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.

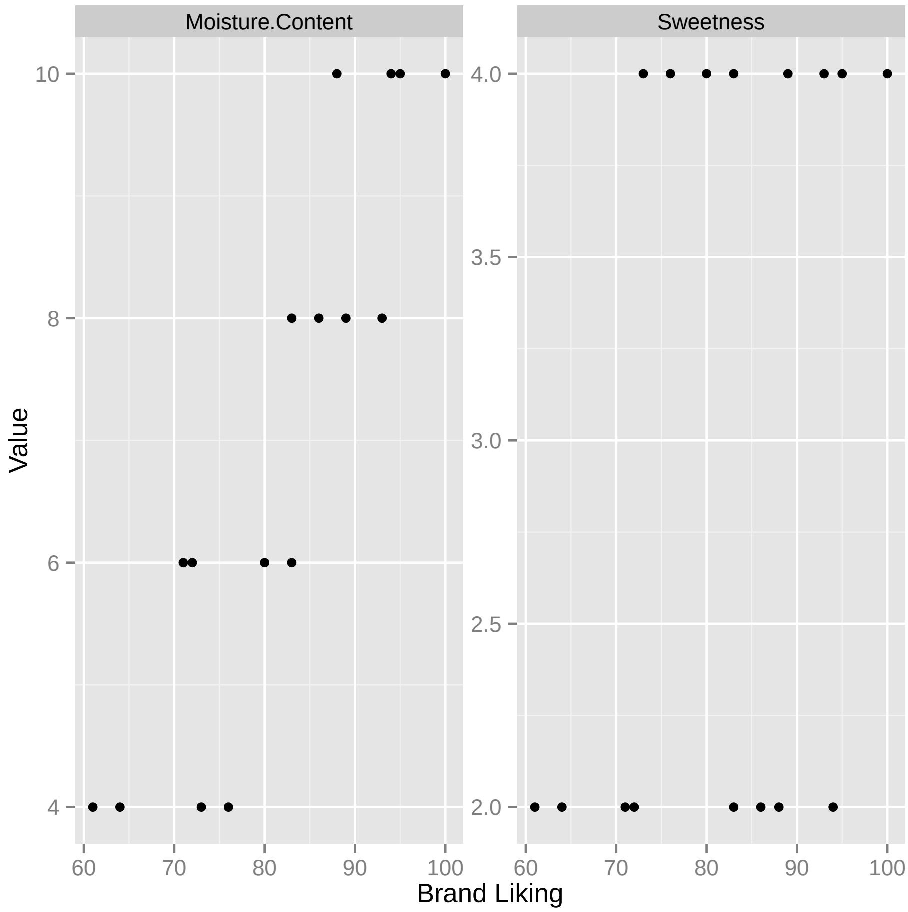
1.1 Answer:

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,scales="free_y") + xlab("Brand Liking") + ylab("Value")
```



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
```

These aids show that while both moisture content and sweetness are correlated with brand liking, that moisture content and sweetness are not correlated with each other.

