Assignment 6

Gilbert Watson

Monday, November 4th, 2013

Contents

1 6.5

In a small-scale experimental study of the relation between degree of brand liking (Y) and moisture content (X_1) and sweetness (X_2) of the product, the following results were obtained from the experiment based on a completely randomized design (data are coded):

a) Obtain the scatter plot matrix and the correlation matrix. What information do these diagnostic aids provide here?

b) Fit regession model (6.1) to the data. State the estimated regression function. How is b_1 interpreted here?

c) Obtain the residuals and prepare a box plot of the residuals. What information does this plot provide?

d) Plot the residuals against Y, X_1, X_2 , and X_1 X_2 on separate graphs. Also prepare a normal probability plot. Interpret the plots and summarize your findings.

e) Conduct the Breusch-Pagan test for constancy of the error variance, assuming $log(\sigma^2) = \gamma_0 + \gamma_1 X_{i1} + \gamma_2 X_{i2}$: use $\alpha = .01$. State the alternaltives, decision rule, and conclusion.

1.1 Answer:

a) First let's load the data:

```
> BrandPreference <- read.table(file="6.5.txt",stringsAsFactors=F)
> names(BrandPreference) <- c("Brand.Liking","Moisture.Content","Sweetness")
```

Then let's generate the scatter plot matrix:

```
> require(ggplot2)
> require(reshape2)
> melted <- melt(BrandPreference,id.vars=c("Brand.Liking"))
> g <- ggplot(data=melted,aes(Brand.Liking,value))
> g + geom_point() + facet_wrap(~variable)
```
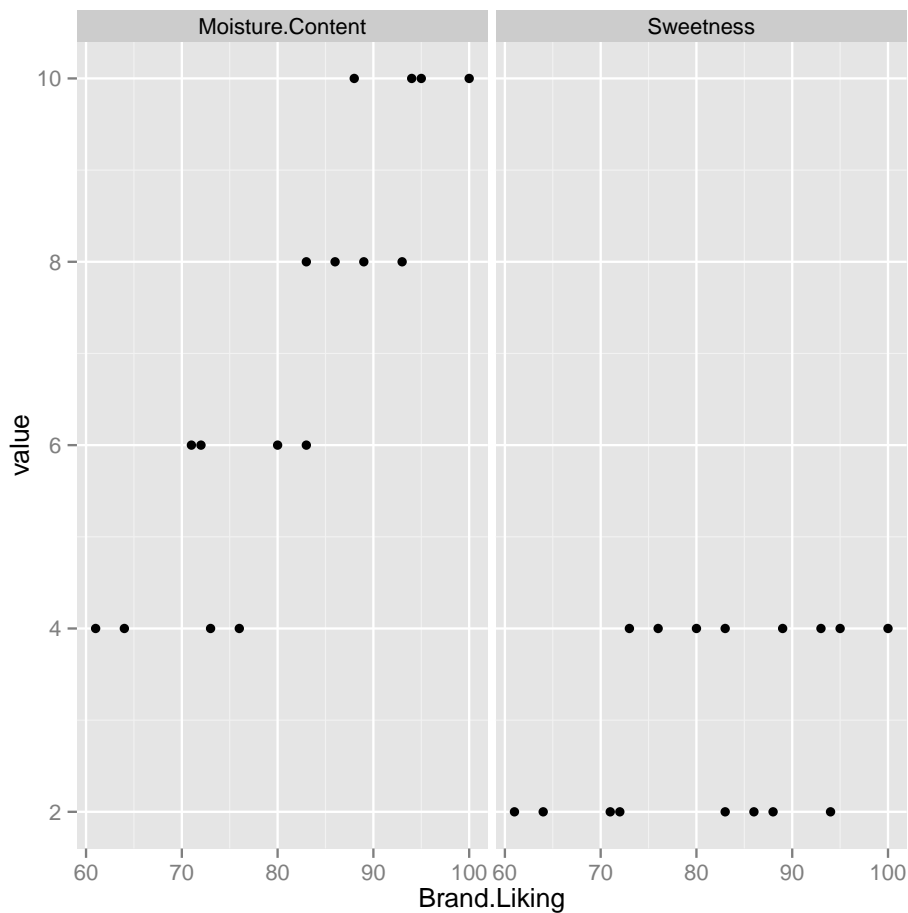
And now let's compute the correlation matrix:

```
> cor(BrandPreference)
```

```
                Brand.Liking Moisture.Content Sweetness
Brand.Liking       1.0000000        0.8923929 0.3945807
Moisture.Content   0.8923929        1.0000000 0.0000000
Sweetness          0.3945807        0.0000000 1.0000000
```

