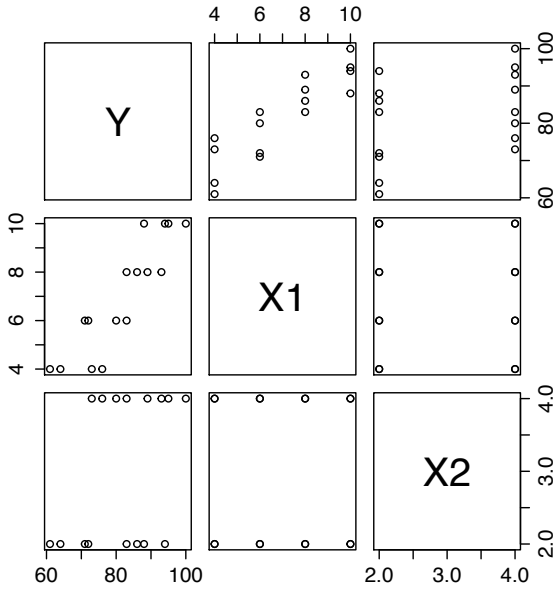
Problem 6.5
Part a

Scatter Plot



Correlation Matrix

	Y	X1	X2
Y	1.0000000	0.8923929	0.3945807
X1	0.8923929	1.0000000	0.0000000
X2	0.3945807	0.0000000	1.0000000

- Both scatter plot and correlation matrix indicate the relationships between X1, X2 and Y, which represent the moisture content, sweetness and the degree of brand liking, respectively. A strong positive association exists between X1 and Y, which is supported with evidences from both scatterplot and correlation matrix. The correlation of X1 and Y is 0.8923929, which is close to 1. Consistently, there is a positive upwards trend shown in the scatter plot of X1and Y. In the similar way, X2 has a relative weak positive relationship with Y with r =0.3945<0.5. However, there is not very intuitive and clear trend shown in the scatter plot between X2 and Y.
- The correlation between X1 and X2 is 0 and thus I conclude X1 and X2 are independent. There is no interaction between X1 and X2.

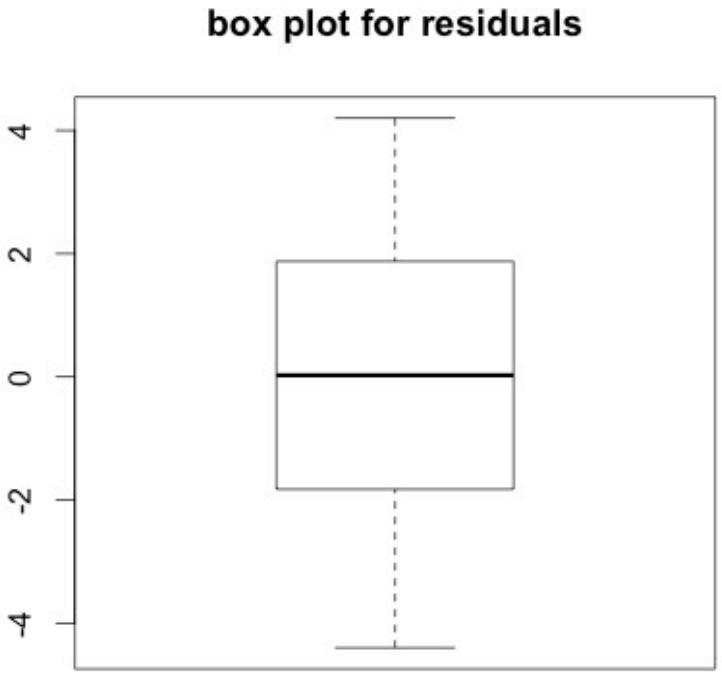
Part b

```
Call:
lm(formula = Y ~ X1 + X2, data = data)

Residuals:
   Min     1Q Median     3Q    Max
-4.400 -1.762  0.025  1.587  4.200

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  37.6500     2.9961  12.566 1.20e-08 ***
X1            4.4250     0.3011  14.695 1.78e-09 ***
X2            4.3750     0.6733   6.498 2.01e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.693 on 13 degrees of freedom
Multiple R-squared:  0.9521,    Adjusted R-squared:  0.9447
F-statistic: 129.1 on 2 and 13 DF,  p-value: 2.658e-09
```

