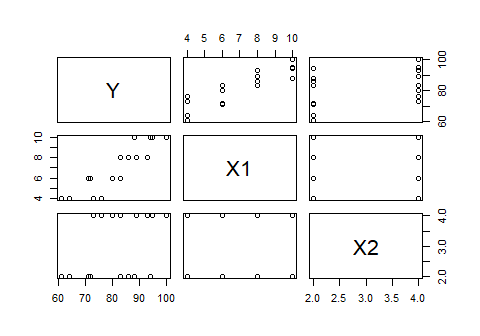
Chapter 6 Problems 6.5-6.8, Brand Preference

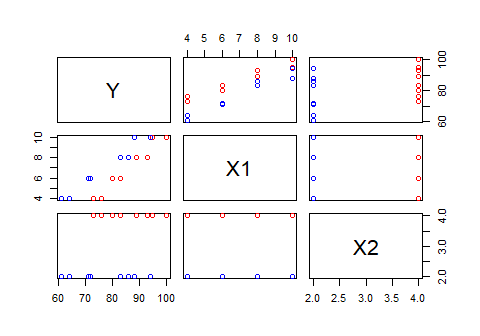
### Michael Streyle

A small-scale experimental study was conducted of the relation of degree of brand liking (Y) vs. moisture content (X1) and sweetness (X2) of a product. The scatterplot matrix and correlation matrices appear below.

mydata <- read.table(file.choose(),header=F,col.names=c("Y", "X1", "X2"))  
  
x1name = "Moisture Content"  
x2name = "Sweetness"  
yname = "Brand Preference"   
  
attach(mydata)  
  
plot.new()  
plot (mydata)



plot (mydata, col=ifelse(X2==2, 'blue', 'red'))



cor (mydata)

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

Summarize the scatterplot matrix and correlation matrix... The scatterplot matrix shows the bivariate relationships for each predictor and response variable. The noteworthy relationships seems to be between Y and X1 (Brand Preference and Moisture Content). The correlation matrix shows the correlation coefficients for each bivariate relationship. The noteworthy relationships from the correlation matrix seem to also be for Y and X1 with a correlation coefficient of 0.8923929. Both of these matrices indicate the relationship between sweetness and brand preference is not very noteworthy.

Next, a multiple regression model is fit to the data:

myfit <- lm (Y ~ X1 + X2)  
myfit

##   
## Call:  
## lm(formula = Y ~ X1 + X2)  
##   
## Coefficients:  
## (Intercept) X1 X2   
## 37.650 4.425 4.375

confint(myfit, level=0.95)

## 2.5 % 97.5 %  
## (Intercept) 31.177312 44.122688  
## X1 3.774470 5.075530  
## X2 2.920372 5.829628

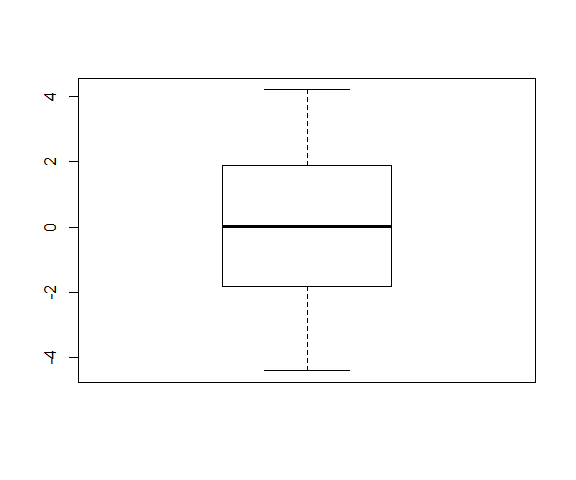
summary(myfit)

##   
## Call:  
## lm(formula = Y ~ X1 + X2)  
##   
## 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 = 37.65 + 4.425X1 + 4.375X2

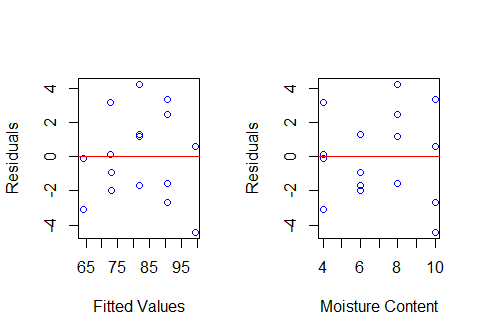
The interpretations of and are:b\_1 is the change in mean responses per unit increase in moisture content when sweetness is held constant. b\_2 is the change in mean responses per unit increase in sweetness when moisture content is held constant.

