HW4 GR5205 - Multiple Linear Regression Model

Mathieu Sauterey - UNI: mjs2364

6 November 2017

Problem 1 (50p) (Problems 6.18 (b-f), 6.21, & 7.7 in ALRM book)

A commercial real estate company evaluates vacancy rates, square footage, rental rates, and operating expenses for commercial properties in a large metropolitan area in order to provide clients with quantitative information upon which to make rental decisions. The data are taken from 81 suburban commercial properties that are the newest, best located, most attractive, and expensive for five specific geographic areas. The variables are: rental rates Y , the age X1, operating expanses and taxes X2, vacancy rates X3, and total square footage X4.

(a)(5p) Obtain the scatter plot matrix and the correlation matrix. Interpret these and state your principal findings.

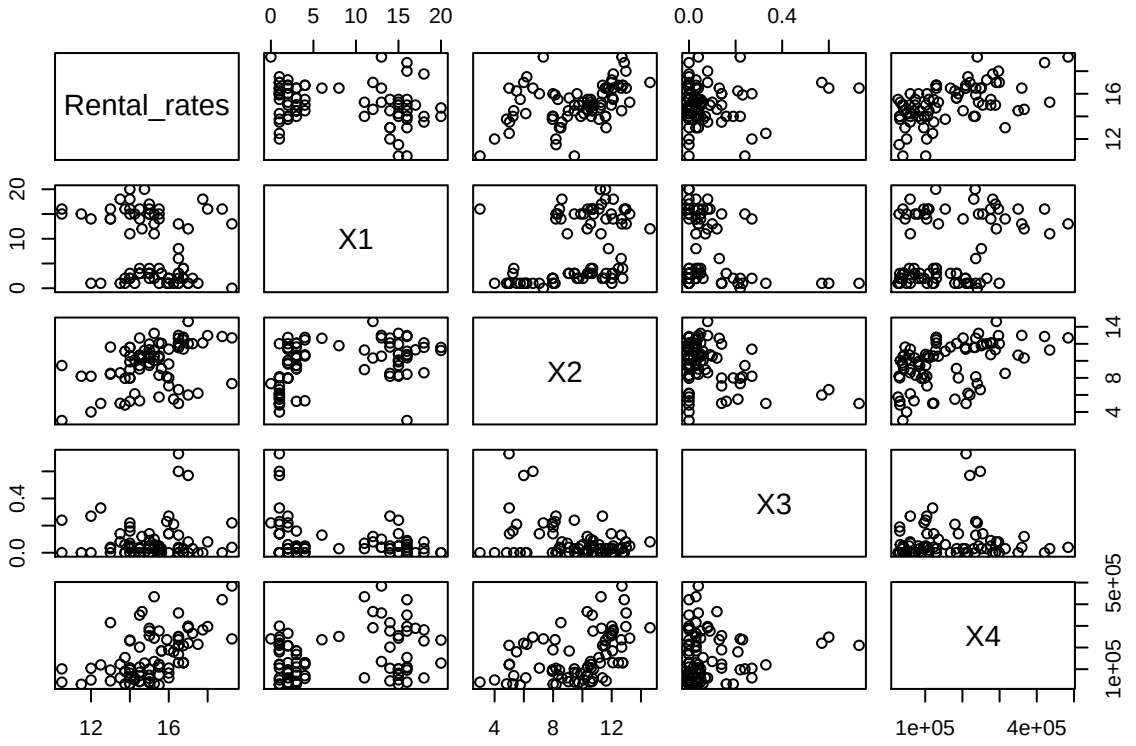
First, we read the data using data.table function and store it in a data frame named *data*. Then we rename the columns and obtain the scatter plot and correlation matrices.

```
data <- read.table("Homework_4_data_Problem1.txt", header = FALSE, as.is =TRUE)
names(data) <- c("Rental_rates","X1", "X2", "X3", "X4")

library(graphics)
cor(data)
```

```
##              Rental_rates         X1         X2          X3         X4
## Rental_rates   1.00000000 -0.2502846  0.4137872  0.06652647 0.53526237
## X1            -0.25028456  1.0000000  0.3888264 -0.25266347 0.28858350
## X2             0.41378716  0.3888264  1.0000000 -0.37976174 0.44069713
## X3             0.06652647 -0.2526635 -0.3797617  1.00000000 0.08061073
## X4             0.53526237  0.2885835  0.4406971  0.08061073 1.00000000
```

```
pairs(data)
```