b) Now let's estimate the model $Y_i = b_0 + b_1 X_{i1} + b_2 X_{i2} + \epsilon_i$:

```
> fit <- lm(Brand.Liking~Moisture.Content+Sweetness,data=BrandPreference)
> fitsum <- summary(fit)
> fitsum

Call:
lm(formula = Brand.Liking ~ Moisture.Content + Sweetness, data = BrandPreference)

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 ***
Moisture.Content 4.4250     0.3011  14.695 1.78e-09 ***
Sweetness       4.3750     0.6733   6.498 2.01e-05 ***
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 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
```
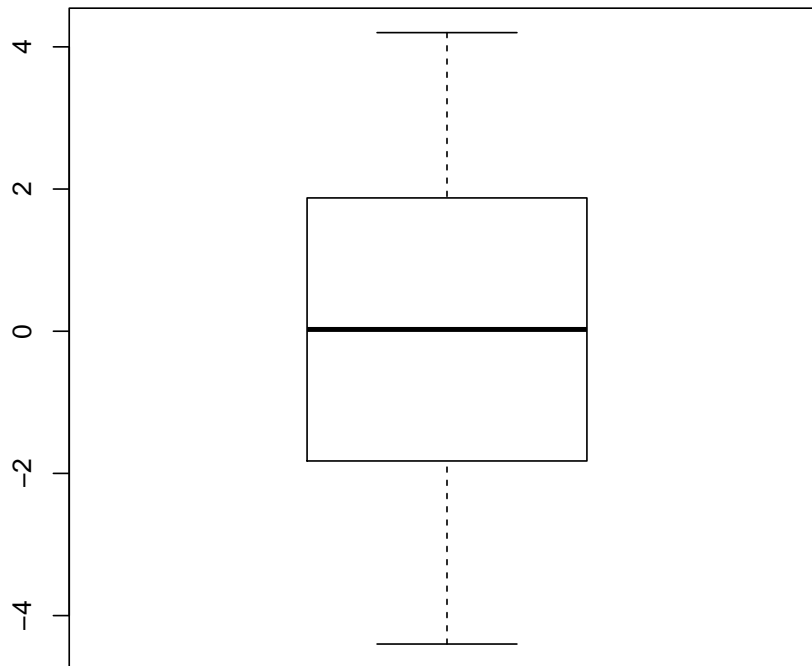
So, the estimated model is:
$$Y_i = 37.65 + 4.425X_{i1} + 4.375X_{i2}$$

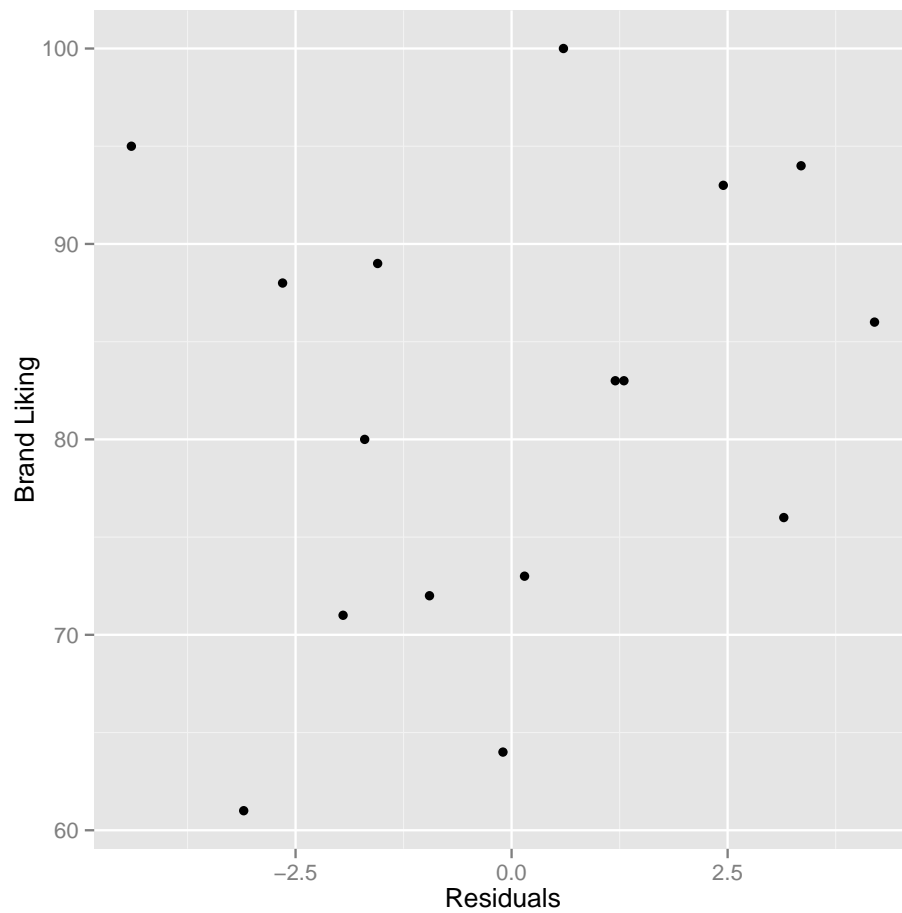c) Now let's examine the residuals:

