

Estimation of the total surface occupied by fruit trees in a Region of Navarra

UGARTE, M.D.(*)

(*)DEPARTAMENTO DE ESTADÍSTICA E I. O., UNIVERSIDAD PÚBLICA DE NAVARRA, PAMPLONA, SPAIN

E-MAIL: LOLA@UNAVARRA.ES

Material from the book *Probability and Statistics with R*
by Ugarte, Militino, and Arnholt. Chapman and Hall/CRC, 2008

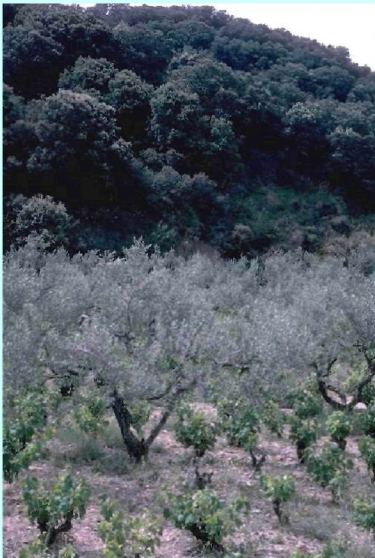
ÍNDICE

1. Introduction	1
2. Aim	5
3. Aplication	6
4. Data File Description	17
5. Linear Regression Model	20
6. Solution in R	21
7. Descriptive Analysis	22
8. Descriptive Analysis	25
9. Summary Models Comparison	48
10. Questions	58
11. Checking normality	60

12. Heteroscedasticity	61
13. Spatial Autocorrelation?	62
14. Coefficients of Confidence Intervals	65

Introduction

- Fruit trees is not a major crop in Navarra

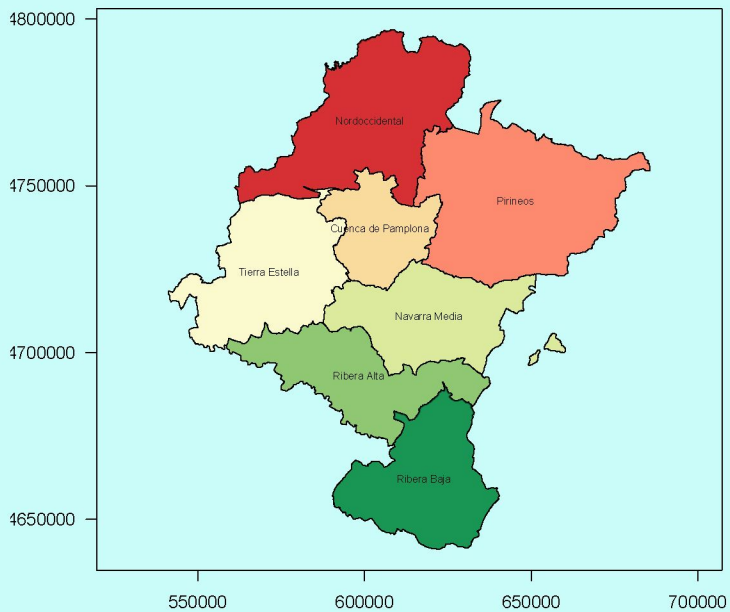


AUTONOMOUS REGIONS IN SPAIN



· NAVARRA

Comarcas Agrarias de Navarra



Aim

- This work aims to estimate the total area occupied by fruit trees in a region called Comarca VII, located in the South of Navarra, Spain, using as auxiliary information, classified data provided by satellite images.

Application

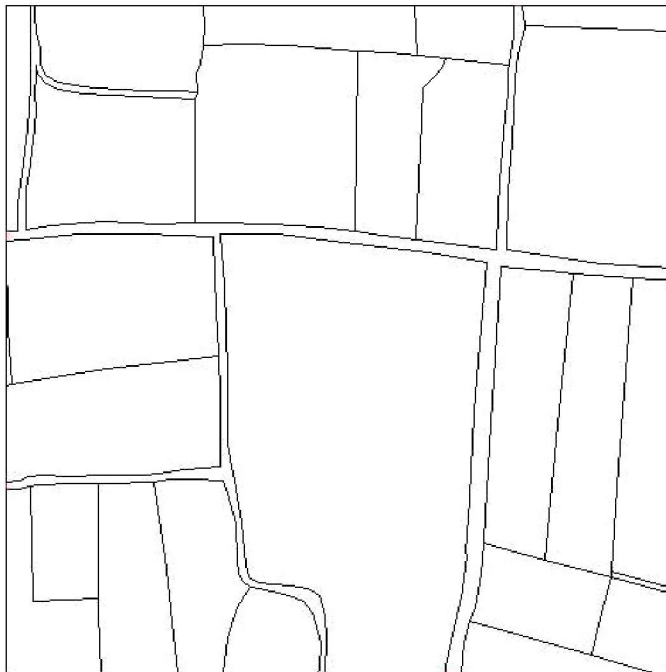
- Definition of the study domain (defined by using cropland maps depicted over past records, but recently updated in 1999 by the local Government)
- The sample consists of 47 segments of four hectares in three areas drawn by simple random sampling.
- The ground survey was carried out in July and August 2001 by an agricultural engineer
- The locations of the sampled segments were determined by a Navarra cropland map and several orthophoto maps provided by the local government

- Square segments are defined by overlapping a regular square grid on the area
- The surveyed segments were later digitized to weigh the land surface occupied by the sampled fruit trees and to integrate the information into the software of satellite images processing.
- In this work the satellite images were processed using ER Mapper 6.3 software
- The auxiliary information consists of classified fruit trees by remote sensing for the whole population of segments