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
```

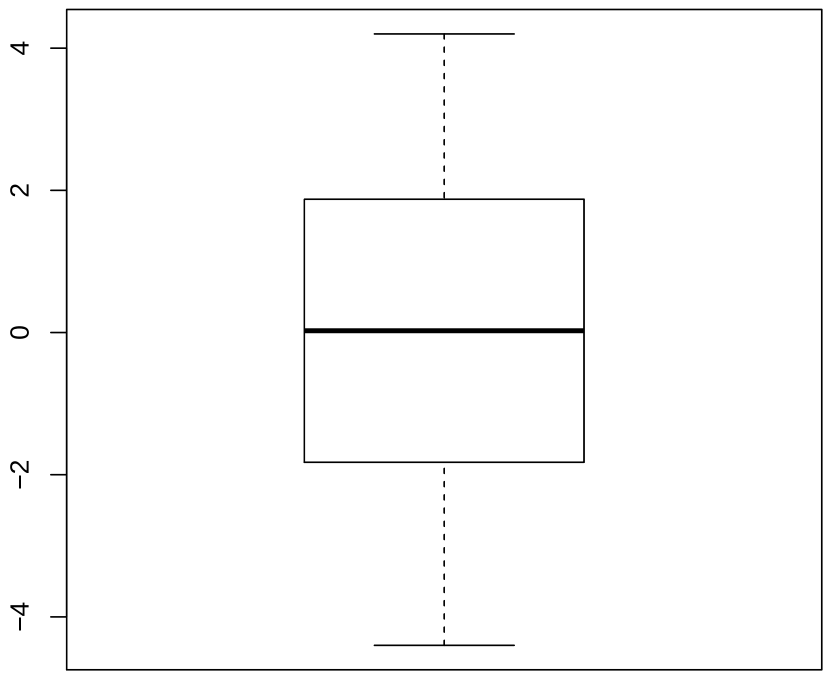
So, the estimated model is:

$$Y_i = 37.65 + 4.425X_{i1} + 4.375X_{i2}$$

An interpretation for b_1 is that for every one unit increase in moisture content, brand liking increases by 4.425 units.

c) Now let's examine the residuals:

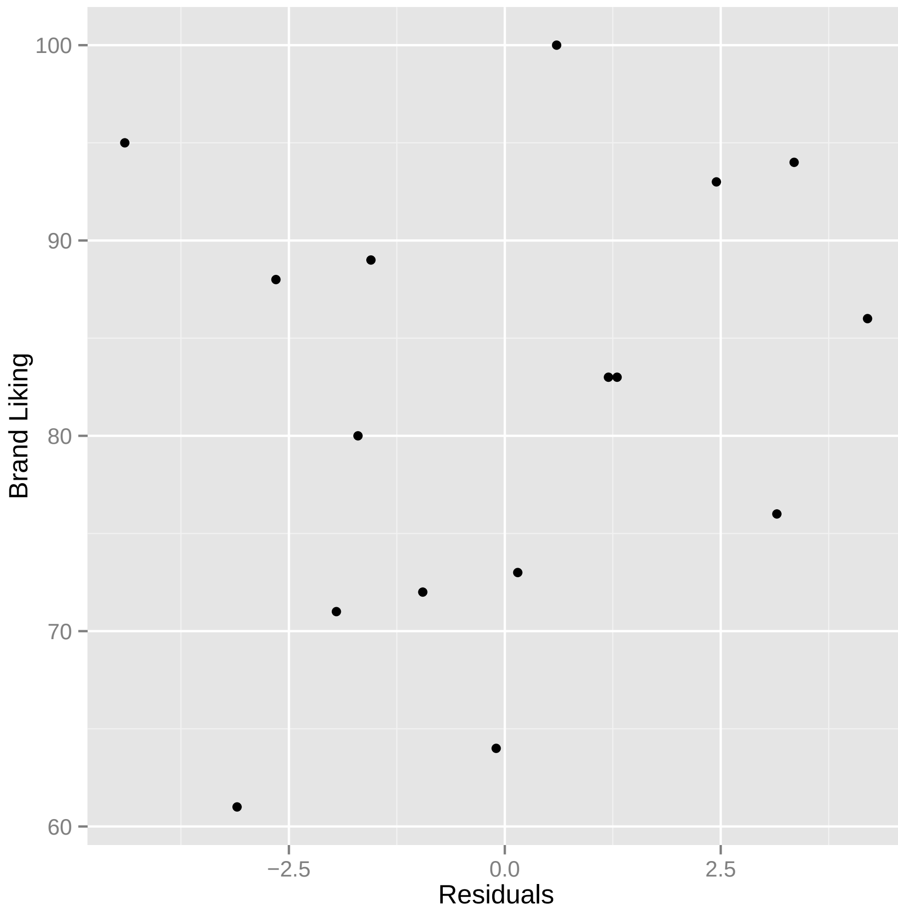
```
> boxplot(fit$residuals)
```



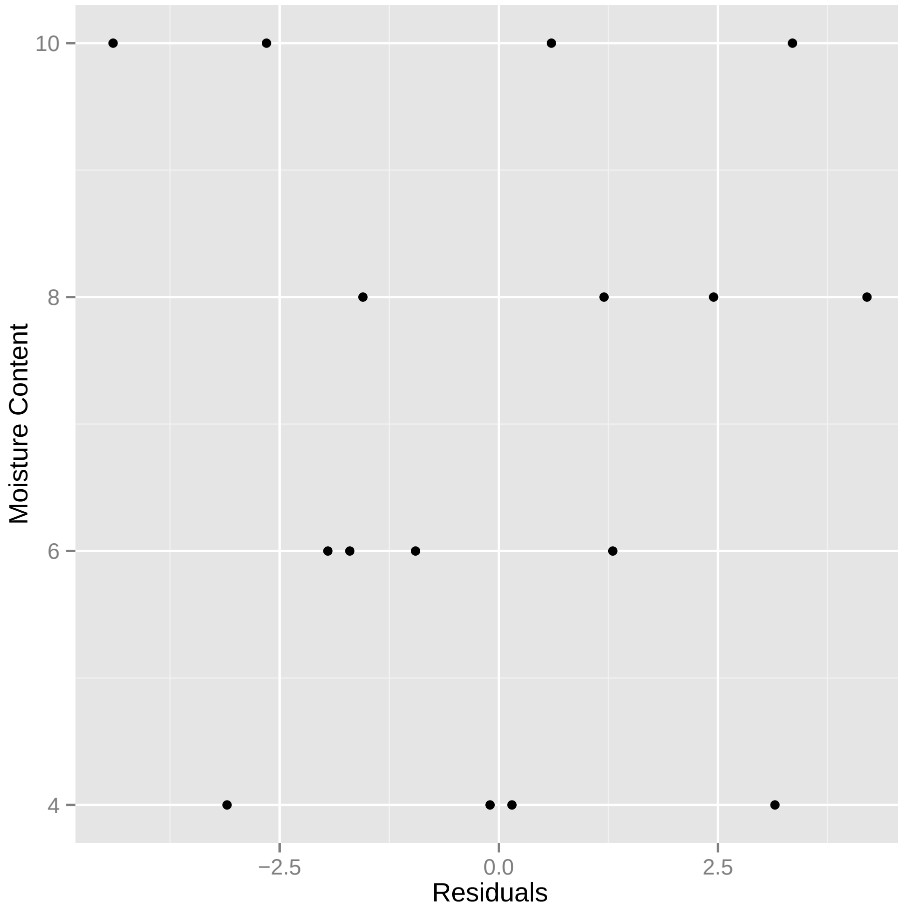
This plot shows that the residuals are roughly normally distributed around 0.

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

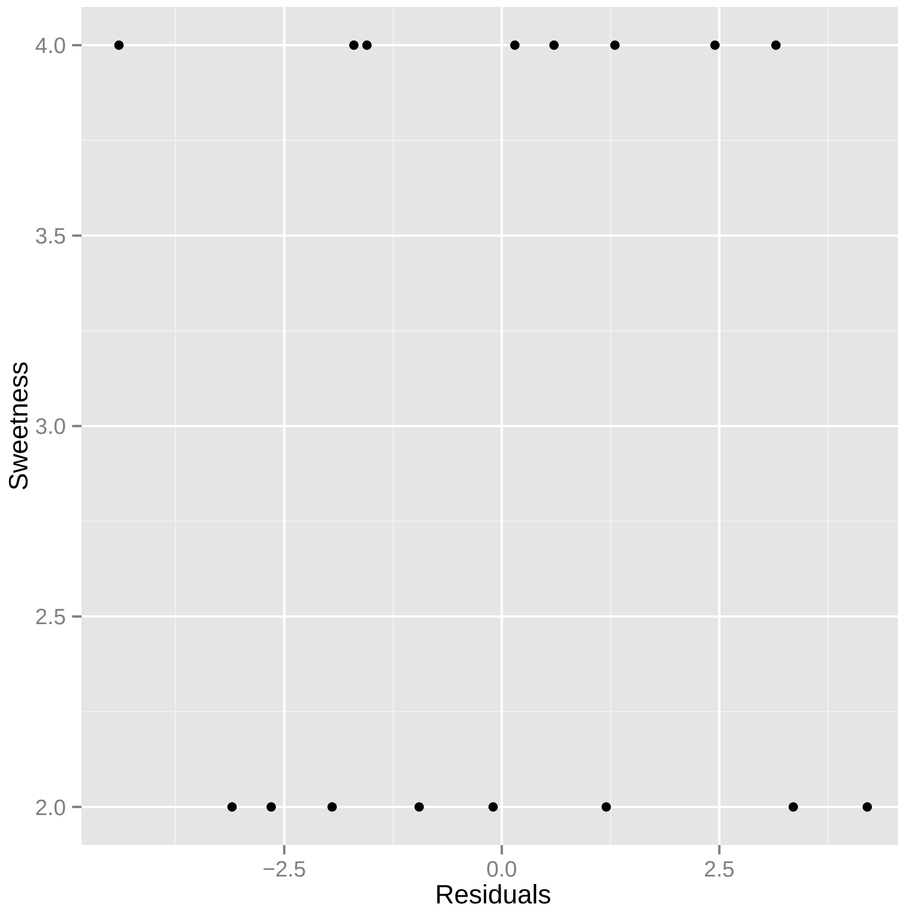
```
> qplot(x=fit$residuals,
+       y=fit$model$Brand.Liking,
+       xlab="Residuals",
+       ylab="Brand Liking")
```



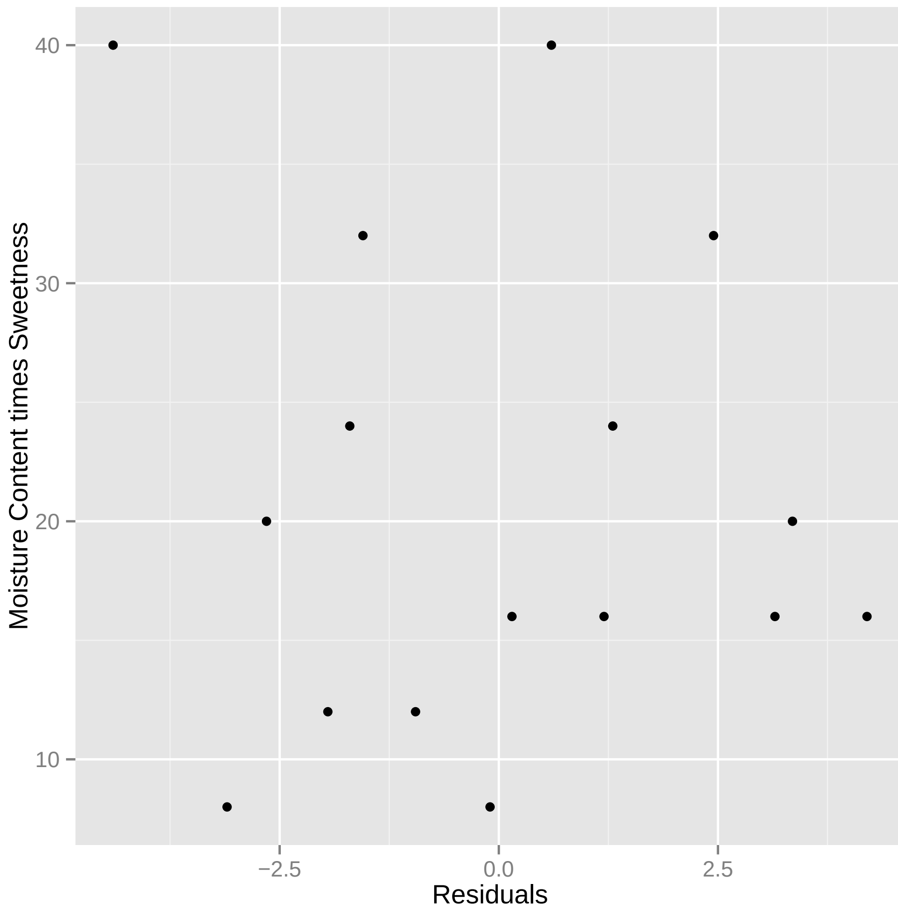