> *boxplot(fit$residuals)*

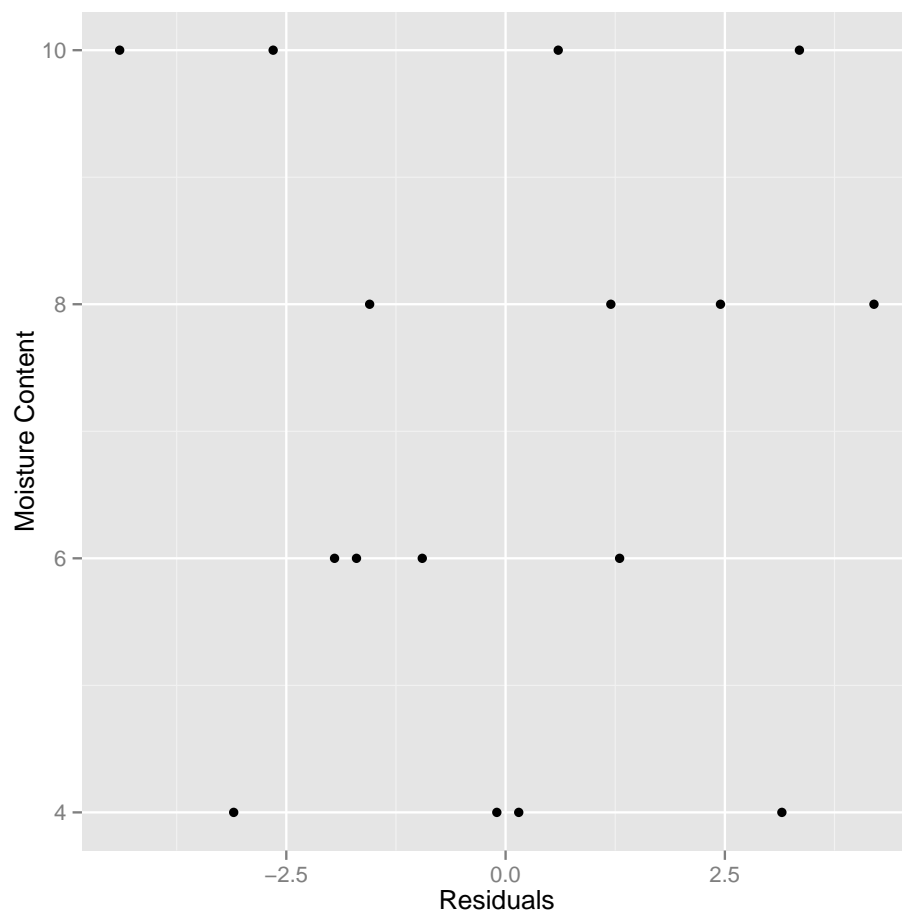This plot shows that the residuals are roughly normally distributed around 0.