ORTHOFOTO E:1:5000

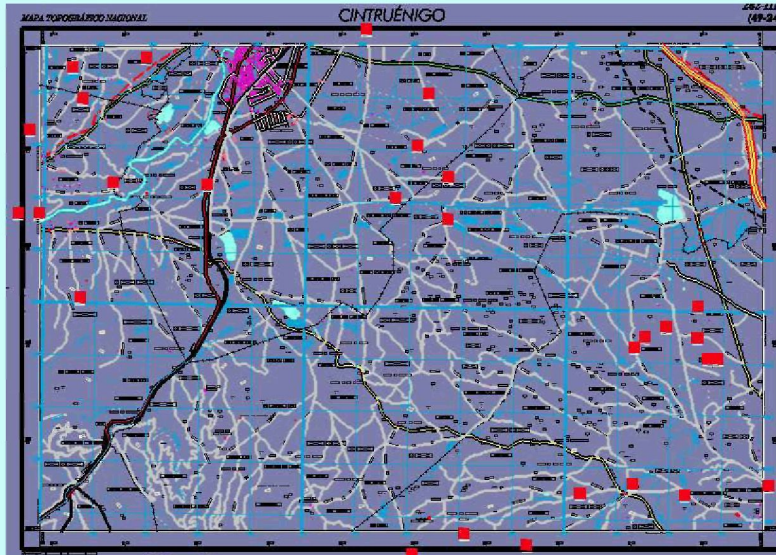


CATASTRO- LAND REGISTRY



60746544

TOPOGRAPHIC MAP

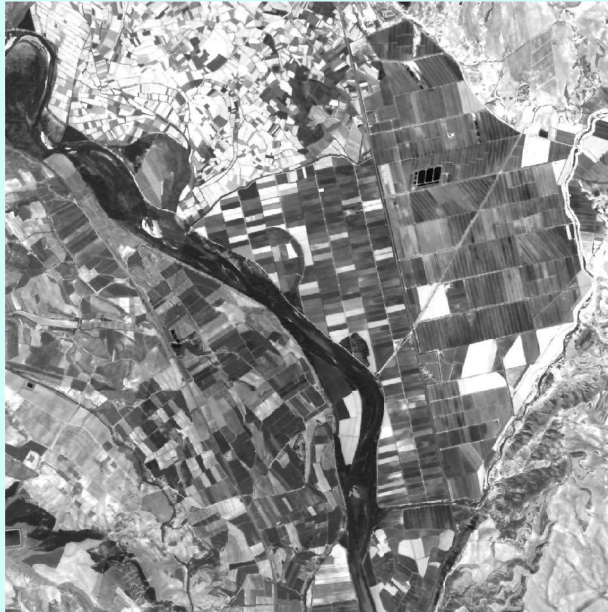


**Color Aerial Photograph with all of the segments in the sample using a grid E
1:10.000**



REMOTE SENSING

ALLOWS TO OBTAIN GROUND INFORMATION IN SEGMENTS OUT OF THE
SAMPLE Multispectral Image IKONOS

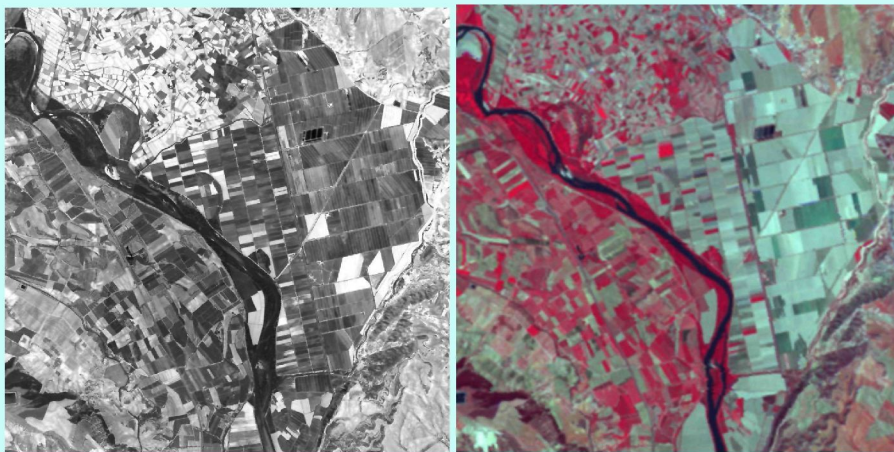


FUSION OF IMAGES

- Three multispectral images, collected in different seasons during 2001 were two Landsat 7ETM images, taken on April and November and one IRS LISS-III image taken on August.
- The ETM and LISS-III sensors, characterized by a high spectral resolution, do not show an optimal spatial resolution.
- The availability of high spectral and spatial resolution images is important when undertaking studies in highly parceled agricultural areas.
- First, a high spectral resolution eases discrimination of different land cover types and second, a high spatial resolution is necessary to delimit accurately the area occupied by each land cover type.
- Fusion of multispectral and panchromatic images, with complementary spectral and spatial characteristics, is a widely used technique for this aim.

- Three IRS Pan images, also collected during 2001, have been used to improve the spatial quality of the ETM and LISS-III multispectral images.
- Prior to being merged, all the images were ortho-rectified.
- Ortho-rectification is the process by which the geometric distortions of the image are modeled and accounted for, resulting in a planimetrically correct image.
- To preserve the spectral and radiometric information of the original multispectral images, the fusion method used in this work is based on the multiresolution wavelet transform

- Auxiliar Information: satellite images
 - multispectral and panchromatic images



SAMPLING



- Sample of 47 segments of 4 hectares (irrigated land). (Comarca VII)
- The study domain in three small areas.

