Prediction and Time Series Computer Exercise Week 2

**Name: Nguyen Xuan Binh Student ID: 887799**

**EXERCISE 2.3**

Continuation to Exercise 2.2. Use backward elimination to choose the model. Perform the backward elimination using the permutation test. You may utilize lecture slides and demo exercises of the previous week. Compare results with part (b) of Problem 2.2. Use level of significance α = 5%.

In backward elimination, the first step is to estimate the full model and examine statistical significance of the explanatory variables. The least significant variable is removed from the model and after that, a new model is estimated. Variables are removed from the model one at a time, until all remaining variables are statistically significant.

hald=read.table("hald.txt",header=T)  
fullmodel=lm(HEAT~CHEM1+CHEM2+CHEM3+CHEM4,data=hald)  
summary(fullmodel)

##   
## Call:  
## lm(formula = HEAT ~ CHEM1 + CHEM2 + CHEM3 + CHEM4, data = hald)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.1750 -1.6709 0.2508 1.3783 3.9254   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 62.4054 70.0710 0.891 0.3991   
## CHEM1 1.5511 0.7448 2.083 0.0708 .  
## CHEM2 0.5102 0.7238 0.705 0.5009   
## CHEM3 0.1019 0.7547 0.135 0.8959   
## CHEM4 -0.1441 0.7091 -0.203 0.8441   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.446 on 8 degrees of freedom  
## Multiple R-squared: 0.9824, Adjusted R-squared: 0.9736   
## F-statistic: 111.5 on 4 and 8 DF, p-value: 4.756e-07

As seen from exercise 2.2, the variable HEAT correlates strongly with all explanatory candidates from the summary above. However, the CHEM elements are not assumed to be normally distributed, so instead of AIC model, we should use the permutation test

Now permutation tests are carried out for each models of different explanatory variables. We will start with the model.

HEAT = β0 + β1CHEM1 + β2CHEM2 + β3CHEM3 + β4CHEM4 + ε (1)

n\_permute <- 2000  
alpha <- 0.05  
tested\_vars = c("CHEM1","CHEM2","CHEM3", "CHEM4") # Choose Chem1 to Chem4  
rSqr\_full = summary(fullmodel)$r.squared  
  
# Explanatory variables: 4 chems  
X <- as.matrix(cbind(rep(1, nrow(hald)), hald[,c("CHEM1","CHEM2","CHEM3", "CHEM4")]))  
# Response variable heat  
y <- hald$HEAT  
  
fit\_helper\_i <- function(tested\_var, X,y) {  
 # Sample with replacement  
 # In testing significance of a explanatory variable, only the column of the variable is permuted, the rest stays unchanged  
 X[,tested\_var] <- sample(X[,tested\_var])  
 # plus 1 in the index above because X is combined with vector of 1s in  
 # X <- as.matrix(cbind(rep(1, nrow(hald)), hald[,c(1,2,3,4)]))  
 y\_fitted <- X %\*% solve((t(X) %\*% X)) %\*% t(X) %\*% y  
 cor(y\_fitted,y)^2  
}  
  
rSqr <-sapply(tested\_vars, function(tested\_var) replicate(n\_permute, fit\_helper\_i(tested\_var,X,y)))  
  
p\_value <- apply(rSqr, 2, function(col\_rSqr) sum(col\_rSqr > rSqr\_full)/length(col\_rSqr))  
  
t(cbind(colnames(hald)[c(1,2,3,4)], p\_value))

## CHEM1 CHEM2 CHEM3 CHEM4   
## "CHEM1" "CHEM2" "CHEM3" "CHEM4"  
## p\_value "0.0725" "0.4805" "0.89" "0.851"

We see that p\_value of Chem3 is the largest and it is bigger than 0.05 => We need to drop it from our model. Now the new model is

HEAT = β0 + β1CHEM1 + β2CHEM2 + β4CHEM4 + ε (2)

n\_permute <- 2000  
alpha <- 0.05  
tested\_vars = c("CHEM1","CHEM2","CHEM4")   
rSqr\_full = summary(fullmodel)$r.squared

# Explanatory variables: 4 chems  
X <- as.matrix(cbind(rep(1, nrow(hald)), hald[,c("CHEM1","CHEM2", "CHEM4")]))  
# Response variable heat  
y <- hald$HEAT  
  
fit\_helper\_i <- function(tested\_var, X,y) {  
 # Sample with replacement  
 # In testing significance of a explanatory variable, only the column of the variable is permuted, the rest stays unchanged  
 X[,tested\_var] <- sample(X[,tested\_var])  
 # plus 1 in the index above because X is combined with vector of 1s in  
 # X <- as.matrix(cbind(rep(1, nrow(hald)), hald[,c(1,2,3,4)]))  
 y\_fitted <- X %\*% solve((t(X) %\*% X)) %\*% t(X) %\*% y  
 cor(y\_fitted,y)^2  
}  
  
rSqr <-sapply(tested\_vars, function(tested\_var) replicate(n\_permute, fit\_helper\_i(tested\_var,X,y)))  
  
p\_value <- apply(rSqr, 2, function(col\_rSqr) sum(col\_rSqr > rSqr\_full)/length(col\_rSqr))  
  
t(cbind(colnames(hald)[c(1,2,4)], p\_value))