resid = myfit$residuals  
fitted = myfit$fitted.values  
  
plot.new()  
boxplot (resid)

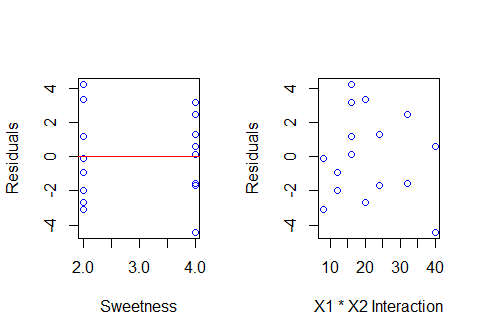


The box plot tells us ... that the data is relatively normally distributed and has constant variance with no obvious outliers.

par(mfrow=c(1,2))  
  
plot (fitted, resid, xlab="Fitted Values", ylab="Residuals", col='blue')  
abline(h=0, col='red')  
  
plot (X1, resid, xlab=x1name, ylab="Residuals", col='blue')  
abline(h=0, col='red')

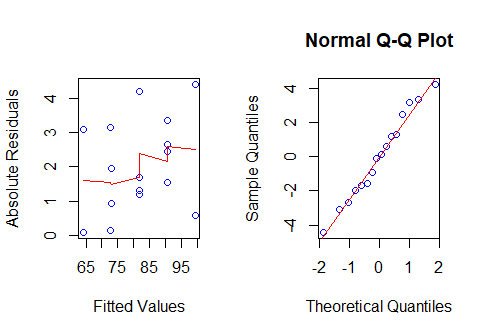


plot (X2, resid, xlab=x2name, ylab="Residuals", col='blue')  
abline(h=0, col='red')  
  
mydata$X1X2 = X1 \* X2  
plot (mydata$X1X2, resid, xlab="X1 \* X2 Interaction", ylab="Residuals", col='blue')



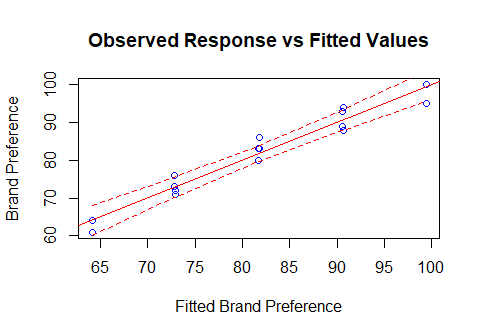
These residual plots above tell us that there seems to be constant variance, a linear relationship, and no obvious outliers in the residuals vs fitted values and also in the residuals vs the moisture content. The plots of residuals vs sweetness and resiuals vs X1\*X2 are not as clear, but show no obvious outliers.

plot.new()  
par(mfrow=c(1,2))  
plot (fitted, abs(resid), xlab="Fitted Values", ylab="Absolute Residuals", col='blue')  
low1 = lowess (fitted, abs(resid))  
lines (low1$x, low1$y, col='red')  
  
qqnorm (resid, col='blue')  
qqline (resid, col='red')



These plots tell us... The Absolute Residuals vs Fitted Values plot tells us that there is not constant variance. The Normal Q-Q Plot shows the data is relatively normally distributed and there are no obvious outliers.

plot.new()  
par(mfrow=c(1,1))  
plot (fitted, Y, xlab=paste("Fitted", yname), ylab=yname, col='blue', main="Observed Response vs Fitted Values")  
abline(0, 1, col='red')  
  
# Working-Hotelling limits (confidence band for the regression line)   
  
error.df = summary(myfit)$df[2]  
num.pred = summary(myfit)$df[1]  
  
fcrit <- qf (0.95, num.pred, error.df)   
#pred <- predict (myfit, mydata [, c("X1", "X2")], se.fit=TRUE)   
wh.fit = lm (Y ~ fitted)  
wh.pred = predict (wh.fit, data.frame (fitted = wh.fit$fitted.values), se.fit=TRUE)   
wh.fitted = wh.pred$fit   
  
w <- sqrt (num.pred \* fcrit)   
marg.err <- w \* wh.pred$se.fit   
  
# Sort the fitted values from smallest to largest  
wh.order = order (wh.fitted)  
wh.ordered = wh.fitted [wh.order]  
  
wh.lower = wh.ordered - marg.err [wh.order]  
wh.upper = wh.ordered + marg.err [wh.order]  
  
lines (wh.ordered, wh.lower, lty=2, col='red')  
lines (wh.ordered, wh.upper, lty=2, col='red')



# Breusch-Pagan test  
library(lmtest)

## Loading required package: zoo

##   
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':  
##   
## as.Date, as.Date.numeric

bptest(myfit, studentize=FALSE)

##   
## Breusch-Pagan test  
##   
## data: myfit  
## BP = 1.0422, df = 2, p-value = 0.5939

The alternatives are H0: y1 = 0, and Ha:y1 != 0 where y1 = 0 corresponds to constancy of error variance. The decision rule states that if p-value > 0.01 then conclude H0, and otherwise conclude Ha. The Breusch-Pagan test indicates that BP = 1.0422 with a p-value of 0.5939 with level of significance set to 0.01. So we conclude H0 which is constancy of variance.