Data File Description

- QUADRAT is the number of sampled segment or quadrat
- SArea is the small area
- WH is the classified surface of wheat in the sampled segment (in squared meters)
- BA is the classified surface of barley in the sampled segment (in squared meters)
- NAR is the classified surface of fallow or non arable land in the sampled segment
- COR is the classified surface of corn in the sampled segment
- SF is the classified surface of sunflower in the sampled segment
- VI is the classified surface of vineyard in the sampled segment

- PS is the classified surface of grass in the sampled segment
- ES is the classified surface of asparagus in the sampled segment
- AF is the classified surface of lucerne in the sampled segment
- CO is the classified surface of rape in the sampled segment
- AR is the classified surface of rice in the sampled segment
- AL is the classified surface of almonds in the sampled segment
- OL is the classified surface of olives in the sampled segment
- FR is the classified surface of fruit trees in the sampled segment
- OBS is the observed surface of fruit trees in the sampled segment

This is the content of the first 8 variables and 10 rows of file `satfruit`

```
>satfruit[1:10, 1:8]
```

	QUADRAT	SArea	WH	BA	NAR	COR	SF	VI
1	59106566	R68	0.00000	0	1933.912	0.0000	0.0000000	0.00000
2	59086560	R68	0.00000	0	1392.159	690.8583	0.0000000	399.05674
3	59406568	R68	0.00000	0	2026.149	0.0000	0.0000000	54.21483
4	59406562	R68	0.00000	0	1310.520	0.0000	0.0000000	0.00000
5	59486566	R68	0.00000	0	1684.034	203.6149	0.0000000	0.00000
6	59446566	R68	0.00000	0	3366.676	0.0000	0.0000000	68.70976
7	60006620	R68	0.00000	0	1596.651	0.0000	0.0000000	0.00000
8	59886642	R68	0.00000	0	1037.096	0.0000	0.0000000	0.00000
9	59846648	R68	41.79581	0	5090.172	1379.6287	0.1715065	0.00000
10	61286548	R63	0.00000	0	0.000	4042.8066	1058.1888445	48.04504

Linear Regression Model

$$y_{ij} = \beta_0 + \beta_1 x_{ij1} + \cdots + \beta_p x_{ijp} + \epsilon_{ij}, \quad i = 1, \dots, t, \quad j = 1, \dots, n_i$$

- ϵ_{ij} are the random errors $N(0, \sigma^2)$
- y_{ij} : fruit hectares in the j -th segment of the i -th area
- n_i is the number of sampled segments in i -th area
- t number of small areas
- x_{ijk} : classified crop hectares in the j -th segment of the i -th area
 $k = 1, \dots, p$

SOLUTION IN R

Load the library PASWR from the menu, type satfruit, calculate its dimension, and show the names of the variables contained in the file

```
library(PASWR)
attach(satfruit)
> dim(satfruit)
[1] 47 17
> names(satfruit)
 [1] "QUADRAT" "SArea"  "WH"     "BA"     "NAR"    "COR"    "SF"
 [8] "VI"      "PS"     "ES"     "AF"     "CO"     "AR"     "AL"
[15] "OL"      "FR"     "OBS"
```


Descriptive Analysis

- Do a descriptive analysis of data in file satfruit. Calculate the means, quartiles, and range of the numerical variables.
- What is the maximum number of m^2 of classified fruits by segment?
- How many observations are there by small area?

```
summary(satfruit) #descriptive analysis
```

QUADRAT	SArea	WH	BA
Min. :59086560	R63: 3	Min. : 0.00	Min. : 0.00
1st Qu.:60676695	R67:32	1st Qu.: 0.00	1st Qu.: 0.00
Median :61406658	R68:12	Median : 0.00	Median : 0.00
Mean :61087866		Mean : 78.36	Mean : 92.28
3rd Qu.:61656512		3rd Qu.: 0.00	3rd Qu.: 0.00
Max. :63006502		Max. :2377.70	Max. :3964.03

NAR	COR	SF	VI
Min. : 0.00	Min. : 0.0	Min. : 0.0	Min. : 0.00
1st Qu.: 77.18	1st Qu.: 0.0	1st Qu.: 0.0	1st Qu.: 0.00
Median : 508.41	Median : 0.0	Median : 0.0	Median : 0.00
Mean :1309.05	Mean : 761.1	Mean : 149.3	Mean : 36.18
3rd Qu.:1896.00	3rd Qu.: 292.3	3rd Qu.: 0.0	3rd Qu.: 0.00
Max. :5206.75	Max. :7123.1	Max. :5459.4	Max. :1128.25

PS	ES	AF	CO
Min. : 0.00	Min. : 0.00	Min. : 0.000	Min. : 0
1st Qu.: 0.00	1st Qu.: 0.00	1st Qu.: 0.000	1st Qu.: 0
Median : 0.00	Median : 0.00	Median : 1.827	Median : 0
Mean : 58.43	Mean : 64.76	Mean : 731.052	Mean : 100

AR	AL	OL	FR
Min. : 0.00	Min. : 0.0	Min. : 0.0	Min. : 0
1st Qu.: 0.00	1st Qu.: 0.0	1st Qu.: 0.0	1st Qu.: 4241
Median : 0.00	Median : 0.0	Median : 0.0	Median : 8536
Mean : 20.72	Mean : 489.9	Mean : 601.8	Mean : 7827
3rd Qu.: 0.00	3rd Qu.: 355.2	3rd Qu.: 569.3	3rd Qu.:11356
Max. :973.97	Max. :6745.3	Max. :6922.6	Max. :13969

OBS

```

Min. : 0
1st Qu.: 3382
Median : 7173
Mean : 7414
3rd Qu.:11563
Max. :13548

```

There are 3 observations in R63, 32 in R67 and 12 in R68.