The estimated regression function is Yi_hat = 37.6550+4.4250X1+4.3750X2
The estimator b1 measures the change in the mean response E [Y] per unit increase in X1 when X2 is hold constant. In this example, b1 is equal to 4.4250, which can be interpreted that the expectation of the degree of brand likeliness increases 4.4250 with per unit increasing in its moisture content when its sweetness is hold constant.

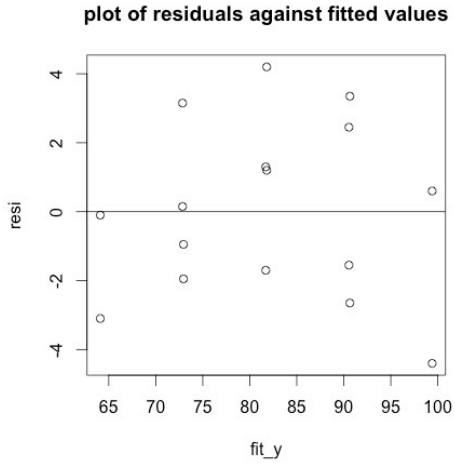
Part c

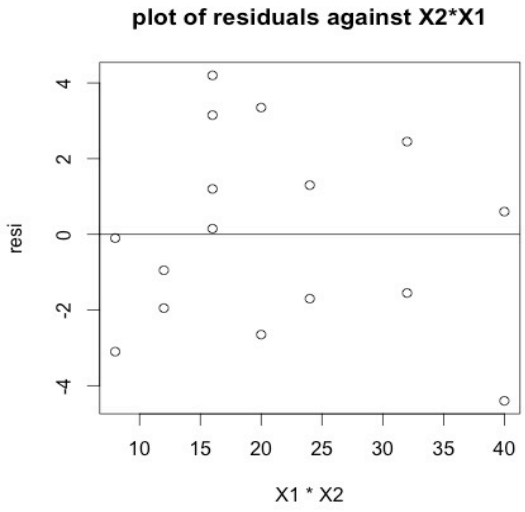
```
> resi
    1     2     3     4     5     6     7     8     9    10    11    12    13    14    15    16
-0.10  0.15 -3.10  3.15 -0.95 -1.70 -1.95  1.30  1.20 -1.55  4.20  2.45 -2.65 -4.40  3.35  0.60
>
```

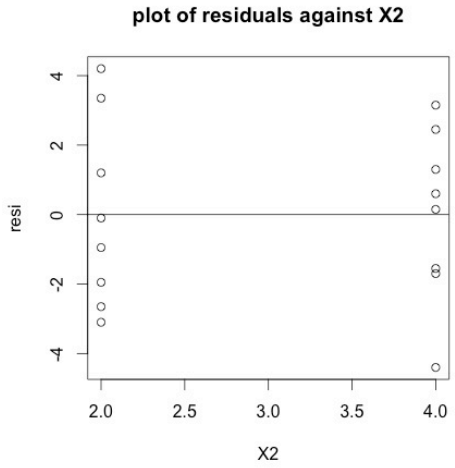


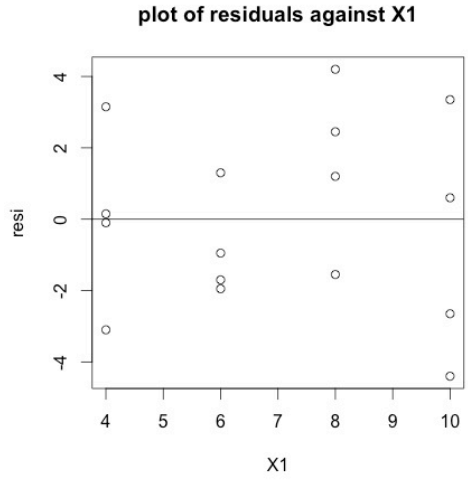
- The boxplot of the residuals shows that the residuals are approximately normally distributed with center around zero.
- The distribution is approximately symmetric with similar distance to both sides
- There is no obvious outlier shown in the boxplot figure.