```
> qplot(x=fit$residuals,
+       y=fit$model$Moisture.Content,
+       xlab="Residuals",
+       ylab="Moisture Content")
```



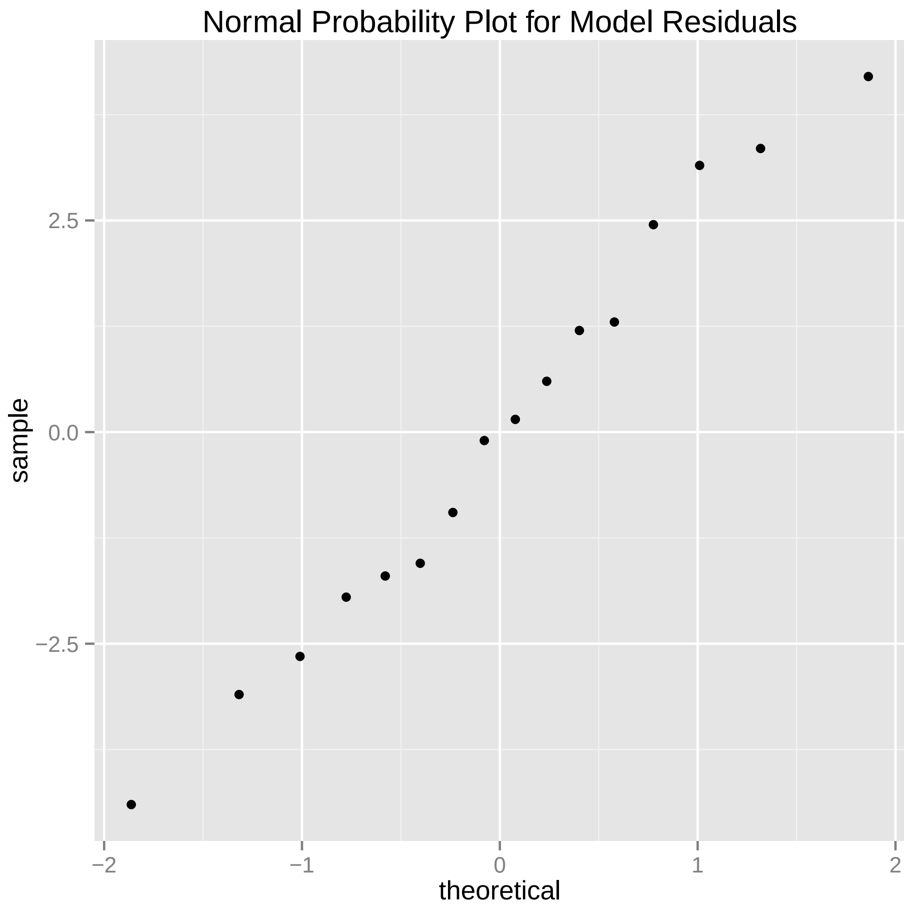