The maximum number of classified fruits by segment is 13969 m^2 .

Descriptive Analysis

Use `pairs()` in R to explore the linear relationships between OBS and the remainder of the exploratory variables. Comment on the results

Type in R

```
> pairs(satfruit[,c(17:10)]) #scatterplot diagrams  
> pairs(satfruit[,c(17,9:3)])
```

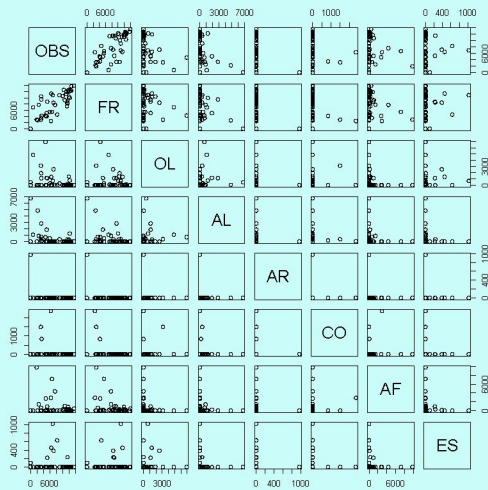


Figura 1: Scatterplot 1

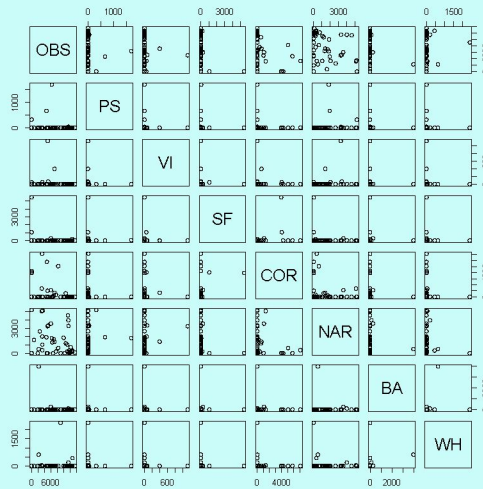


Figura 2: Scatterplot 2

Comments The linear relationship of OBS (number of observed hectares of fruit trees) with FR (number of classified hectares of fruit trees) is clear. Not so with the rest of the variables.

Use `histogram` from `library(lattice)` to show fruits histograms in each small area

Type in R

```
attach(satfruit) #attach the file satfruit for the whole session
library(lattice)
histogram(~OBS|SArea,as.table=TRUE)
```

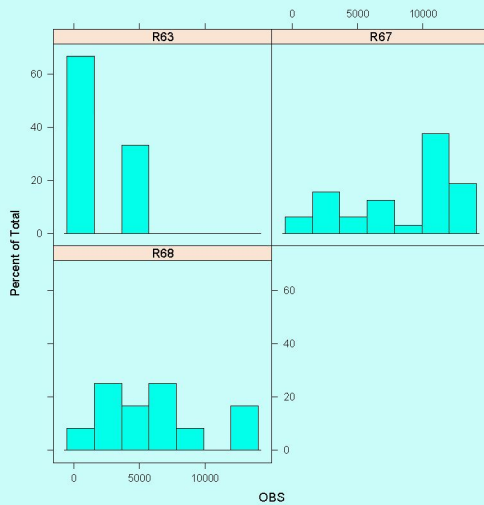


Figura 3: Histogram of OBS by Areas

Use `histogram` from `library(lattice)` to show classified fruits histograms in each small area

Type in R

```
library(lattice)
histogram(~FR|SArea, as.table=TRUE)
```

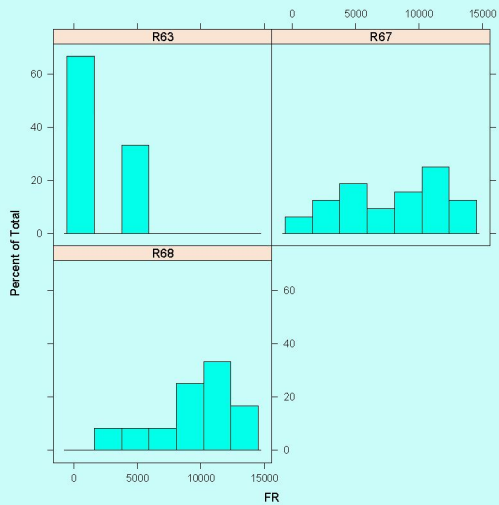


Figura 4: Histogram of FR by Areas

Use `boxplot` to show the observed fruits variability per areas and use `barplot` to show the observed fruits total per area and their standard errors

```
par(mfrow=c(2,2))
attach(datos)
boxplot(split(OBS,SArea),col="blue",main="Observed Fruits")
boxplot(split(FR,SArea),col="yellow",main="Classified Fruits")
```

```
medias<-sapply(split(OBS,SArea),mean)
des.e<-sapply(split(OBS,SArea),sd)
ee<-des.e/sqrt(table(SArea))
tabla<-rbind(medias,ee)
barplot(tabla,col=c("blue","red"),ylab="OBS",xlab="SArea",
legend=rownames(tabla),main="Observed Fruits Means")
```

```
medias<-sapply(split(FR,SArea),mean)
des.e<-sapply(split(FR,SArea),sd)
ee<-des.e/sqrt(table(SArea))
tabla<-rbind(medias,ee)
barplot(tabla,col=c("yellow","red"),ylab="FR",xlab="SArea",
legend=rownames(tabla),main="Classified Fruits Means")
```