## CHEM1 CHEM2 CHEM4   
## "CHEM1" "CHEM2" "CHEM4"   
## p\_value "0" "0.052" "0.1965"

We see that p\_value of Chem4 is the largest and it is bigger than 0.05 => We need to drop it from our model. Now the new model is

HEAT = β0 + β1CHEM1 + β2CHEM2 + ε (3)

n\_permute <- 2000  
alpha <- 0.05  
tested\_vars = c("CHEM1","CHEM2")   
rSqr\_full = summary(fullmodel)$r.squared  
  
# Explanatory variables: 4 chems  
X <- as.matrix(cbind(rep(1, nrow(hald)), hald[,c("CHEM1","CHEM2")]))  
# Response variable heat  
y <- hald$HEAT  
  
  
fit\_helper\_i <- function(tested\_var, X,y) {  
 # Sample with replacement  
 # In testing significance of a explanatory variable, only the column of the variable is permuted, the rest stays unchanged  
 X[,tested\_var] <- sample(X[,tested\_var])  
 # plus 1 in the index above because X is combined with vector of 1s in  
 # X <- as.matrix(cbind(rep(1, nrow(hald)), hald[,c(1,2,3,4)]))  
 y\_fitted <- X %\*% solve((t(X) %\*% X)) %\*% t(X) %\*% y  
 cor(y\_fitted,y)^2  
}  
  
rSqr <-sapply(tested\_vars, function(tested\_var) replicate(n\_permute, fit\_helper\_i(tested\_var,X,y)))  
  
p\_value <- apply(rSqr, 2, function(col\_rSqr) sum(col\_rSqr > rSqr\_full)/length(col\_rSqr))  
  
t(cbind(colnames(hald)[c(1,2)], p\_value))

## CHEM1 CHEM2   
## "CHEM1" "CHEM2"  
## p\_value "0" "0"

Now, both the p\_values of CHEM1 and CHEM2 are smaller than 0.05 => The correct appropriate model is HEAT = β0 + β1CHEM1 + β2CHEM2 + ε (3) with two explanatory variables CHEM1 and CHEM2

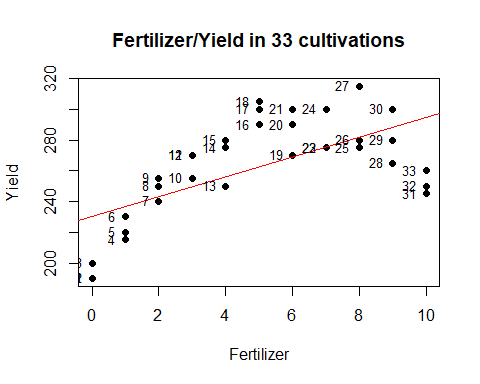
Compare results with part (b) of Problem 2.2: CHEM2 and CHEM4 are not statistically significant in Problem 2.2(b), while in this model, CHEM3 and CHEM4 are not statistically significant. This difference is due to the assumption that the explanatory variables are not normally distributed

**EXERCISE 2.4**

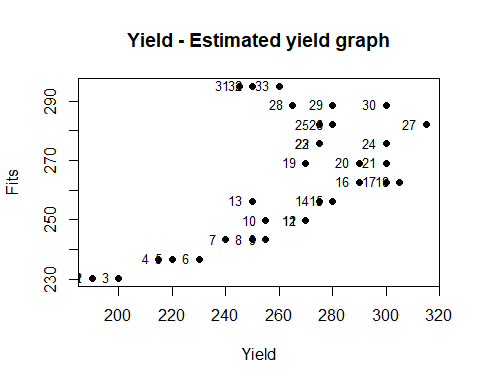
The quantity of a fertilizer affects the yield of wheat. The effect was studied by altering the quantity of the fertilizer (11 levels) in 33 different cultivations (the same amount of fertilizer in 3 cultivations) and by measuring the yield of each cultivation. Results of the study are given in the file crop.txt. The variables are, Yield = Yield (kg/unit of area) Fertilizer = the amount of the fertilizer (kg/unit of area)

**a) Estimate a linear regression model, where Yield is a response variable and Fertilizer is an explanatory variable. Using regression graphics, study whether the model is sufficient.**

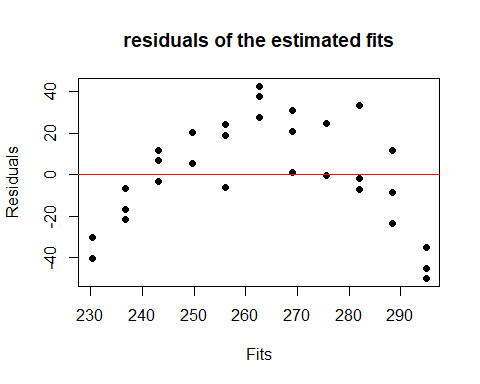
crop <- read.table("crop.txt",header=T,sep="\t")  
model1 <- lm(Yield~Fertilizer,data=crop)  
  