```
> qplot(x=fit$residuals,
+       y=fit$model$Sweetness,
+       xlab="Residuals",
+       ylab="Sweetness")
```




```
> 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")
```



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

e) Let's perform the Breush-Pagan test for heteroskedasticity of the error terms. The alternatives are as follows:

$$H_0 : \gamma_1 = \gamma_2 = 0$$
$$H_a : \gamma_1 \neq 0 \text{ or } \gamma_2 \neq 0$$

The decision rule is reject H_0 if:

$$\chi^2_{BP} > \chi^2_{BP}(0.99, 2)$$

Let's compute the test statistic and critical value:

```
> # compute test statistic
> efit <- lm(I(fit$residuals^2)~fit$model$Moisture.Content+fit$model$Sweetness)
> SSR_star <- sum(anova(efit)$`Sum Sq`) - deviance(efit)
> SSEfit <- deviance(fit)
> chisq_bp <- (SSR_star/2)/(SSEfit/length(fit$model$Brand.Liking))^2
> # compute the critical value
```

```
> cv <- qchisq(0.99,2)
> print(paste0("ChiSqr^2_{BP} is ",chisq_bp," and the critical value is ",cv,"."))
```

```
[1] "ChiSqr^2_{BP} is 1.04223856310212 and the critical value is 9.21034037197618."
```

Clearly, we cannot reject the null hypothesis - the model has constant error variance.

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, $SSR(X_2|X_1) = 306.25$.

The extra sum of squares table would look like the following:

Source of Variation	SS	df	MS
Regression	1872.7	2	1872.7
Moisture Content	1566.45	1	1566.45
Moisture Content|Sweetness	306.25	1	306.25
Error	94.3	13	7.25384615384616
Total	1967	12	

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$Brand.Liking) - fitsum$df[1]))) /
+   (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^2_{Y1}, R^2_{Y2}, R^2_{12}, $R^2_{Y1|2}$, $R^2_{Y2|1}$, and R^2. Explain what each coefficient measures and interpret your results.

3.1 Answer:

a) R^2_{Y1}:

```
> anovafit <- anova(fit)
> r2_y1 <- anovafit["Moisture.Content","Sum Sq"]/sum(anovafit$`Sum Sq`)
> r2_y1
```

```
[1] 0.796365
```

b) R^2_{Y2}:

```
> r2_y2 <- anovafit["Sweetness","Sum Sq"]/sum(anovafit$`Sum Sq`)
> r2_y2
```

```
[1] 0.155694
```

c) R^2_{12}:

```
> r2_12 <- (anovafit["Sweetness","Sum Sq"]+anovafit["Moisture.Content","Sum Sq"])/
+   sum(anovafit$`Sum Sq`)
> r2_12
```

```
[1] 0.952059
```

d) $R^2_{Y1|2}$:

```
> fo2 <- lm(Brand.Liking~Sweetness,data=BrandPreference)
> fosum2 <- summary(fo2)
> r2_y12 <- (deviance(fo2) - deviance(fit))/deviance(fo2)
> r2_y12
```

```
[1] 0.9432184
```

e) $R^2_{Y2|1}$:

```
> r2_y21 <- (deviance(fo) - deviance(fit))/deviance(fo)
> r2_y21
```

```
[1] 0.7645737
```

f) R^2:

```
> r2 <- fitsum$r.squared
> r2
```

```
[1] 0.952059
```

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 $SSR(X_1|X_2)$ and $SSR(X_1)$:

```
> SSR_x1x2 <- deviance(fo2) - deviance(fit)
> SSR_x1 <- sum(anova(fo)$`Sum Sq`) - deviance(fo)
> print(paste0("SSR(X1|X2) is ",SSR_x1x2," and SSR(X1) is ",SSR_x1))

[1] "SSR(X1|X2) is 1566.45 and SSR(X1) is 1566.45"
```

Clearly $SSR(X_1|X_2)$ and $SSR(X_1)$ are both the same.

d) As suggested by the correlation matrix, Moisture Content and Sweetness are uncorrelated, so we would expect the estimated coefficients from part (b) to be the same for both models. We would also expect the $SSR(X_1|X_2)$ to be the same as $SSR(X_1)$ since X_1 and X_2 are unrelated. The presence or absence of X_2 provides no information about X_1. The converse is also true.

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 $\beta_2 = \beta_4 = 0$:

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

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)

$$\begin{gathered} SSR(X_1) + SSR(X_2, X_3|X_1) + SSR(X4|X_1, X_2, X_3) \\ = SSR(X_1) + SSR(X_1, X_2, X_3) - SSR(X_1) + SSR(X_1, X_2, X_3, X_4) - SSR(X_1, X_2, X_3) \\ = SSR(X_1, X_2, X_3, X_4) \end{gathered}$$

as required.

b)

$$\begin{gathered} SSR(X_2, X_3) + SSR(X_1|X_2, X_3) + SSR(X4|X_1, X_2, X_3) \\ = SSR(X_2, X_3) + SSR(X_1, X_2, X_3) - SSR(X_2, X_3) + SSR(X_1, X_2, X_3, X_4) - SSR(X_1, X_2, X_3) \\ = SSR(X_1, X_2, X_3, X_4) \end{gathered}$$

as required.

7 7.30

Refer to Brand Preference Problem 6.5:

a) Regress Y on X_2 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_fo2 <- residuals(fo2)
> print(res_fo2)
```

```
      1       2       3       4       5       6       7       8       9      10
-13.375 -13.125 -16.375 -10.125  -5.375  -6.125  -6.375  -3.125   5.625   2.875
     11      12      13      14      15      16
  8.625   6.875  10.625   8.875  16.625  13.875
```

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_fo2,y=res_xonx)
```

```
[1] 0.9711943
```

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

```
> r_y12 <- sqrt((deviance(fo2) - deviance(fit))/deviance(fo2))
> r_y12
```

```
[1] 0.9711943
```

Which is clearly the same as the coefficient of simple correlation between the two sets of residuals.

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. The alternatives are as follows:

$$H_0 : \beta_4 = 0$$
$$H_a : \beta_4 \neq 0$$

The decision rule is reject H_0 if:

$$F^* > F(0.95, 1, 12)$$

Let's compute the test statistic and critical value:

```
> # test statistic
> F_star <- ((deviance(fit) - deviance(wint)) /
+   ((length(BrandPreference$Brand.Liking) - fitsum$df[1]) -
+      (length(BrandPreference$Brand.Liking) - wint_sum$df[1]))) /
+   (deviance(wint)/(length(BrandPreference$Brand.Liking) - wint_sum$df[1]))
> # critical value
> cv <- qf(0.95,1,length(BrandPreference$Brand.Liking)-wint_sum$df[1])
> print(paste0("F^star is ",F_star," and the critical value is ",cv,"."))