d) Let's plot the residuals against Y, X_1, X_2 , and X_1 X_2:

```
> qplot(x=fit$residuals,
+       y=fit$model$Brand.Liking,
+       xlab="Residuals",
+       ylab="Brand Liking")
```



```
> qplot(x=fit$residuals,
+       y=fit$model$Moisture.Content,
+       xlab="Residuals",
+       ylab="Moisture Content")
```
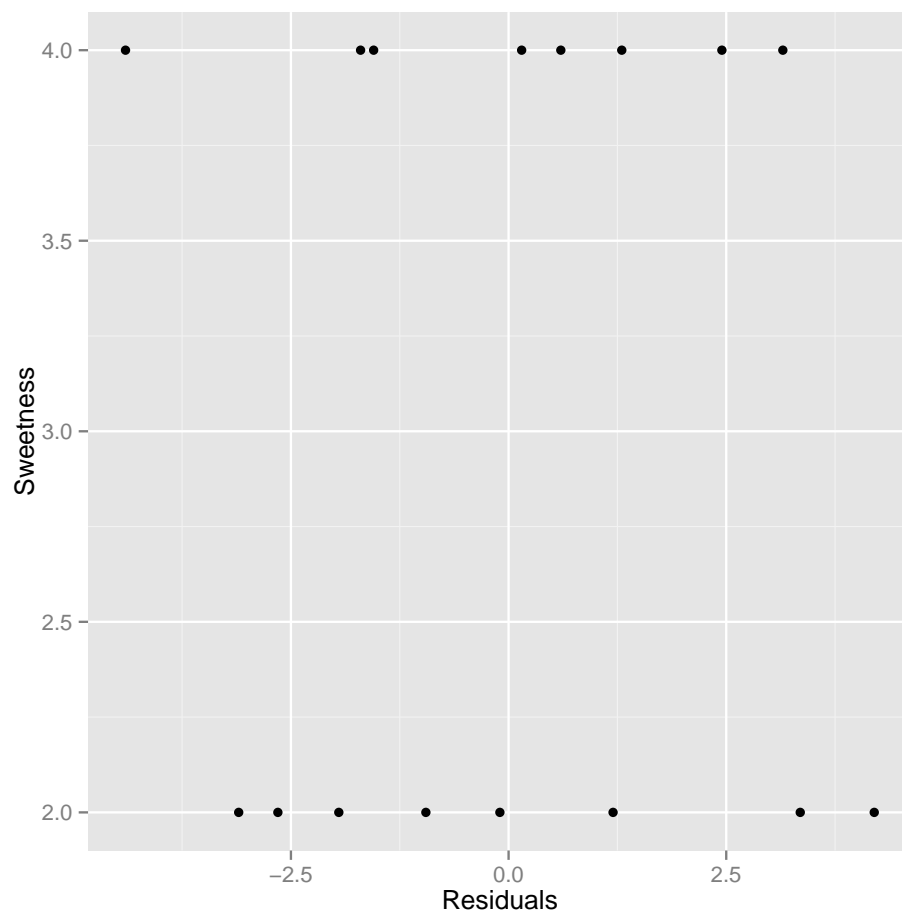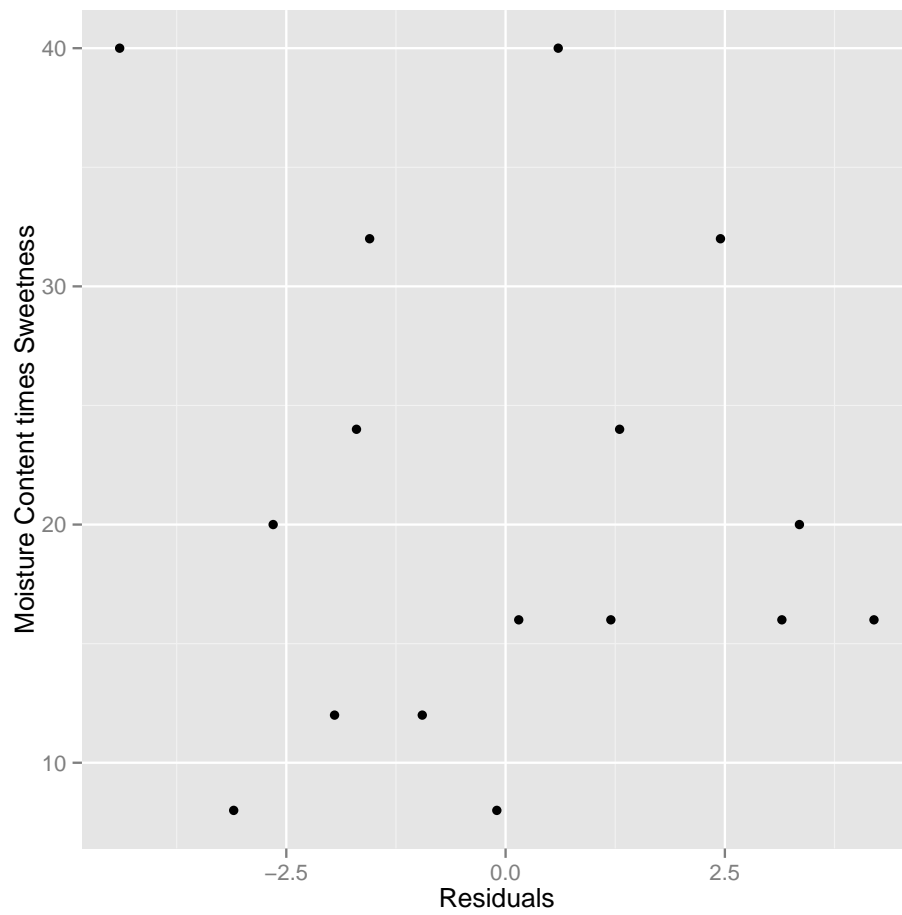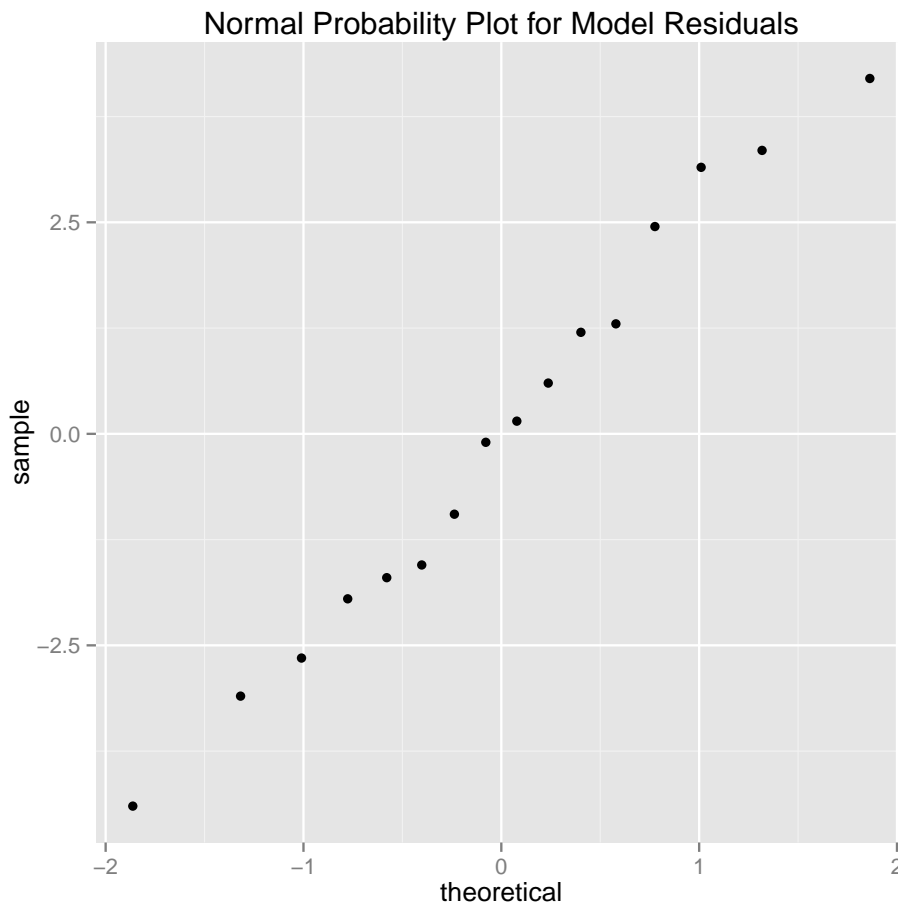
```
> qplot(x=fit$residuals,
+       y=fit$model$Sweetness,
+       xlab="Residuals",
+       ylab="Sweetness")
```

```
> qplot(x=fit$residuals,
+       y=fit$model$Moisture.Content*fit$model$Sweetness,
+       xlab="Residuals",ylab="Moisture Content times Sweetness")
```

```
> qplot(sample=fit$residuals,
+       stat="qq",
+       main="Normal Probability Plot for Model Residuals")
```