Based on the correlation matrix we notice that Total Sq. footage and Rental Rates have the largest positive correlation, while the Rental Rates and Age are negatively correlated. We also note that Total Square Footage is somewhat correlated with Operating expanses and taxes, and with Age, so problems with multicolinearity may arise.

(b)(5p) Fit regression model below for four predictor variables to the data and state the estimated regression function.

$$Y_i = \beta_0 + \beta_1 * X_{i1} + \beta_2 * X_{i2} + \beta_3 * X_{i3} + \beta_4 * X_{i4} + \epsilon_i$$

Considering a multiple linear regression model of the form , we compute the estimated regression coefficients using the function lm.

```
X1 <- data$X1
X2 <- data$X2
X3 <- data$X3
X4 <- data$X4

lm1 <- lm(data$Rental_rates ~ X1 + X2 + X3 + X4)
beta_0 <- lm1$coefficients[1]
beta_1 <- lm1$coefficients[2]
beta_2 <- lm1$coefficients[3]
beta_3 <- lm1$coefficients[4]
beta_4 <- lm1$coefficients[5]
```

```r
names(lm1$coefficients) <- c("Intercept","B1","B2","B3","B4")
lm1$coefficients[1]
```

```
## Intercept
##  12.20059
```

```r
lm1$coefficients[2]
```

```
##         B1
## -0.1420336
```

```r
lm1$coefficients[3]
```

```
##        B2
## 0.2820165
```

```r
lm1$coefficients[4]
```

```
##        B3
## 0.6193435
```

```r
lm1$coefficients[5]
```

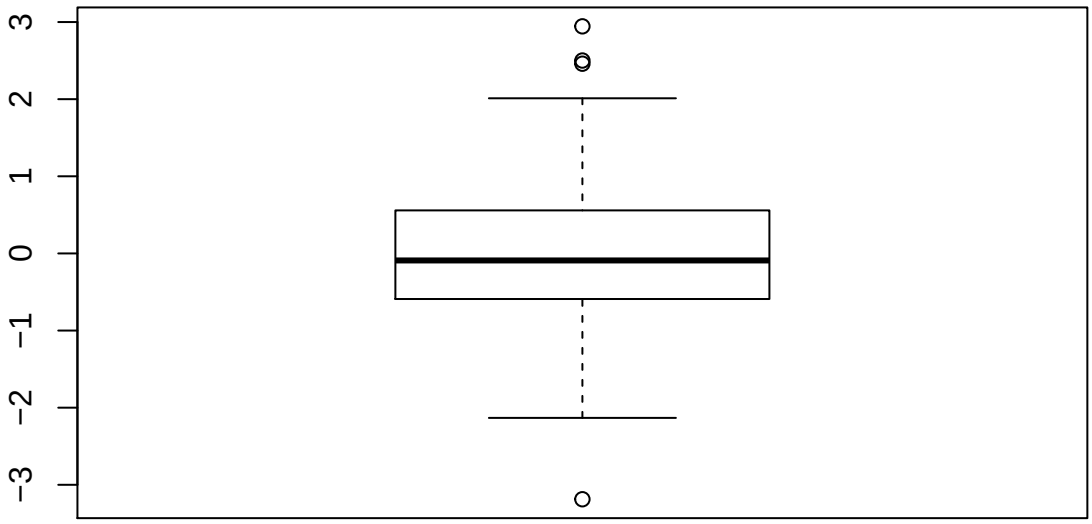
```
##           B4
## 7.924302e-06
```

We finally obtain the estimated regression function

$$\hat{Y_i} = 12.2 - 0.142 * X_{i1} + 0.282 * X_{i2} + 0.619 * X_{i3} + 0.00000792 * X_{i4}$$

(c)(5p) Obtain the residuals and prepare a box plot of the residuals. Does the distribution appear to be fairly symmetrical?