FIT <- model1$fit  
RES <- model1$res  
  
plot(crop$Fertilizer, crop$Yield, ylab="Yield",  
xlab="Fertilizer", pch=16,  
main="Fertilizer/Yield in 33 cultivations")  
abline(model1,col="red")  
text(crop$Fertilizer, crop$Yield, labels=1:33, cex= 0.8,  
pos=2)



plot(crop$Yield,FIT, ylab="Fits",xlab="Yield",pch=16,main="Yield - Estimated yield graph")  
text(crop$Yield, FIT, labels=1:33, cex= 0.8, pos=2)



plot(FIT,RES, xlab="Fits",ylab="Residuals",pch=16, main="residuals of the estimated fits")  
abline(h=0, col="red")



According to the regression graphics above, we can see that the linear model seems to be inappropriate. The residuals are distributed significantly far away from y=0. Also, there are too many outliers in the residuals/fits graph as well as in the Fertilizer/Yield linear regression model => This model is insufficient in demonstrating the affect of fertilizers on crops

**b) Estimate a linear regression model, where you have added the explanatory variable**

**LSqrd = Fertilizer · Fertilizer**

**to the model of the part a). That is, LSqrd consists of the squared elements of the variable Fertilizer. Using regression graphics, study whether the model is sufficient.**

Now we calculate the variable LSqrd and add it to the explanatory variables of Yield together with fertilizer

fertilizer = crop$Fertilizer  
LSqrd <- fertilizer \* fertilizer  
  
model2 <- lm(Yield ~ fertilizer + LSqrd, data=crop)

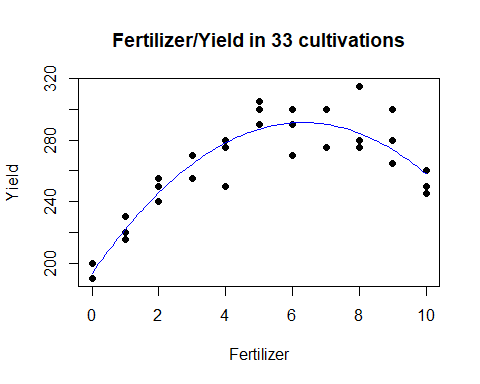
From the model above, we notice that the data points are distributed in a parabolic line => a curve with the coefficients found from the linear model calculated above will be needed

summary(model2)

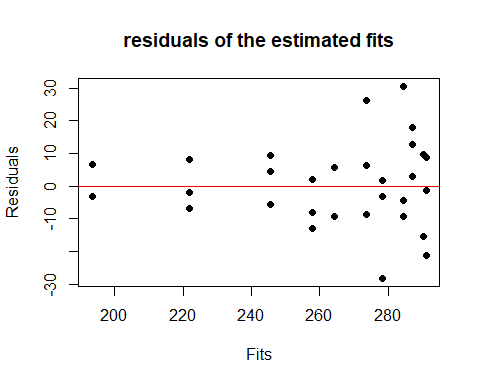
##   
## Call:  
## lm(formula = Yield ~ fertilizer + LSqrd, data = crop)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -28.256 -8.007 -1.196 6.690 30.554   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 193.3100 5.6511 34.207 < 2e-16 \*\*\*  
## fertilizer 31.0812 2.6292 11.821 8.11e-13 \*\*\*  
## LSqrd -2.4611 0.2532 -9.719 8.85e-11 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 12.85 on 30 degrees of freedom  
## Multiple R-squared: 0.8559, Adjusted R-squared: 0.8463   
## F-statistic: 89.07 on 2 and 30 DF, p-value: 2.407e-13

The multiple linear model YIELD = β0 + β1 \* fertilizer + β2 \* LSqrd + ε Besides, we know that LSqrd = fertilzer^2 => YIELD = β0 + β1 \* fertilizer + β2 \* fertilzer^2 + ε From the summary above, the coefficient estimates are β0 = 193.3100 β1 = 31.0812 β2 = -2.4611 Now we will plot the parabolic curve against the Fertilizer/Yield model

plot(crop$Fertilizer, crop$Yield, ylab="Yield", xlab="Fertilizer", pch=16, main="Fertilizer/Yield in 33 cultivations")  
curve((-2.4611\*x^2 + 31.0812\*x + 193.31), col="blue", from=0, to=10, ylab="Yield", xlab="Fertilizer",add=TRUE)

   
=> The parabolic curve fits the model really well. Now we examine its Fits/Residuals plot

FIT <- model2$fit  
RES <- model2$res  
plot(FIT,RES, xlab="Fits",ylab="Residuals",pch=16, main="residuals of the estimated fits")  
abline(h=0, col="red")



The residuals are distributed close to y = 0, which means that the model is quite appropriate, albeit its heteroscedastic distribution towards the right => Overall, this model is sufficient

**c) Compare the results obtained in parts a) and b). Which of the models is more suitable here?**

From the results of a) and b), it is clear than the model in b) is more suitable for the linear regression Yield/Fertilizer