# Lack of fit test  
# How many levels of X1 combined with X2?  
  
aggregate (Y, by=list(X1, X2), FUN='mean')

## Group.1 Group.2 x  
## 1 4 2 62.5  
## 2 6 2 71.5  
## 3 8 2 84.5  
## 4 10 2 91.0  
## 5 4 4 74.5  
## 6 6 4 81.5  
## 7 8 4 91.0  
## 8 10 4 97.5

# Fit the full model:  
full = lm (Y ~ 0 + as.factor(X1) \* as.factor(X2))  
anova (myfit, full)

## Analysis of Variance Table  
##   
## Model 1: Y ~ X1 + X2  
## Model 2: Y ~ 0 + as.factor(X1) \* as.factor(X2)  
## Res.Df RSS Df Sum of Sq F Pr(>F)  
## 1 13 94.3   
## 2 8 57.0 5 37.3 1.047 0.453

The alternatives are H0: The relationship assumed in the model is reasonable (there is no lack of fit) HA: The relationship assumed in the model is not reasonable (there is lack of fit) Using a level of significance of 0.01, the decision rule states that if the p-value > 0.01, then conclude an appropriate fit, and if the p-value < 0.01, then conclude a lack of fit. The lack of fit test indicates a p-value of 0.453 which is greater than 0.01, which indicates a lack of 'lack of fit', or an indication of an appropriate fit.

summary(myfit)

##   
## Call:  
## lm(formula = Y ~ X1 + X2)  
##   
## 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 alternatives for the F test are : H0: = 0 and Ha: != 0. The decision rule states that if the p-value is less than the level of significance (0.01), then it indicates a good regression relation, and if the p-value is greater than the level of significance (0.01), then it indicates a poor regression relation. The p value for the F test is 2.658e-09 which is less than 0.01 which indicates a good regression relation.

confint (myfit, level=0.995)

## 0.25 % 99.75 %  
## (Intercept) 27.545738 47.754262  
## X1 3.409483 5.440517  
## X2 2.104236 6.645764

Joint Bonferroni confidence intervals for and are 3.409483 <= <= 5.440517 and 2.104236 <= <= 6.645764.

Interpretation: With 99% family or joint confidence, the value of is between 3.409483 and 5.440517 and the value of is between 2.104236 and 6.645764. represents the change in mean likeness response per unit increase in moisture content when sweetness is held constant. represents the change in mean likeness response per unit increase in sweetness when moisture content is held constant.

summary(myfit)$r.squared

## [1] 0.952059

The coefficient of multiple determination for the model is: 0.952059 and it measures the proportionate reduction of the total variation in Y associated with the use of the set of X variables (X1 and X2). Even though we have a high coefficient of multiple determination, it does not mean that the fitted model is a useful one.

cor(Y, fitted)^2

## [1] 0.952059

# or:  
summary(lm (Y ~ fitted))$r.squared

## [1] 0.952059

The coefficient of simple determination between Y and Yhat is: 0.952059. Yes, in this example the coefficient of simple determination between Y and Yhat is the same as the coefficient of multiple determination from part (a).

predict (myfit, data.frame (X1=5, X2=4), interval='confidence', level=0.99)

## fit lwr upr  
## 1 77.275 73.88111 80.66889

predict (myfit, data.frame (X1=5, X2=4), interval='prediction', level=0.99)

## fit lwr upr  
## 1 77.275 68.48077 86.06923

99% confidence interval, with interpretation, for the expected response when and :The expected response, with 99% confidence, is between 73.88111 and 80.66889. This means when the moisture content is 5 and the sweetness is 4, the mean expected response would result in a 'likeness' value between 73.88111 and 80.66889 with 99 percent confidence.

99% confidence interval, with interpretation, for a new observation when and : The expected value for a new observation with moisture content of 5 and sweetness of 4, with 99% confidence, will be between 68.48077 and 86.06923 likeness (degree of brand liking).