[1] "F^star is 3.23014804845221 and the critical value is 4.7472253467225."
```

Clearly we fail to reject H_0 and conclude that X_1X_2 should not remain in the model. Let's compute the p-value of the test statistic:

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

[1] "The p-value for F^star is: 0.0974862454531731"
```

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_1X_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) Now let's test and see if X_2 can be removed from the model. The alternatives are as follows:

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

The decision rule is reject H_0 if:

$$F^* > F(0.99, 1, 117)$$

Let's compute the test statistic and critical value:

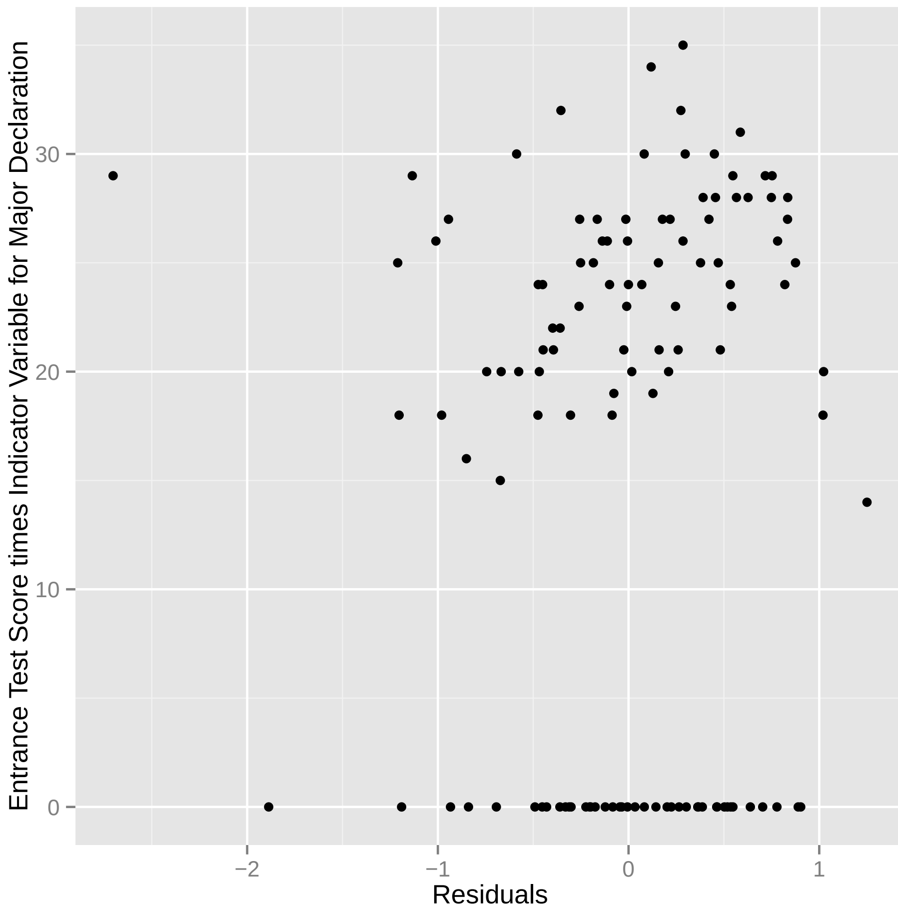
```
> # test statistic
> gpafo <- lm(GPA~Entrance.Test,data=GPA)
> gpafo_sum <- summary(gpafo)
> F_star <- ((deviance(gpafo) - deviance(gpafit)) /
+   ((length(GPA$GPA) - gpafo_sum$df[1]) -
+      (length(GPA$GPA) - gpafit_sum$df[1]))) /
+   (deviance(gpafit)/(length(GPA$GPA) - gpafit_sum$df[1]))
> # critical value
> cv <- qf(0.99,1,length(GPA$GPA)-gpafit_sum$df[1])
> print(paste0("F^star is ",F_star," and the critical value is ",cv,"."))
```

```
[1] "F^star is 0.617931372908239 and the critical value is 6.85656380811069."
```

Clearly we fail to reject H_0 and conclude that X_2 should not remain in the model.

d) Let's plot the residuals vs the variable X_1X_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 $X_2 = 1$. 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:

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

b) Let's test whether the interaction term is beneficial to the model. The alternatives are as follows:

$$H_0 : \beta_4 = 0$$
$$H_a : \beta_4 \neq 0$$

The decision rule is reject H_0 if:

$$F^* > F(0.95, 1, 117)$$

Let's compute the test statistic and critical value:

```
> # test statistic
> F_star <- ((deviance(gpafit) - deviance(intfit)) /
+   ((length(GPA$GPA) - gpafit_sum$df[1]) -
+      (length(GPA$GPA) - intfit_sum$df[1]))) /
+   (deviance(intfit)/(length(GPA$GPA) - intfit_sum$df[1]))
> # critical value
> cv <- qf(0.95,1,length(GPA$GPA)-intfit_sum$df[1])
> print(paste0("F^star is ",F_star," and the critical value is ",cv,"."))

[1] "F^star is 5.52263534905967 and the critical value is 3.92287936161707."
```

Clearly we reject H_0 and conclude that $X_1 X_2$ should remain in the model. The nature of the interaction effect is that if an incoming student declared a major at time of admittance, then the marginal effect of a one point increase in entrance test score on first year GPA is 0.0622446509862124 higher than the baseline marginal effect when a student doesn't initially indicate a major.