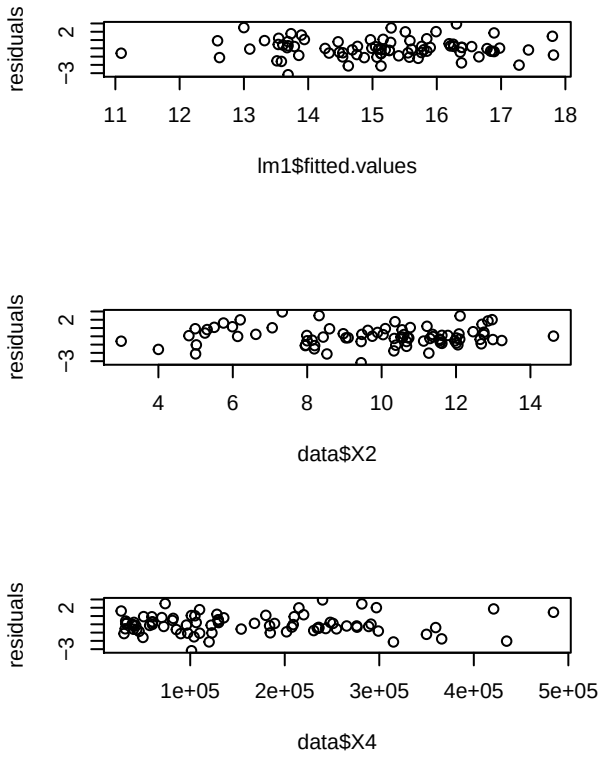
```r
residuals <- lm1$residuals
boxplot(residuals)
```



We create a boxplot of the residuals and we observe that the distribution seems to be fairly symmetrical around the median.

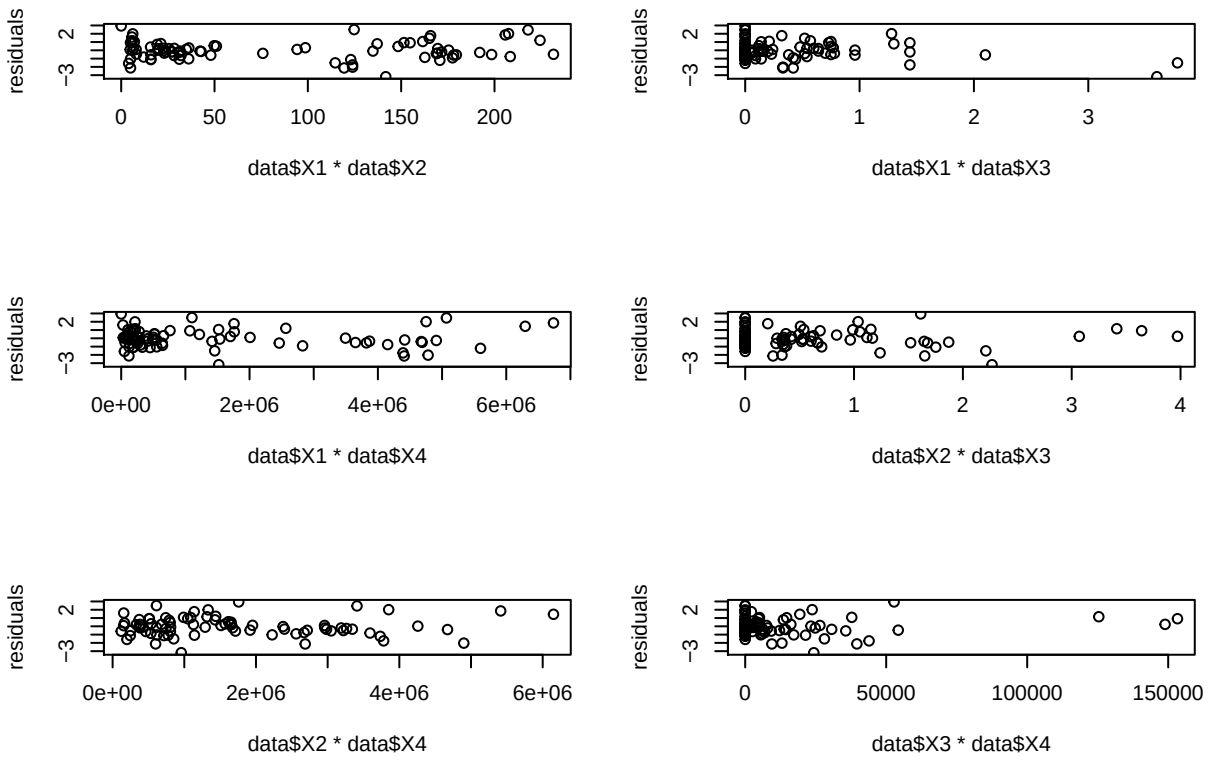
(d)(5p) Plot the residuals against the fitted line, each predictor variable, and each two-factor interaction term on separate graphs. Also prepare a normal probability plot. Analyse your plots and summarize your findings.

```
par(mfrow=c(3,2))
plot(lm1$fitted.values,residuals)
plot(data$X1,residuals)
plot(data$X2, residuals)
plot(data$X3, residuals)
plot(data$X4, residuals)
```