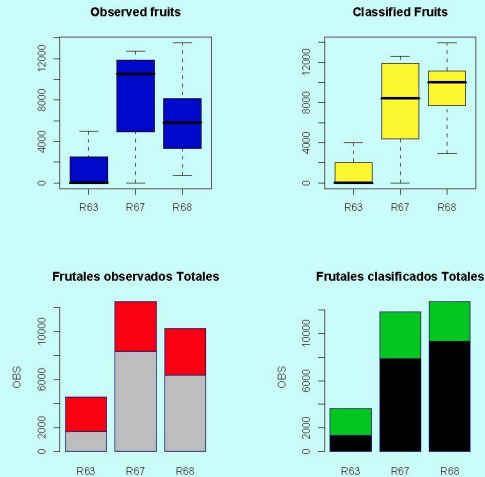


Figura 5: Boxplots of observed and classified fruits

Comments We observe a clear difference between the medians of the observed surfaces per small areas. The same occurs with the classified surfaces

Cuadro 1: Means and Sd of observed fruits

	variable	Freq	mean	sd
1	R63	3.00	1668.00	2889.00
2	R67	32.00	8346.00	4168.00
3	R68	12.00	6364.00	3886.00

Cuadro 2: Means and Sd of classified fruits

	variable	Freq	mean	sd
1	R63	3.00	1338.00	2317.00
2	R67	32.00	7861.00	3979.00
3	R68	12.00	9359.00	3363.00

Determine the variables with higher marginal correlation with OBS.

```
> round(cor(satfruit[,17], satfruit[,3:16]),2)
  WH    BA   NAR  COR   SF   VI   PS    ES   AF
0.05 -0.18 -0.24 -0.4 -0.3 -0.1 -0.11 -0.02 -0.17

  CO    AR   AL    OL   FR
-0.15 -0.26 -0.4 -0.29 0.82
```

Solution The marginal correlation of OBS with FR =0.82, with COR=-0.40, with AL=-0.4 and with SF=-0.3.

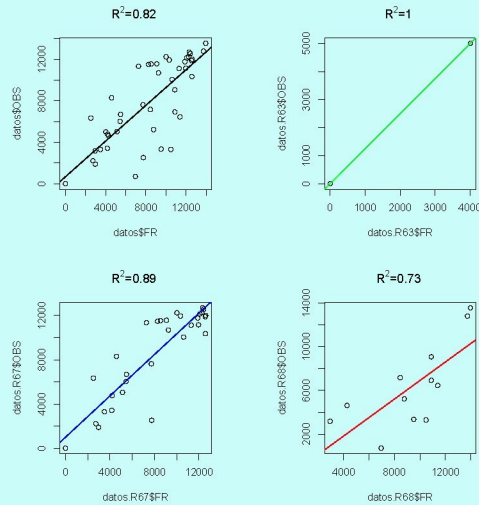


Figura 6: Linear regression of OBS vs. FR per small areas and coefficients of determination.

To make graphics in R

```
par(mfrow=c(2,2))
r2=summary(lm(satfruit$OBS~satfruit$FR))$r.squared
r=sqrt(r2)
plot(satfruit$FR, satfruit$OBS,main=expression(paste(plain(R^2),plain("=0.82"))))
abline(lsfite(satfruit$FR,satfruit$OBS),lwd=2,col=1)

datos.R63=satfruit[satfruit$SArea=="R63",]
cor(datos.R63$OBS,datos.R63$FR)
plot(datos.R63$FR,datos.R63$OBS,main=expression(paste(plain(R^2),plain("=1"))))
abline(lsfite(datos.R63$FR,datos.R63$OBS),col="green",lwd=2)

datos.R67=satfruit[satfruit$SArea=="R67",]
cor(datos.R67$OBS,datos.R67$FR)
plot(datos.R67$FR,datos.R67$OBS,main=expression(paste(plain(R^2),plain("=0.89"))))
abline(lsfite(datos.R67$FR,datos.R67$OBS),col="blue",lwd=2)
```



```
datos.R68=satfruit[satfruit$SArea=="R68",]  
cor(datos.R68$OBS,datos.R68$FR)  
plot(datos.R68$FR,datos.R68$OBS,main=expression(paste(plain(R^2),plain("=0.73"))))  
abline(lsfit(datos.R68$FR,datos.R68$OBS),col="red",lwd=2)
```

To make simultaneously graphics of linear regression and robust regression

```
\small{\begin{verbatim}  
panel.scatreg=function(x,y) {panel.xyplot(x,y)  
panel.abline(lm(y~x),col=1,lwd=2)  
panel.abline(lqs(y~x),col=3,lty=3,lwd=2)}  
xyplot(FR~OBS|SArea,as.table=T,panel=panel.scatreg)  
xyplot(FR~OBS,as.table=T,panel=panel.scatreg)
```

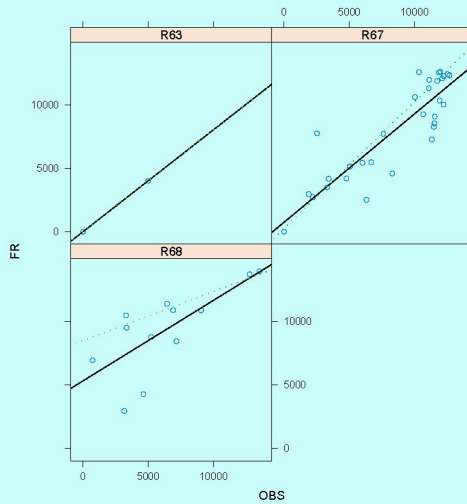


Figura 7: Linear regression of OBS vs. FR per small areas

Fit the linear regression model, called model (a) of OBS vs. the rest of the numerical variables in the same order as they are recorded in the file. Do the analysis of variance and decide what variables are statistically significant.