Normal Probability Plot for Model Residuals

The plots indicate the the residuals are not only distributed normally, but are uncorrelated with either any of the predictors or the outcome. The model assumptions hold.

e) Let's perform the Breush-Pagan test for heteroskedasticity of the error terms

```
> require(lmtest)
> pb <- bptest(fit,studentize=F)
> pb

        Breusch-Pagan test

data:  fit
BP = 1.0422, df = 2, p-value = 0.5939
```

Clearly, we cannot reject the null hypothesis that the model exhibits homoskedasticity.

2   7.3

Refer to Brand Preference Problem 6.5:

a) Obtain the analysis of variance table that decomposes the regression sum of squares into extra sums of squares associated with X_1, and with X_2 , given X_1.

b) Test whether X_2 can be dropped from the regression model given that X_1 is retained. Use the F^* test statistic and level of significance .01. State the alternatives, decision rule, and conclusion. What is the P-value of the test?

2.1 Answer:

a) First we must fit the first order model:

```
> fo <- lm(Brand.Liking~Moisture.Content,data=BrandPreference)
> fosum <- summary(fo)
> fosum

Call:
lm(formula = Brand.Liking ~ Moisture.Content, data = BrandPreference)

Residuals:
   Min     1Q Median     3Q    Max
-7.475 -4.688 -0.100  4.638  7.525

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)        50.775      4.395  11.554 1.52e-08 ***
Moisture.Content    4.425      0.598   7.399 3.36e-06 ***
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 1

Residual standard error: 5.349 on 14 degrees of freedom
Multiple R-squared:  0.7964,        Adjusted R-squared:  0.7818
F-statistic: 54.75 on 1 and 14 DF,  p-value: 3.356e-06
```

Now we can find $SSR(X_2|X_1)$:

```
> ssr_x2x1 <- deviance(fo) - deviance(fit)
> ssr_x2x1

[1] 306.25
```

So, the extra sum of squares, $SSR(X_2|X_1) = 306.25$.

b) Now let's test to see if we can drop X_2 given X_1. The alternatives are as follows:

$$H_0 : \beta_3 = 0$$
$$H_a : \beta_3 \neq 0$$

The decision rule is reject H_0 if:

$$F^* > F(0.99, 1, 13)$$

Let's compute the test statistic and critical value:

```
> # test statistic
> F_star <- ((deviance(fo) - deviance(fit)) /
+    ((length(BrandPreference$Brand.Liking) - fosum$df[1]) - (length(BrandPreference$Brand.Liking) -
+    (deviance(fit)/(length(BrandPreference$Brand.Liking) - fitsum$df[1]))
> # critical value
> cv <- qf(0.99,1,length(BrandPreference$Brand.Liking)-fitsum$df[1])
> print(paste0("F^star is ",F_star," and the critical value is ",cv,"."))

[1] "F^star is 42.2189819724284 and the critical value is 9.07380572851566."
```

Clearly we reject H_0 and conclude that X_2 should remain in the model. Let's compute the p-value of the test statistic:

```
> pval <- 1 - pf(F_star,1,length(BrandPreference$Brand.Liking)-fitsum$df[1])
> print(paste0("The p-value for F^star is: ",pval))

[1] "The p-value for F^star is: 2.01104739359081e-05"
```

3 7.12

Refer to Brand Preference Problem 6.5: Calculate R_{Y1}^2, R_{Y2}^2, R_{12}^2, $R_{Y1|2}^2$, $R_{Y2|1}^2$, and R^2. Explain what each coefficient measures and interpret your results.

3.1 Answer:

4 7.24

Refer to Brand Preference Problem 6.5:

a) Fit first-order simple linear regression model (2.1) for relating brand liking (Y) to moisture content (X_1). State the fitted regression function.

b) Compare the estimated regression coefficient for moisture content obtained in part (a) with the corresponding coefficient obtained in Problem 6.5b. What do you find?

c) Does $SSR(X_1)$ equal $SSR(X_1|X_2)$ here? If not, is the difference substantial?

d) Refer to the correlation matrix obtained in Problem 6.5a. What bearing does this have on your findings in parts (b) and (c)?

4.1 Answer:

a) Let's first fit the first order model:

```
> fo <- lm(Brand.Liking~Moisture.Content,data=BrandPreference)
> fosum <- summary(fo)
> fosum

Call:
lm(formula = Brand.Liking ~ Moisture.Content, data = BrandPreference)

Residuals:
    Min     1Q Median     3Q    Max
```

```
-7.475 -4.688 -0.100  4.638  7.525


Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)       50.775      4.395  11.554 1.52e-08 ***
Moisture.Content   4.425      0.598   7.399 3.36e-06 ***
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 1

Residual standard error: 5.349 on 14 degrees of freedom
Multiple R-squared: 0.7964,       Adjusted R-squared:  0.7818
F-statistic: 54.75 on 1 and 14 DF,  p-value: 3.356e-06
```