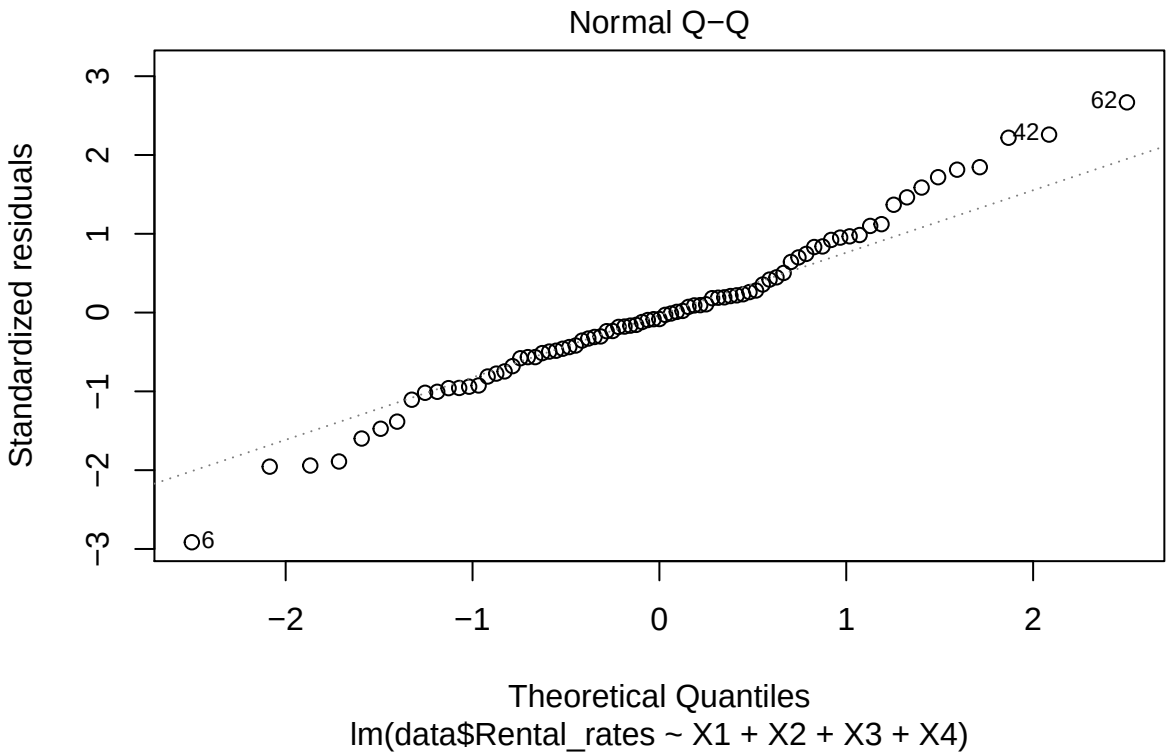
We firt plotted the fitted values against the residuals. The plot does not suggest any systematic deviation from the response hyperplane, and there is no sign of nonconstancy of the error variance. This conclusion is consistent with the plots of each predictor variable versus residuals.

```
par(mfrow=c(3,2))
plot(data$X1*data$X2,residuals)
plot(data$X1*data$X3,residuals)
plot(data$X1*data$X4,residuals)
plot(data$X2*data$X3, residuals)
plot(data$X2*data$X4, residuals)
plot(data$X3*data$X4, residuals)
```