Type in R

```
model.A<-lm(OBS~WH+BA+NAR+COR+SF+VI+PS+ES+AF+CO+AR+AL+OL+FR)
```

```
> summary.aov(model.A)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
WH	1	2213285	2213285	0.3770	0.5435441	
BA	1	32652460	32652460	5.5621	0.0246272	*
NAR	1	49947383	49947383	8.5082	0.0064147	**
COR	1	197235474	197235474	33.5978	1.959e-06	***
SF	1	35592550	35592550	6.0630	0.0193748	*
VI	1	4630651	4630651	0.7888	0.3810904	
PS	1	13478087	13478087	2.2959	0.1395323	
ES	1	381673	381673	0.0650	0.8003693	
AF	1	66400430	66400430	11.3109	0.0020115	**
CO	1	2434603	2434603	0.4147	0.5241735	
AR	1	41940340	41940340	7.1443	0.0117361	*
AL	1	78135145	78135145	13.3098	0.0009301	***
OL	1	99650224	99650224	16.9748	0.0002497	***
FR	1	48852251	48852251	8.3217	0.0069554	**
Residuals	32	187855866	5870496			

```
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Solution: The statistically significant variables are BA, NAR, COR, SF, AF, AR, AL, OL, FR.

Compute (R^2 , R_{ajus}^2 AIC and BIC) for model (A)

Type in R

```
summary(model.A)
> summary(model.A)$r.squared
[1] 0.781918
> summary(model.A)$adj.r.squared
[1] 0.6865072
> AIC(model.A)
[1] 879.829
> AIC(model.A, k = log(nrow(datos)))
[1] 909.4314
```

Find the best regression model using `leaps` from library `leaps` and step to determine the best subset regression. Call them model (B) and (C) respectively.

Type in R

```
library(leaps)
a<-leaps(cbind(WH, BA, NAR, COR, SF, VI, PS, ES, AF, CO,
AR, AL, OL, FR), OBS, method="adjr2")

which(a$adjr2==max(a$adjr2))
#[1] 81
dim(a$which)
#[1] 131 14
a$which[81,]
  1     2     3     4     5     6     7
FALSE FALSE FALSE FALSE TRUE FALSE TRUE
  8     9     A     B     C     D     E
TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
```

The selected model using R^2 is the one that does not consider the variables: WH, BA, NAR, COR, and VI. If we fit this model, the result is

```
model.B<-lm(OBS~SF+PS+ES+AF+CO+AR+AL+OL+FR)
```

```
summary.aov(model.B)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
SF	1	76175219	76175219	14.7480	0.0004652	***
PS	1	12941875	12941875	2.5056	0.1219516	
ES	1	1597921	1597921	0.3094	0.5814168	
AF	1	29553573	29553573	5.7218	0.0219505	*
CO	1	9217582	9217582	1.7846	0.1897468	
AR	1	69983857	69983857	13.5493	0.0007367	***
AL	1	226458747	226458747	43.8439	9.101e-08	***
OL	1	108342642	108342642	20.9758	5.122e-05	***
FR	1	136019618	136019618	26.3343	9.384e-06	***
Residuals	37	191109387	5165119			

```
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Compute R^2 , R_a^2 , AIC, and BIC for model B

```
> summary(model.B)$r.squared
[1] 0.778141
> summary(model.B)$adj.r.squared
[1] 0.7241754
> AIC(model.B)
[1] 870.636
> AIC(model.B, k = log(nrow(satfruit))) #BIC criterion
[1] 890.9877
```


Selection of auxiliary variables using step

```
step(model.A)
```

```
Step:  AIC= 731.72
```

```
OBS ~ PS + AL + OL + FR
```

	Df	Sum of Sq	RSS	AIC
<none>			219296954	732
- AL	1	10196514	229493468	732
- PS	1	19596107	238893061	734
- OL	1	39717334	259014288	738
- FR	1	449327366	668624320	782

Solution: The function *leaps* select the variables SE, PS, ES, AF, CO, AR, AL, OL and FR. The function *step* select PS, AL, OL, and FR.

Fit a model -Model C- with the variables selected by *step* and compute AIC, and BIC

```
model.C<-lm(OBS~PS+AL+OL+FR)
summary.aov(model.C)
AIC(model.C)
AIC(model.C, k = log(nrow(satfruit))) #BIC criterion
```

Summary Models Comparison

Model	R^2	R^2_{ajus}	AIC	BIC
model (A)	0.78	0.69	880	909
model (B)	0.78	0.72	871	891
model (C)	0.75	0.72	867	878

With the variables selected with **leaps**, AIC=871 and with the variables selected with **step** AIC=867. So, we choose the model selected by **step** that is simpler.