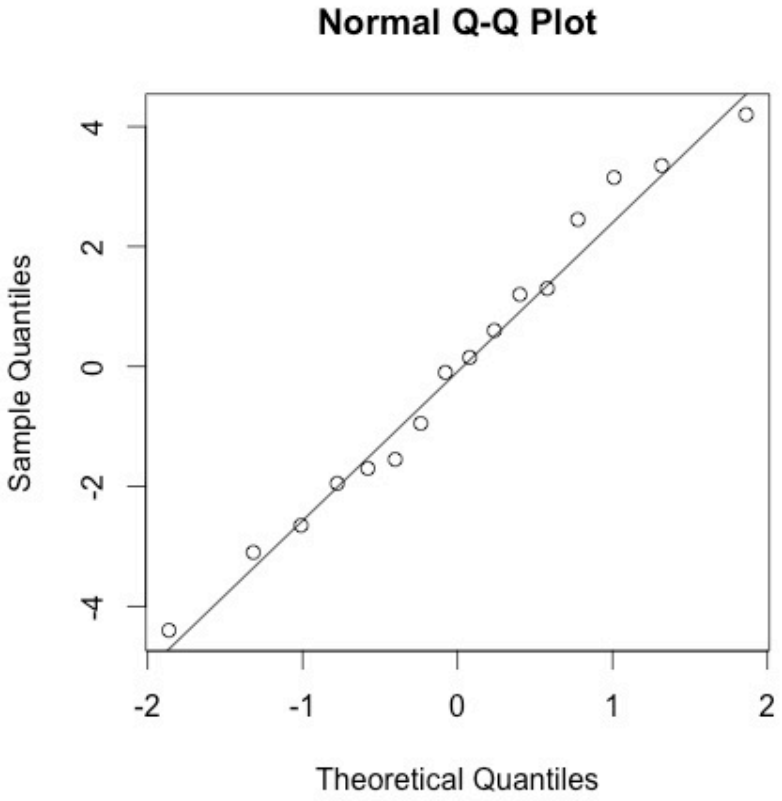
Part d


plot of residuals against fitted values
resi
fit_y
65 70 75 80 85 90 95 100
-4 -2 0 2 4


plot of residuals against X2*X1
resi
X1 * X2
10 15 20 25 30 35 40
-4 -2 0 2 4


plot of residuals against X2
resi
X2
2.0 2.5 3.0 3.5 4.0
-4 -2 0 2 4


plot of residuals against X1
resi
X1
4 5 6 7 8 9 10
-4 -2 0 2 4



- The residuals fall within a horizontal bond centered around zero, displaying no systematic tendencies to be positive or negative. These are prototype situations of the residual plots against predictors when a linear regression model is appropriate. The residual plot against fitted values supports the same conclusion.
- The residual plots does not illustrate any non-constant error variance by not showing obvious trend
- There is no obvious outlier
- The residuals do not vary with different level of X1*X2 and thus adding this new variable will not provide important additional descriptive and predictive power to the original model. No need to introduce the new variable X1*X2.
- The Normal QQ plot also indicates that there is not strong departure from the normality of error term