We plotted the residuals against each two-factor $X_y * X_z$ interaction term. The plots do not exhibit any systematic pattern, therefore no interaction effects reflected by the term $\beta_5 * X_y * X_z$ appear to be present.

```
plot(lm1, which=2)
```



The QQ plot above is moderately linear so we can infer that the distribution of residuals is approximately normal with heavy tails.

(e)(5p) Can you conduct a formal test for lack of fit here?

We cannot conduct a formal test for lack of fit here because we there are no replicate observations on Y corresponding to levels of each of the predictor variables kept constant across trials.

(f)(10p) The commercial real estate company obtained information about additional three properties. Find separate prediction intervals for the rental rates for each of the new properties. Use 95% confidence coefficient in each case. Can the rental rates of these three properties be predicted fairly precisely? What is the family confidence level for the set of three predictions?

```
newdata1 <- data.frame(X1 = 4, X2=10,X3=0.1,X4=80000)
newdata2 <- data.frame(X1 = 6, X2=11.5,X3=0,X4=120000)
newdata3 <- data.frame(X1 = 12, X2=12.5,X3=0.32,X4=340000)

predict.lm(lm1, newdata =newdata1, interval = "prediction", level=0.95)
```

```
##       fit      lwr     upr
## 1 15.1485 12.85249 17.4445
```

```
predict.lm(lm1, newdata =newdata2, interval = "prediction", level=0.95)
```

```
##        fit      lwr      upr
## 1 15.54249 13.24504 17.83994
```

```
predict.lm(lm1, newdata =newdata3, interval = "prediction", level=0.95)
```

```
##        fit      lwr      upr
## 1 16.91384 14.53469 19.29299
```

Hence for the first properties the rate is between 12.85 and 17.444. For the second property it is between 13.25 and 17.84. For the third property it is between 14.54 and 19.29. The prediction interval for each of these properties is quite large so we cannot make a precise prediction.