The estimated function is:
$$Y_i = 50.775 + 4.425X_{i1}$$

b) Let's compare the estimated models:

```
> # two predictor model
> fitsum

Call:
lm(formula = Brand.Liking ~ Moisture.Content + Sweetness, data = BrandPreference)

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 ***
Moisture.Content  4.4250     0.3011  14.695 1.78e-09 ***
Sweetness         4.3750     0.6733   6.498 2.01e-05 ***
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 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

> # one predictor model
> fosum

Call:
lm(formula = Brand.Liking ~ Moisture.Content, data = BrandPreference)

Residuals:
   Min     1Q Median     3Q    Max
-7.475 -4.688 -0.100  4.638  7.525


Coefficients:
                Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept)         50.775      4.395  11.554 1.52e-08 ***
Moisture.Content     4.425      0.598   7.399 3.36e-06 ***
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 1

Residual standard error: 5.349 on 14 degrees of freedom
Multiple R-squared:  0.7964,      Adjusted R-squared:  0.7818
F-statistic: 54.75 on 1 and 14 DF,  p-value: 3.356e-06
```

We find that the coefficients are the same in both models for Sweetness.

c) Let's find $SSR(X_1|X_2)$:

5 7.28

a Define each of the following exta sums of squares: (I) $SSR(X_5|X_1)$, (2) $SSR(X_3, X_4|X_1)$, (3) $SSR(X_4|X_1, X_2, X_3)$.

b For a multiple regression model with five X variables, what is the relevant exta sum of squares for testing whether or not $\beta_5 = 0$? whether or not $\beta_2 = \beta_4 = 0$?

5.1 Answer:

a) 1)
$$SSR(X_5|X_1) = SSR(X_1, X_5) - SSR(X_1)$$

2)
$$SSR(X_3, X_5|X_1) = SSR(X_1, X_3, X_5) - SSR(X_1)$$

3)
$$SSR(X_4|X_1, X_2, X_3) = SSR(X_1, X_2, X_3, X_4) - SSR(X_1, X_2, X_3)$$

b) The relevant sum of squares for whether or not $\beta_5 = 0$:

$$SSR(X_5|X_1, X_2, X_3, X_4, X_5)$$

The relevant sum of squares for whether or not $\beta_2 = \beta_4 = O$:

$$SSR(X_2, X_4|X_1, X_3, X_5)$$

6 7.29

Show that:

a) $SSR(X_1, X_2, X_3, X_4) = SSR(X_1) + SSR(X_2, X_3|X_1) + SSR(X4|X_1, X_2, X_3)$.

b) $SSR(X_1, X_2, X_3, X_4) = SSR(X_2, X_3) + SSR(X_1|X_2, X_3) + SSR(X4|X_1, X_2, X_3)$.

6.1 Answer:

a)
$$SSR(X_1) + SSR(X_2, X_3|X_1) + SSR(X4|X_1, X_2, X_3)$$
$$= SSR(X_1) + SSR(X_1, X_2, X_3) - SSR(X_1) + SSR(X_1, X_2, X_3, X_4) - SSR(X_1, X_2, X_3)$$
$$= SSR(X_1, X_2, X_3, X_4)$$

as required.

b)
$$SSR(X_2, X_3) + SSR(X_1|X_2, X_3) + SSR(X4|X_1, X_2, X_3)$$
$$= SSR(X_2, X_3) + SSR(X_1, X_2, X_3) - SSR(X_2, X_3) + SSR(X_1, X_2, X_3, X_4) - SSR(X_1, X_2, X_3)$$
$$= SSR(X_1, X_2, X_3, X_4)$$

as required.

7 7.30

Refer to Brand Preference Problem 6.5:

a) Regress Y on X_1 using simple linear regression model (2.1) and obtain the residuals.

b) Regress X_1 on X_2 using simple linear regression model (2.1) and obtain the residuals.

c) Calculate the coefficient of simple correlation between the two sets of residuals and show that it equals $R_{Y1|2}$.

7.1 Answer:

a) We have already previously estimated this model, so we can obtain the residuals:

```
> res_fo <- residuals(fo)
> print(res_fo)

     1      2      3      4      5      6      7      8      9     10     11
-4.475  4.525 -7.475  7.525 -5.325  2.675 -6.325  5.675 -3.175  2.825 -0.175
    12     13     14     15     16
 6.825 -7.025 -0.025 -1.025  4.975
```

b) Now let's regress X_1 on X_2 and obtain the residuals:

```
> xonx <- lm(Moisture.Content~Sweetness,data=BrandPreference)
> res_xonx <- residuals(xonx)
> print(res_xonx)

 1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16
-3 -3 -3 -3 -1 -1 -1 -1  1  1  1  1  3  3  3  3
```

c) The simple correlation coefficient between the residuals for the two models is:

```
> cor(x=res_fo,y=res_xonx)

[1] -3.980956e-17
```

Now let's show that is is roughly equal to $R_{Y1|2}$:

```
>
```

8   8.11

Refer to Brand Preference Problem 6.5:

a) Fit regression model (8.22).

b) Test whether or not the interaction term can be dropped from the model; use $\alpha = .05$. State the alternatives, decision rule, and conclusion.