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_2, 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:

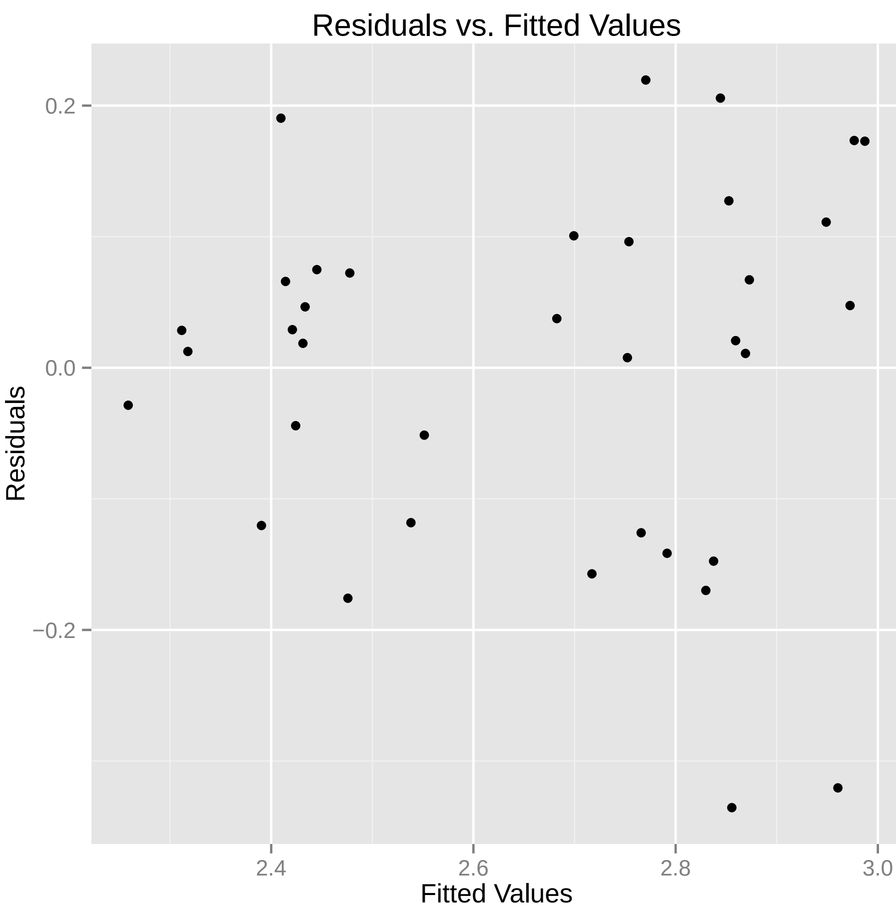
a) First, let's read in the data and recode the year variable into an indicator with a base of the year 2000:

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

Now fit the first order regression model and plot the residuals:

```
> msfit <- lm(Market.Share~Merchandise.Price+Nielson.Rating+
+               Wholesale.Price.Discount+Package.Promotion+
+               i1999+i2001+i2002,
+             data = MarketShare)
> msfit_sum <- summary(msfit)
> msfit_sum

Call:
lm(formula = Market.Share ~ Merchandise.Price + Nielson.Rating +
    Wholesale.Price.Discount + Package.Promotion + i1999 + i2001 +
    i2002, data = MarketShare)

Residuals:
     Min       1Q   Median       3Q      Max
-0.33558 -0.11872  0.02459  0.08020  0.21952

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              3.021e+00  4.705e-01   6.421 5.94e-07 ***
```

```
Merchandise.Price        -2.470e-01  1.982e-01  -1.246   0.2229
Nielson.Rating           -9.653e-05  1.914e-04  -0.504   0.6181
Wholesale.Price.Discount  4.093e-01  5.385e-02   7.601 2.80e-08 ***
Package.Promotion         1.240e-01  5.484e-02   2.261   0.0317 *
i1999                     1.324e-02  9.304e-02   0.142   0.8879
i2001                    -1.088e-01  7.133e-02  -1.525   0.1385
i2002                    -8.306e-02  8.657e-02  -0.959   0.3456
---
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: 0.1529 on 28 degrees of freedom
Multiple R-squared:  0.7326,        Adjusted R-squared:  0.6657
F-statistic: 10.96 on 7 and 28 DF,  p-value: 1.382e-06

> # plot the residuals
> qplot(x=msfit$fitted.values,y=msfit$residuals,
+       xlab="Fitted Values",
+       ylab="Residuals",
+       main="Residuals vs. Fitted Values")
```



The first order model appears to fit reasonably well. The residuals appear to have no relationship with the fitted values. The size of the residuals may get larger with increasing size of the fitted values, but it is not clear that this is the case.