(g)(10p)Obtain the analysis of variance table that decomposes the regression sum of squares into extra sums of squares associated with X4; with X1 given X4; with X2 given X1 and X4; and with X3, given X1, X2 and X4.

```
lm2 <- lm(data$Rental_rates ~ data$X4 + data$X1 + data$X2 + data$X3)
anova(lm2)
```

```
## Analysis of Variance Table
##
## Response: data$Rental_rates
##           Df Sum Sq Mean Sq F value    Pr(>F)
## data$X4    1 67.775  67.775 52.4369 3.073e-10 ***
## data$X1    1 42.275  42.275 32.7074 2.004e-07 ***
## data$X2    1 27.857  27.857 21.5531 1.412e-05 ***
## data$X3    1  0.420   0.420  0.3248    0.5704
## Residuals 76 98.231   1.293
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We read that SSR associated with X4 is 67.775, SSR of X1 given X4 is 42.275, SSR of X2 given X1 and X4 is 27.857, and SSR of X3 given X1, X2, and X4 is 0.42.

(h)(5p) Test whether X3 can be dropped from the regression model given that X1, X2, and X4 are retained. Use the F* test statistic and level of significance 0.01. State the alternatives decision rule, and conclusion. What is the p-value of the test?

We test the hypothesis H0 : $\beta_3 = 0$ versus the alternative hypothesis Ha : $\beta_3 \neq 0$. To do so we need to compute the reduced linear regression model that does not contain the predictor variable X3. Then we can use the general linear test statistic F* = (SSE(R)-SSE(F))/(dfR-dfF)/(SSE(F)/dfF) with level of confidence $\alpha = 0.01$.

The decision rule is if F* <= F(1 - alpha ; dfR - dfF, dfF) then we conclude H0. If F* > F(1 - alpha ; dfR - dfF, dfF) then we conclude Ha.

```
lm2 <- lm(data$Rental_rates ~ X1 + X2 + X4)

SSE_F <- sum((lm1$residuals)^2)
SSE_R <- sum((lm2$residuals)^2)
df_F <- lm1$df.residual
```

```
df_R <- lm2$df.residual

F_test <- ((SSE_R-SSE_F)/(df_R-df_F))/(SSE_F/df_F)
F_test
```

```
## [1] 0.3247534
```

```
F_test_crit<- qf(0.99, df1=1, df2=df_F)
F_test_crit
```

```
## [1] 6.980578
```

```
p_value <- 1-pf(F_test,1,df_F)
p_value
```

```
## [1] 0.5704457
```

The critical F score is equal to 6.98. The F statistic is equal to 0.32 which is less than 6.98 so we conclude H0 meaning that $\beta_3 = 0$. The p-value is equal to 0.5704 which is clearly above 0.01 so it makes sense to conclude the null hypothesis H0 : $\beta_3 = 0$ at the 1% level of confidence.

Problem 2 (Problems 8.15 & 8.19 in ALRM book)

The users of the copiers are either training institutions that use a small model, or business firms that use a large, commercial model. An analyst at Tri-City wishes to fit a regression model including both number of copiers serviced (X1) and type of copier (X2) as predictor variables and estimate the effect of copier model (S-small, L-large) on number of minutes spent on the service call. Assume that the regression model

$$Y_i = \beta_0 + \beta_1 * X_{i1} + \beta_2 * X_{i2} + \epsilon_i$$

is appropriate, and let X2 = 1 if small model and 0 if large, commercial model.

(a)(5p) Explain the meaning of all regression coefficients in the model.

β_0 is the intercept of the regression model. It predicts the value of the response variable when the variables X1 and X2 are e=both equal to zero, meaning when the copier is large and there are zero copiers serviced. However in this situation if zero copiers are serviced then regardless of the copier model, we would expect that zero minutes are spent on the service call. Therefore β_0 does not have a meaningful interpretation here.

β_1 is the quantitative regression coefficient associated with the numbers of copiers serviced. If we fix X2, the other varibale in the model, then for each unit increment of X1, the number of minutes spent on the service call will increase by β_1.

β_2 is the qualitative regression coefficient associated with the type copiers serviced (X2 = 1 if small model and 0 if large). If we fix X1, the other varibale in the model, then the number of minutes spent on the service call will increase either by 0 or by β_2 respectively if the copier is large or small.

(b)(5p) Fit the regression model and state the estimated regression function.

