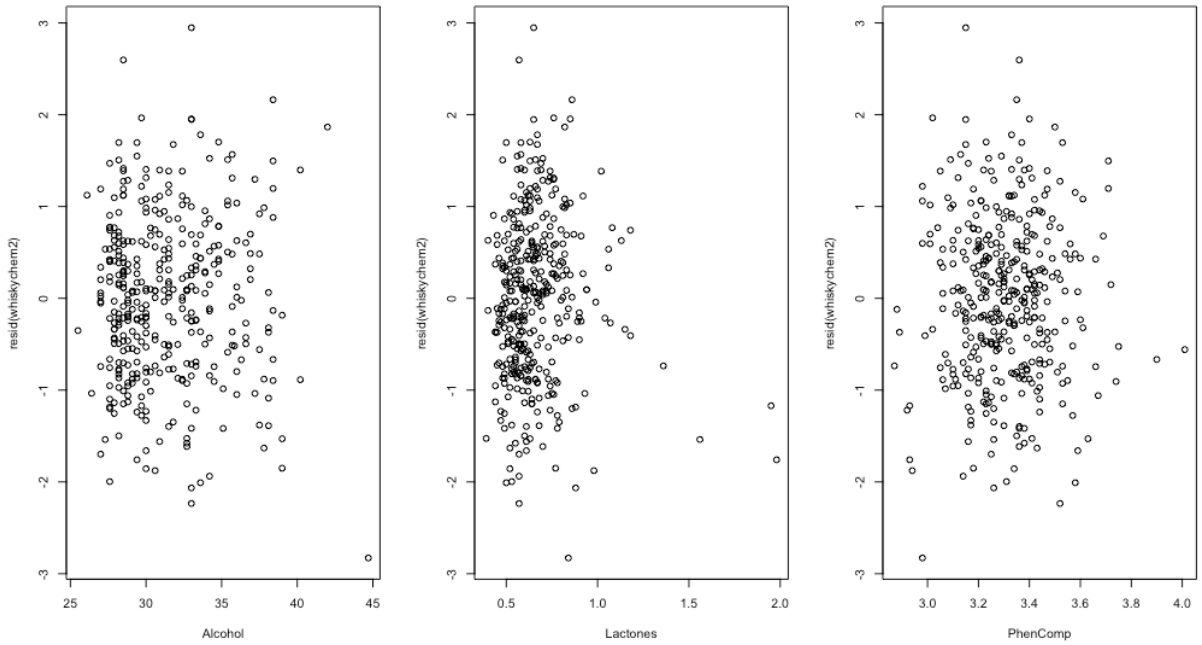
```
> whiskychem2 = lm(Taste ~ Alcohol + Lactones + PhenComp, data = whiskychem)
> whiskychem2 = update(whiskychem1, . ~ . - Esters)
> summary(whiskychem2)
```

```
Call:
lm(formula = Taste ~ Alcohol + Lactones + PhenComp, data = whiskychem)

Residuals:
     Min       1Q   Median       3Q      Max
-2.82854 -0.66383  0.03279  0.59275  2.94784

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.59963    1.03184   3.489 0.000544 ***
Alcohol      0.12010    0.01428   8.408    9e-16 ***
Lactones     0.81899    0.26514   3.089 0.002160 **
PhenComp    -0.68031    0.28966  -2.349 0.019362 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8988 on 372 degrees of freedom
Multiple R-squared:  0.1962,    Adjusted R-squared:  0.1897
F-statistic: 30.26 on 3 and 372 DF,  p-value: < 2.2e-16
```

```
> par(mfrow = c(1, 2))
> plot(whiskychem2, which = 1:2)
```




```
> par(mfrow = c(1, 3))
> plot(resid(whiskychem2) ~ Alcohol + Lactones + PhenComp, data = whiskychem)
```



- The Normal Q-Q plot of residuals shows a reasonable linear relationship with very little deviation, suggesting errors normally distributed.
- The Residuals vs. Fitted plot does not show a discernable pattern
- The Residuals vs. Predictor plots do not show an obvious pattern. Therefore, linear model also seems adequate.

ii) [4 marks] Write down the final model and interpret it in the context of the data.