Graph the default diagnostic regression plots of Model (C). Plot the standardized residuals, the student residuals, the Cook distances, the diagonal elements of the hat matrix, the DFFITS, and DFBETAS of Model (C).

```
#Default diagnostic plots in R

par(mfrow=c(2,2))
plot(model.C)
```

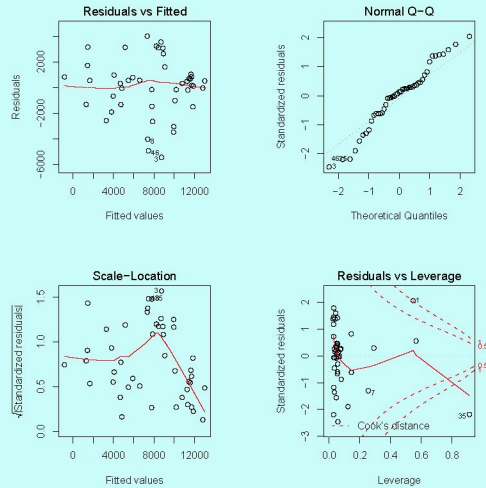


Figure 8: Default Diagnostic Plots. Model C

The diagonal elements of the hat matrix, the standardized residuals, and the studentized residuals of Model (C) can be computed in R as

```
# Hat values, residuals, observed versus fitted
a<-model.C
iden<-function(y, a = 3, c=0.05)
{
  n <- length(y)
  oy <- order(abs(y))
  b<-y*c
  which <- oy[(n - a + 1):n]
  text(seq(1:n)[which], y[which]+b[which], as.character(which))
  list(y=y,b=b) }
par(mfrow=c(2,2),pty="s")
plot(hatvalues(a),type="h",xlab="",ylim=c(0,1),
ylab="diagonales de la matriz hat")
X<-model.matrix(a)
abline(h=2*(ncol(X))/nrow(X))
iden(hatvalues(a))
title("a) Elementos
diagonales \n de la matriz hat")
```

```
plot(rstandard(a),type="n",xlab="",ylab="r_i")
text(rstandard(a))
title("b) Residuales estandarizados \n internamente")
```

```
plot(rstudent(a),type="n",xlab="",ylab="r_i^*")
text(rstudent(a))
abline(h=qt(0.025, a$df.residual-1))
abline(h=qt(0.975,a$df.residual-1))
title("c) Residuales estandarizados \n externamente")
```

```
prediccion<-predict.lm(a)
plot(prediccion, satfruit$OBS,xlab="v. ajustados",ylab="y",type="n",
main="d) Observados vs. ajustados")
abline(0,1)
text(prediccion, satfruit$OBS)
```

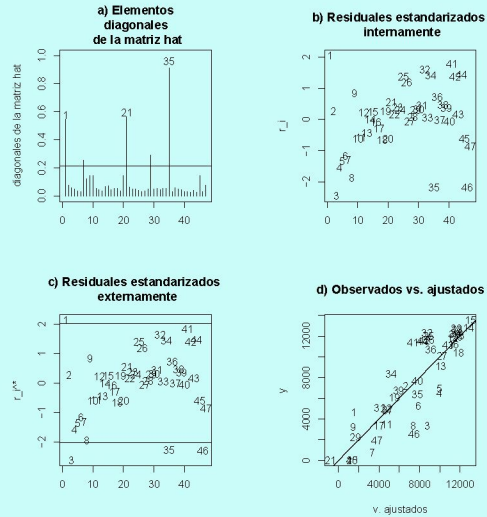


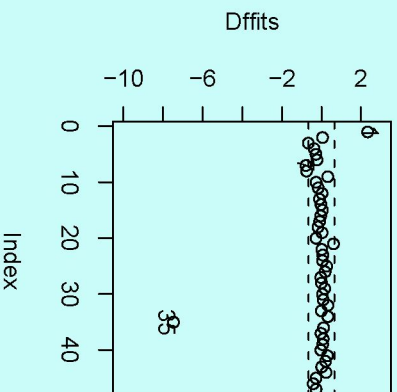
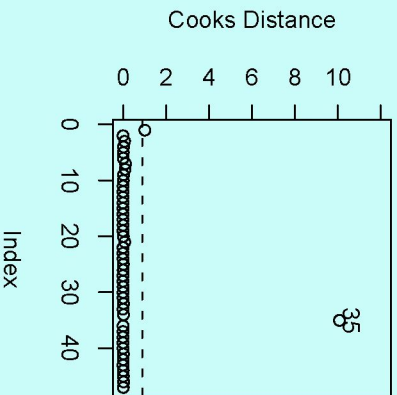
Figura 9: Diagnostic Plots. Model C