8.1   Answer:

a) Let's fit the model:

```
> wint <- lm(Brand.Liking~Moisture.Content+Sweetness+Moisture.Content*Sweetness,
+            data=BrandPreference)
> wint_sum <- summary(wint)
> wint_sum

Call:
lm(formula = Brand.Liking ~ Moisture.Content + Sweetness + Moisture.Content *
    Sweetness, data = BrandPreference)

Residuals:
   Min     1Q Median     3Q    Max
-4.150 -1.488  0.125  1.700  3.700

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                27.1500     6.4648   4.200  0.00123 **
Moisture.Content            5.9250     0.8797   6.735 2.09e-05 ***
Sweetness                   7.8750     2.0444   3.852  0.00230 **
Moisture.Content:Sweetness -0.5000     0.2782  -1.797  0.09749 .
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 1

Residual standard error: 2.488 on 12 degrees of freedom
Multiple R-squared:  0.9622,       Adjusted R-squared:  0.9528
F-statistic: 101.9 on 3 and 12 DF,  p-value: 8.379e-09
```

b) Now let's test and see if the interaction term can be removed from the model:

9   8.16

Refer to Grade point average Problem 1.19. An assistant to the director of admissions conjectured that the predictive power of the model could be improved by adding information on whether the student had chosen a major field of concentration at the time the application was submitted. Assume that regression model (8.33) is appropriate, where X_1 is entrance test score and $X_2 = 1$ if student had indicated a major field of concentration at the time of application and 0 if the major field was undecided.

a) Explain how each regression coefficient in model (8.33) is interpreted here.

b) Fit the regression model and state the estimated regression function.

c) Test whether the X_2 variable can be dropped from the regression model; use $\alpha = .01$. State the alternatives. decision rule. and conclusion.

d) Obtain the residuals for regression model (8.33) and plot them against $X_1 X_2$. Is there any evidence in your plot that it would be helpful to include an interaction term in the model?

9.1 Answer:

a) β_1 can be interpreted as the effect of a one unit increase in entrance test score - a one unit increase in entrance test score yeilds a β_1 unit increase in a student's first year GPA, on average. β_2 can be interpreted as the average difference in GPA points associated with students who stated thier major field at the time of application irrespective of entrance test score.

b) Fit the model:

```
> GPA <- read.table(file='8.16.txt',stringsAsFactors=F)
> names(GPA) <- c("GPA","Entrance.Test","Major.Indicated")
> gpafit <- lm(GPA~Entrance.Test+Major.Indicated,data=GPA)
> gpafit_sum <- summary(gpafit)
> gpafit_sum

Call:
lm(formula = GPA ~ Entrance.Test + Major.Indicated, data = GPA)

Residuals:
     Min       1Q    Median       3Q      Max
-2.70304 -0.35574  0.02541  0.45747  1.25037

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      2.19842    0.33886   6.488 2.18e-09 ***
Entrance.Test    0.03789    0.01285   2.949  0.00385 **
Major.Indicated -0.09430    0.11997  -0.786  0.43341
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 1

Residual standard error: 0.6241 on 117 degrees of freedom
Multiple R-squared:  0.07749,       Adjusted R-squared:  0.06172
F-statistic: 4.914 on 2 and 117 DF,  p-value: 0.008928
```
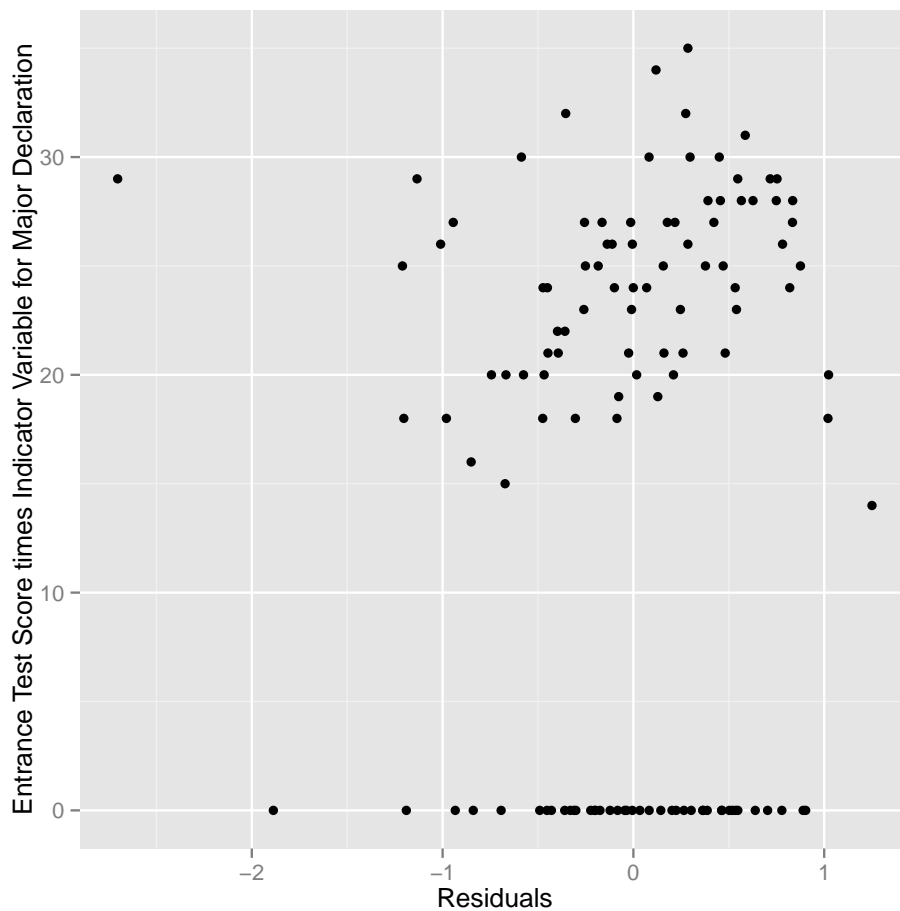
The estimated model is:

$$Y_i = 2.19841928804895 + 0.0378939641261244X_{i1} + 0.0378939641261244X_{i2}$$

c) Construct a test:

d) Let's plot the residuals vs the variable $X_1 X_2$:

```
> qplot(x=residuals(gpafit),
+       y=Entrance.Test*Major.Indicated,data=GPA,
+       xlab="Residuals",
+       ylab="Entrance Test Score times Indicator Variable for Major Declaration")
```