b) Now let's refit with quadratic and interaction terms for the quantitative predictors.

```
> msfitint <- lm(Market.Share~Merchandise.Price+Nielson.Rating+
+                    I(Merchandise.Price^2)+I(Nielson.Rating^2)+
+                    Merchandise.Price*Nielson.Rating+Wholesale.Price.Discount+
+                    Package.Promotion+
+                    i1999+i2001+i2002,
+               data = MarketShare)
> msfitint_sum <- summary(msfitint)
> msfitint_sum

Call:
lm(formula = Market.Share ~ Merchandise.Price + Nielson.Rating +
    I(Merchandise.Price^2) + I(Nielson.Rating^2) + Merchandise.Price *
    Nielson.Rating + Wholesale.Price.Discount + Package.Promotion +
    i1999 + i2001 + i2002, data = MarketShare)

Residuals:
     Min       1Q   Median       3Q      Max
-0.33455 -0.08692  0.01892  0.07039  0.23931

Coefficients:
                                   Estimate Std. Error t value Pr(>|t|)
(Intercept)                       8.698e+00  6.818e+00   1.276   0.2138
Merchandise.Price                -4.803e+00  5.380e+00  -0.893   0.3805
Nielson.Rating                   -9.508e-04  3.492e-03  -0.272   0.7877
I(Merchandise.Price^2)            9.221e-01  1.069e+00   0.863   0.3965
I(Nielson.Rating^2)               5.518e-07  7.375e-07   0.748   0.4613
Wholesale.Price.Discount          3.941e-01  6.098e-02   6.463 9.09e-07 ***
Package.Promotion                 1.149e-01  5.772e-02   1.991   0.0575 .
i1999                             1.236e-02  1.006e-01   0.123   0.9031
i2001                            -1.006e-01  7.476e-02  -1.345   0.1906
i2002                            -5.807e-02  9.541e-02  -0.609   0.5483
Merchandise.Price:Nielson.Rating  1.629e-04  1.393e-03   0.117   0.9078
---
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: 0.1583 on 25 degrees of freedom
Multiple R-squared:  0.744,	Adjusted R-squared:  0.6417
F-statistic: 7.267 on 10 and 25 DF,  p-value: 2.837e-05
```

Now let's test to see if these added terms should be removed. The alternatives are as follows:

$$H_0 : \beta_4 = \beta_5 = \beta_6 = 0$$

$$H_a : \beta_4 \neq 0 \text{ or } \beta_5 \neq 0 \text{ or } \beta_6 \neq 0$$

The decision rule is reject H_0 if:

$$F^* > F(0.95, 1, 25)$$

Let's compute the test statistic and critical value:

```
> # test statistic
> F_star <- ((deviance(msfit) - deviance(msfitint)) /
+   ((length(MarketShare$id) - msfit_sum$df[1]) -
+      (length(MarketShare$id) - msfitint_sum$df[1]))) /
+   (deviance(msfitint)/(length(MarketShare$id) - msfitint_sum$df[1]))
> # critical value
> cv <- qf(0.95,1,length(MarketShare$id)-msfitint_sum$df[1])
> print(paste0("F^star is ",F_star," and the critical value is ",cv,"."))
```

```
[1] "F^star is 0.374002331978545 and the critical value is 4.24169905027715."
```

Clearly, we fail to reject H_0 and conclude that we should drop quadratic and interaction terms from the model.

c) Now let's test to see if we can remove year indicators and the Nielson rating: The alternatives are as follows:

$$H_0 : \beta_3 = \beta_6 = \beta_7 = \beta_8 = 0$$
$$H_a : \beta_3 \neq 0 \text{ or } \beta_6 \neq 0 \text{ or } \beta_7 \neq 0 \text{ or } \beta_8 \neq 0$$

The decision rule is reject H_0 if:

$$F^* > F(0.95, 1, 28)$$

Let's compute the test statistic and critical value. We must also fit a reduced model:

```
> # reduced model:
> msfitrd <- lm(Market.Share~Merchandise.Price+Wholesale.Price.Discount+Package.Promotion,
+             data = MarketShare)
> msfitrd_sum <- summary(msfitrd)
> # test statistic
> F_star <- ((deviance(msfitrd) - deviance(msfit)) /
+   ((length(MarketShare$id) - msfitrd_sum$df[1]) -
+      (length(MarketShare$id) - msfit_sum$df[1]))) /
+   (deviance(msfit)/(length(MarketShare$id) - msfit_sum$df[1]))
> # critical value
> cv <- qf(0.95,1,length(MarketShare$id)-msfit_sum$df[1])
> print(paste0("F^star is ",F_star," and the critical value is ",cv,"."))
```

```
[1] "F^star is 0.681718770588163 and the critical value is 4.19597181855776."
```

Clearly, we fail to reject H_0 and conclude that we should drop year indicators and the Nielson rating. This model estimation follows:

```
> msfitrd_sum
```

```
Call:
lm(formula = Market.Share ~ Merchandise.Price + Wholesale.Price.Discount +
    Package.Promotion, data = MarketShare)

Residuals:
      Min        1Q    Median        3Q       Max
-0.286376 -0.100465 -0.002259  0.104174  0.240020

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)               3.18527    0.36505   8.726  5.7e-10 ***
Merchandise.Price        -0.35269    0.15738  -2.241   0.0321 *
Wholesale.Price.Discount  0.39914    0.05125   7.787  7.0e-09 ***
Package.Promotion         0.11803    0.05149   2.292   0.0286 *
---
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: 0.1498 on 32 degrees of freedom
Multiple R-squared:  0.7065,	Adjusted R-squared:  0.679
F-statistic: 25.68 on 3 and 32 DF,  p-value: 1.191e-08
```

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] 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] munsell_0.4.2      plyr_1.8           proto_0.3-10       scales_0.2.3
[13] stringr_0.6.2      tools_3.0.2
```