```
minutes <- read.table("Homework_4_data_Problem2.txt", header = FALSE, as.is =TRUE)
names(minutes) <- c("Y","X1", "X2")
```

```
Y  <- minutes$Y
X1 <- minutes$X1
X2 <- minutes$X2

lm3 <- lm(Y ~ X1 + X2)
beta_0 <- lm3$coefficients[1]
beta_1 <- lm3$coefficients[2]
beta_2 <- lm3$coefficients[3]


names(lm3$coefficients) <- c("Intercept","B1","B2")
lm3$coefficients[1]
```

```
##  Intercept
## -0.9224729
```

```
lm3$coefficients[2]
```

```
##       B1
## 15.04614
```

```
lm3$coefficients[3]
```

```
##        B2
## 0.7587218
```

We finally obtain the estimated regression function

$$\hat{Y}_i = -0.92 + 15.05 * X_{i1} + 0.76 * X_{i2}$$

(c)(5p) Estimate the effect of copier model on mean service time with a 95 percent confidence interval. Interpret your interval estimate.

```
confint(lm3, level = .95)
```

```
##               2.5 %    97.5 %
## Intercept -7.177891  5.332945
## B1        14.057283 16.035004
## B2        -4.851254  6.368698
```

This means that with confidence 95% we estimate that the mean number of minutes spent either decreases by 4.85 minutes or increases by 6.37 minutes if the copier is small.

(d)(10p) Why would the analyst wish to include X1, number of copiers, in the regression model when interest is in estimating the effect of type of copier model on service time?

```
cor(minutes)
```

```
##              Y          X1          X2
## Y   1.00000000  0.97851698 -0.07107219
## X1  0.97851698  1.00000000 -0.08146852
## X2 -0.07107219 -0.08146852  1.00000000
```

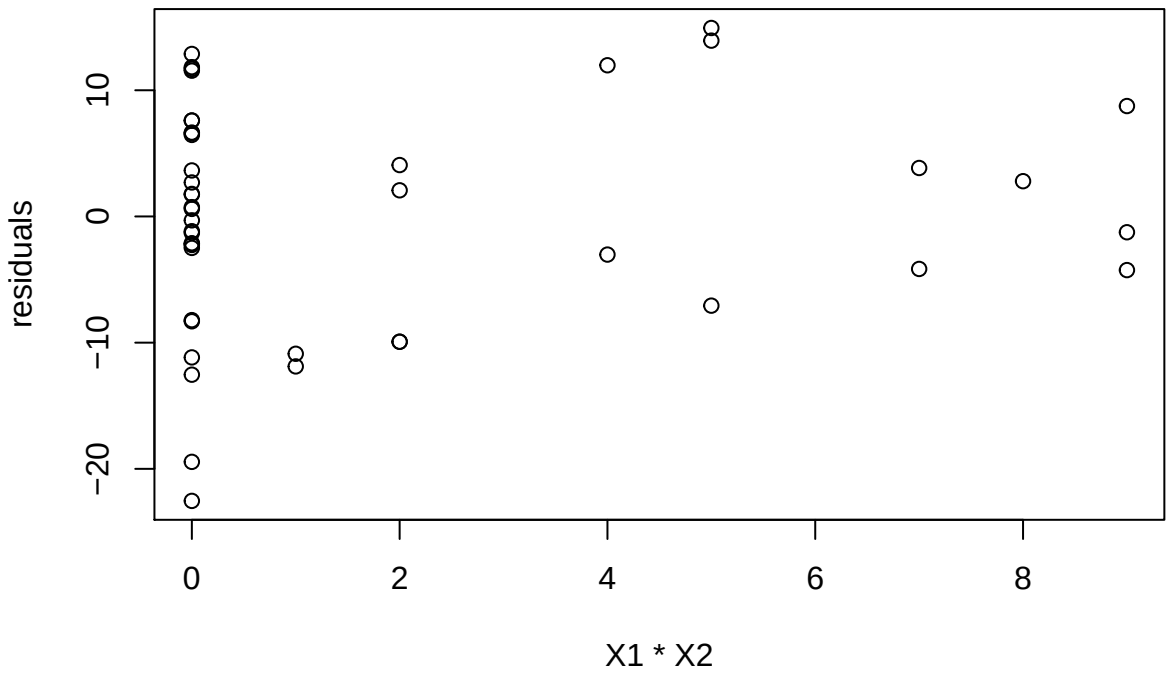
```r
lm4 <- lm(Y ~  X2) #Regression model excluding X1
names(lm4$coefficients) <- c("Intercept","B2")
confint(lm4, level = .95) #Confidence interval of Beta2 when excluding X1
```