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \varepsilon_i \text{ where } \varepsilon_i \sim N(0, \sigma^2)$$

- Y_i : Taste Response
- X_{ji} : The value of predictor Alcohol, Lactones and PhenComp for j = 1, 2, 3 respectively.
- ε_i : The unexplained variation

$$\widehat{Taste} = 3.59963 + (0.12010)Alcohol + (0.81899)Lactones + (-0.68031)PhenComp$$

c**) For both the full model considered in a) and the reduced model in b), answer the questions below.**

i) [2 marks] State the R^2 and explain what it means in the context of the data.
The $ R^2 in the full model is 0.1963. The R^2 indicates how close the data are to the fitted regression line i.e. explained variation over total variation. The R^2 in the full model indicates that the model explains roughly 19.63%% of the variability of the Taste response data around its mean. The R^2 in the reduced model is 0.1962. The R^2 in the reduced model indicates that the model explains roughly 19.62% of the variability of the Taste response data around its mean.

iii) [2 marks] Explain why the adjusted R^2 should be reported over the R^2 for assessing the goodness of fit.
When a predictor is added to a model, the R^2 increases. Therefore, a model containing more predictors may give the appearance of a better fit simply because it contains more predictors. The adjusted R^2, which is always lower than the R^2 is modified for the number of predictors a model contains. When a predictor is added to the model, the adjusted R^2 only increases if that new predictor improves the model, not because there are now more predictors. The loss in R^2 from the full model to the reduced model is only 0.01%, suggesting the reduced model is better.

Question 2

a) [1 mark] State the appropriate hypotheses.

$$H_0: C_1 = 0;\ C_1 \neq 0$$
$$H_0: C_2 = 0;\ H_1: C_2 \neq 0$$

b) [1 mark] Calculate the observed value of the (raw) contrast.

$$C_{1obs} = \frac{\mu_A + \mu_B + \mu_C}{3} - \mu_D = \frac{49.99789 + 48.09186 + 49.826687}{3} - 50.0932 = -0.787721$$

$$C_{2obs} = \frac{\mu_A + \mu_B}{2} - \frac{\mu_C + \mu_D}{2} = \frac{49.99789 + 48.09186}{2} - \frac{49.826687 - 50.0932}{2} = -0.9150685$$

c) [2 marks] Calculate the standard error of the contrast.

$$s.e.(c_1) = s_p\sqrt{\frac{k_A^2}{n_A} + \frac{k_B^2}{n_B} + \frac{k_C^2}{n_C} + \frac{k_D^2}{n_D}} = 2.30338731\sqrt{\frac{1^2}{31} + \frac{1^2}{28} + \frac{1^2}{27} + \frac{-1^2}{40}} = 0.15956583$$

$$s.e.(c_2) = s_p\sqrt{\frac{k_A^2}{n_A} + \frac{k_B^2}{n_B} + \frac{k_C^2}{n_C} + \frac{k_D^2}{n_D}} = 2.30338731\sqrt{\frac{1^2}{31} + \frac{1^2}{28} + \frac{-1^2}{27} + \frac{1^2}{40}} = 0.15956583$$

d) [1 mark] State the null distribution of the test statistic.

$$t_{obs} = \frac{c}{s.e.(c)} \sim t_{N-g}$$

$$For\ C_1: t_{obs} = \frac{-0.787721}{0.15956583} = -4.936652164$$

$$For\ C_2: t_{obs} = \frac{-0.9150685}{0.15956583} = -5.7347397$$

e) [2 marks] Compute or give the best approximate bound on the P-Value.

For C_1: > pt(-4.936652164, 115)

1.356802e-06

For C_2: > pt(-5.7347397, 115)

4.009004e-08

f) [3 marks] Draw your overall conclusions (both statistical and contextual).

Both p-values are less than α = 0.05, therefore reject the null hypothesis for both cases.