Code:

```
#6.5
# part a
data<-read.table(file="http://www.stat.lsu.edu/exstweb/statlab/datasets/KNNLData/CH06PR05.txt")
names(data)<-c("Y","X1","X2")
attach(data)
cor(data)
plot(data)

# part b
fit = lm(Y~X1+X2, data=data)
summary(fit)

# part c
resi = fit$resi
resi
boxplot(resi)
title("box plot for residuals")

# part d
fit_y<-fit$fitted.values
plot(fit_y,resi)
abline(h=0)
title("plot of residuals against fitted values")

plot(X1,resi)
abline(h=0)
title("plot of residuals against X1")

plot(X2,resi)
abline(h=0)
title("plot of residuals against X2")


plot(X1*X2,resi)
abline(h=0)
title("plot of residuals against X2*X1")

qqnorm(resi)
qqline(resi)
```

Problem 6.6

Part a

Ho: $\beta 1 = 0 \; and \; \beta 2 = 0$
H1: $not \; both \; \beta 1 \; and \; \beta 2 \; equal \; to \; 0$

Reduced Model: Yi= $\beta 0 + \varepsilon i$
Full Model: Yi= $\beta 0 + \beta 1 X1 + \beta 2 X2 + \varepsilon i$

$F^* = \frac{MSR}{MSE}$

Decision Rule:
If F*<=F (1-a; p-1; n-p), conclude Ho
If F*>=F (1-a; p-1; n-p), conclude H1
F* = 129.1(from figure in problem 6.5 part b), which is larger than F (0.99,2,13) = 6.700965. Therefore I conclude H1. The result implied that not both $\beta 1$ and $\beta 2$ equal to zero. There is a regression relation between X1, X2 and Y.

Part b

P-value = 2.589*10^(-9), which is approximately zero. (Check Figure in Problem 6.5 Part b)

Part c

B=t(1-0.01/4;16-3) = 3.372468
b1 = 4.4250
s{b1} = 0.3011
b2 = 4.3750
s{b2} = 0.6733

$$4.4250 - 3.372468 * 0.3011 \leq \beta 1 \leq 4.4250 + 3.372468 * 0.3011$$
$$3.40955 \leq \beta 1 \leq 5.44045$$
$$4.3750 - 3.372468 * 0.6733 \leq \beta 2 \leq 4.3750 + 3.372468 * 0.6733$$
$$2.104317 \leq \beta 2 \leq 6.645683$$

With family confidence coefficient 0.99, $\beta 1$ is between 3.40955 *and* 5.44045 and $\beta 2$ is between 2.104317 and 6.645683. 99 percent of families of estimates $\beta 1 and \; \beta 2$ are both correct when repeated samples are selected and 99% confidence intervals for the estimates $\beta 1 and \; \beta 2$ are calculated for each sample.

Problem 6.7

Part a

R^2 = 0.9521(See Problem 6.5 Part b)

Interpretation: 95.21% of total variation is reduced when using of predicators X1 and X2. It is very close to 1, which implies that there is a strong degree of linear association between X1, X2 and Y.

Part b

Yes. R^2 is still equal to 0.9521

$R^2 = \frac{SSR}{SSTO}$. SSTO doesn't change since Yi and $\bar{Y}$ are fixed. If the fitted values in both settings are the same, then SSR doesn't change, neither. Then R^2 is the same. So in the simple linear model, if we choose the same fitted values, the normal equation for simple linear model is also satisfied

Problem 6.8

```
> newx = data.frame(X1 = 5, X2 = 4)
> predict.lm(fit, newx, interval="confidence",level=.99)
     fit      lwr      upr
1 77.275 73.88111 80.66889
> predict.lm(fit, newx, interval="prediction",level=.99)
     fit      lwr      upr
1 77.275 68.48077 86.06923
>
```

Part a

$73.880 \le E\{Y_h\} \le 80.670$

Interpretation: with 0.99 confidence coefficient, the mean value of the degree of brand likeliness Y when X1 =5, X2=4 is somewhere between 73.880 and 80.670.

Part b

$68.483 \le Y_{h(new)} \le 86.067$

Interpretation: with 0.99 confidence coefficient, we predict that the degree of brand likeliness Y for a product with X1=5, X2=4 will be somewhere between 68.483 and 86.067.