```
##                2.5 %   97.5 %
## Intercept  62.17130 95.04299
## B2        -32.93615 20.54539
```

From the correlation matrix above we see that X1, number of copiers, is strongly correlated with the response variable Y. Therefore X1 explains the majority of the response variable. However X2 is very lightly (negatively) correlated with Y so it doesn't explain a lot for the response variable Y. Hence when we exclude X1, the confidence interval of β_2 becomes excessively large and it is more difficult to estimate the effect of X2 on Y. We condlude that including X1 allows to make a narrower estimate of the confidence interval of the effect of X2 on Y.

(e)(10p) Obtain the residuals and plot them against X1X2. Is there any indication that an interaction term in the regression model would be helpful?

```r
residuals <- lm3$residuals
plot(X1*X2,residuals)
```



From the plot of the interaction term against the residuals we do not observe any systematic pattern, therefore no interaction effects reflected by the term $\beta_3 * X_{i1} * X_{i2}$ appear to be present. Hence it would probably not be helpful to include this interaction term in the model.

(f)(5p) Fit regression model with interaction term as an additional explanatory variable, i.e.,

$$Y_i = \beta_0 + \beta_1 * X_{i1} + \beta_2 * X_{i2} + \beta_3 * X_{i1} * X_{i2} + \epsilon_i$$

```
X3 <- X1*X2

lm5 <- lm(Y ~ X1 + X2 + X3)

names(lm5$coefficients) <- c("Intercept","B1","B2", "B3")
lm5$coefficients[1]
```

```
## Intercept
##  2.813114
```

```
lm5$coefficients[2]
```

```
##       B1
## 14.33941
```

```
lm5$coefficients[3]
```

```
##        B2
## -8.141198
```

```
lm5$coefficients[4]
```

```
##       B3
## 1.777387
```

We finally obtain the estimated regression function

$$\hat{Y}_i = 2.81 + 14.34 * X_{i1} - 8.14 * X_{i2} + 1.78 * X_{i1} * X_{i2}$$

(g)(10p) Test whether the interaction term can be dropped from the model; control the alpha risk at 0.10. State the alternatives, decision rule, and conclusion. What is the p-value of the test? If the interaction the cannot be dropped from the model, describe the nature of the interaction effect.

We test the hypothesis H0 : $\beta_3 = 0$ versus the alternative hypothesis Ha : $\beta_3 \neq 0$. To do so we need to compute the reduced linear regression model that does not contain the predictor variable X3. Then we can use the general linear test statistic F* = (SSE(R)-SSE(F))/(dfR-dfF)/(SSE(F)/dfF) with level of confidence $\alpha = 0.01$.

The decision rule is if F* <= F(1 - alpha ; dfR - dfF, dfF) then we conclude H0. If F* > F(1 - alpha ; dfR - dfF, dfF) then we conclude Ha.

```
SSE_F <- sum((lm5$residuals)^2)
SSE_R <- sum((lm3$residuals)^2)
df_F <- lm5$df.residual
df_R <- lm3$df.residual

F_test <- ((SSE_R-SSE_F)/(df_R-df_F))/(SSE_F/df_F)
F_test
```

```
## [1] 3.325989
```

```
F_test_crit<- qf(0.99, df1=1, df2=df_F)
F_test_crit
```

```
## [1] 7.29638
```

```
p_value <- 1-pf(F_test,1,df_F)
p_value
```

```
## [1] 0.07548825
```

The critical F score is equal to 7.30. The F statistic is equal to 3.32 which is less than 7.30 so we conclude H0 meaning that $\beta_3 = 0$. The p-value is equal to 0.075 which is above 0.01 so it makes sense to conclude the null hypothesis H0 : $\beta_3 = 0$ at the 1% level of confidence.