There appears to be a positive relationship between the residuals and X_1X_2 when $X_2 = 0$. It may be beneficial to include the term.

10 8.20

Refer to Grade point average Problems 1.19 and 8.16:

a) Fit regression model (8.49) and state the estimated regression function.

b) Test whether the interaction term can be dropped from the model; use $\alpha = .05$. State the alternatives. decision rule, and conclusion. If the interaction term cannot be dropped from the model, describe the nature of the interaction effect.

10.1 Answer:

a) Fit the model:

```
> intfit <- lm(GPA~Entrance.Test+Major.Indicated+Entrance.Test*Major.Indicated,
+              data=GPA)
> intfit_sum <- summary(intfit)
> intfit_sum
```

```
Call:
lm(formula = GPA ~ Entrance.Test + Major.Indicated + Entrance.Test *
    Major.Indicated, data = GPA)

Residuals:
    Min      1Q   Median      3Q      Max
-2.80187 -0.31392  0.04451  0.44337  1.47544

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                   3.226318   0.549428   5.872 4.18e-08 ***
Entrance.Test                -0.002757   0.021405  -0.129   0.8977
Major.Indicated              -1.649577   0.672197  -2.454   0.0156 *
Entrance.Test:Major.Indicated 0.062245   0.026487   2.350   0.0205 *
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 1

Residual standard error: 0.6124 on 116 degrees of freedom
Multiple R-squared:  0.1194,       Adjusted R-squared:  0.09664
F-statistic: 5.244 on 3 and 116 DF,  p-value: 0.001982
```

The estimated model is:

$$Y_i = 3.2263184991274 + -0.00275741710296684X_{i1} + -1.64957722409641X_{i2} + 0.0622446509862124X_{i1}X_{i2}$$

b) Test whether the interaction term is beneficial to the model:

11   8.42

Refer to Market share data set in Appendix C.3. Company executives want to be able to predict market share of their product (Y) based on merchandise price (X_1), the gross Nielsen rating points $(X_1,$ an index of the amount of advertising exposure that the product received); the presence or absence of a wholesale pricing discount $(X_3 = 1$ if discount present: otherwise $X_3 = 0)$; the presence or absence of a package promotion during the period $(X_4 = 1$ if promotion present: otherwise $X_4 = 0)$: and year (X_5). Code year as a nominal level variable and use 2000 as the reference year.

a) Fit a first-order regression model. Plot the residuals against the fitted values. How well does the first-order model appear to fit the data?

b) Re-fit the model in part (a). After adding all second-order terms involving only the quantitative predictors. Test whether or not all quadratic and interaction terms can be dropped from the regression model: use $\alpha = .05$. State the alternatives. decision rule, and conclusion.

c) In part (a), test whether advertising index (X_2) and year (X_5) can be dropped from the model; use $\alpha = .05$. State the alternatives, decision rule, and conclusion.

11.1   Answer:

a) First, let's read in the data and recode the year variable into an indicator with a base of the year 2000:

```
> MarketShare <- read.table('MarketShare.txt',stringsAsFactors=F)
> names(MarketShare) <- c("id","Market.Share","Merchandise.Price","Nielson.Rating","Wholesale.Price
> MarketShare$i1999 <- 1*(MarketShare$Year == 1999)
> MarketShare$i2001 <- 1*(MarketShare$Year == 2001)
> MarketShare$i2002 <- 1*(MarketShare$Year == 2002)
```

12   System Information

```
> sessionInfo();
```

```
R version 3.0.2 (2013-09-25)
Platform: x86_64-pc-linux-gnu (64-bit)

locale:
 [1] LC_CTYPE=en_US.UTF-8 LC_NUMERIC=C           LC_TIME=C
 [4] LC_COLLATE=C         LC_MONETARY=C          LC_MESSAGES=C
 [7] LC_PAPER=C           LC_NAME=C              LC_ADDRESS=C
[10] LC_TELEPHONE=C       LC_MEASUREMENT=C       LC_IDENTIFICATION=C


attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods   base


other attached packages:
[1] lmtest_0.9-32   zoo_1.7-10       reshape2_1.2.2  ggplot2_0.9.3.1


loaded via a namespace (and not attached):
 [1] MASS_7.3-29        RColorBrewer_1.0-5 colorspace_1.2-4   dichromat_2.0-0
 [5] digest_0.6.3       grid_3.0.2         gtable_0.1.2       labeling_0.2
 [9] lattice_0.20-23    munsell_0.4.2      plyr_1.8           proto_0.3-10
[13] scales_0.2.3       stringr_0.6.2      tools_3.0.2
```