To compute in R the Cook distances, the DFFITS, and DFBETAS of Model (C)

```
# Cook distance and Dffits

par(mfrow=c(2,2))
cd.C<-cooks.distance(a)
plot(cd.C, ylab="Cooks Distance", ylim=c(0,12))
iden(cd.C, a=1)
crit.value<-qf(0.5, ncol(X), nrow(X)-ncol(X))
abline(h=crit.value, lty=2)

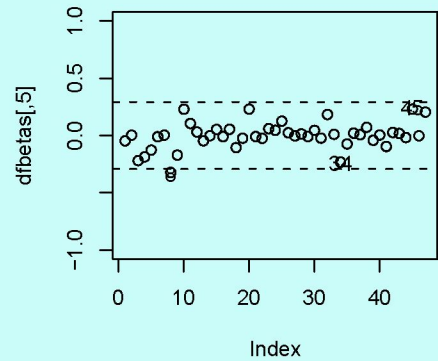
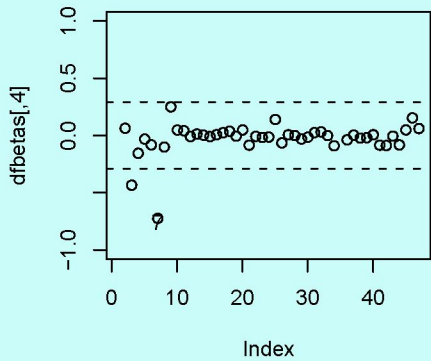
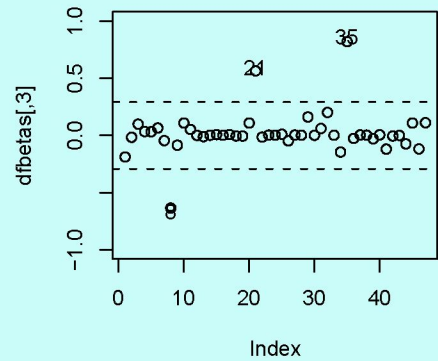
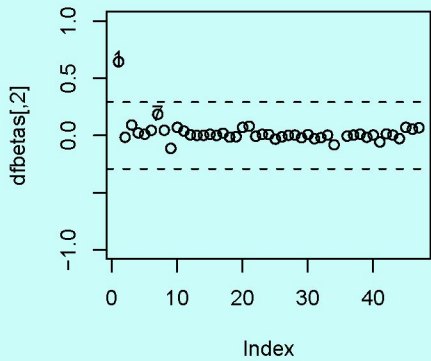
dffits.modelC<-dffits(a)
plot(dffits.modelC, ylab="Dffits")
iden(dffits.modelC, a=3)
crit.value<-2*sqrt(ncol(X)/nrow(X))
abline(h=c(-crit.value, crit.value), lty=2)
```



```
# DFbetas for the first two predictors

par(mfrow=c(2,2))
dfbetas.modelC<-dfbetas(a)
plot(dfbetas.modelC[,2], ylab="dfbetas[,2]", ylim=c(-1,1))
iden(dfbetas.modelC[,2], a=3)
crit.value<-2/sqrt(nrow(X))
abline(h=c(-crit.value, crit.value), lty=2)

plot(dfbetas.modelC[,3], ylab="dfbetas[,3]", ylim=c(-1,1))
iden(dfbetas.modelC[,3], a=3)
crit.value<-2/sqrt(nrow(X))
abline(h=c(-crit.value, crit.value), lty=2)
```



Questions

Are there any outliers or leverage points?

Do you detect any problems in the diagnostic plots?

Test the normality hypothesis with shapiro.test and the absence of heteroscedasticity using Breush-Pagan test

Checking normality

```
> shapiro.test(residuals(model.C))
```

```
Shapiro-Wilk normality test
```

```
data: residuals(model.C)
```

```
W = 0.9632, p-value = 0.1447
```

We accept the normality hypothesis

Heteroscedasticity

```
library(lmtest)  
> bptest(model.C)
```

```
studentized Breusch-Pagan test
```

```
data: model.C
```

```
BP = 4.0998, df = 4, p-value = 0.3927
```

We can not reject the heteroscedasticity hypothesis.

Spatial Autocorrelation?

In this type of problems it makes sense to think about some type of spatial autocorrelation.

If it exists a natural solution to correct for it is to introduce the area as a fixed effect.

Introduce SArea in Model (A). Choose the best model using *step* and call it Model (D).

```
model.A1<-lm(OBS~WH+BA+NAR+COR+SF+VI+PS+ES+AF+CO+AR+
model.D<-step(model.A1)
formula(model.D)
> formula(model.D)
OBS ~ PS + AL + FR + SArea
```

The selected model is $OBS = PS + AL + FR + SArea$.

Do the ANOVA for Model (D). What are the variables statistically significant? Calculate 95 % confidence intervals for the coefficients of the explanatory variables.

```
> summary.aov(model.D)
              Df      Sum Sq   Mean Sq  F value    Pr(>F)
PS              1   10862355   10862355    2.7424    0.1054
AL              1  131460498  131460498   33.1892 9.461e-07 ***
FR              1  460063280  460063280  116.1501 1.558e-13 ***
SArea           2   96615883   48307942   12.1961 6.979e-05 ***
Residuals     41 162398404    3960937
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Coefficients of Confidence Intervals

```
> confint(model.D)
                2.5 %      97.5 %
(Intercept) -1888.3999583 2779.9912039
PS           -0.2039942   4.5243304
AL           -0.9378670   0.1030052
FR           0.7251398    1.1019592
SAreaR67     -2008.5562686 3627.5420881
SAreaR68     -5652.0362374 518.4584211
```

Compute the coefficient of determinations R^2 , and R^2_a , the AIC, and the BIC statistic of Model (D)

```
> summary(model.D)$r.squared
[1] 0.8114716
> summary(model.D)$adj.r.squared
[1] 0.7884804
> AIC(model.D)
[1] 854.9848
> AIC(model.D, k = log(nrow(satfruit)))
[1] 867.9358
```

In summary

Models	R^2	R_{adj}^2	AIC	BIC
model A)	0.78	0.69	880	909
model B)	0.78	0.72	871	891
model C)	0.75	0.72	867	878
model D)	0.81	0.79	855	868

Use the function `drop1()` to test the statistically significant presence of PS and AL.

```
> drop1(model.D, test="Chisq")
```

```
Single term deletions
```

```
Model:
```

```
OBS ~ PS + AL + FR + SArea
```

	Df	Sum of Sq	RSS	AIC	Pr(Chi)
<none>		162398404	720		
PS	1	13487261	175885665	721	0.05282 .
AL	1	10392921	172791326	721	0.08773 .
FR	1	379805229	542203634	774	5.174e-14 ***
SArea	2	96615883	259014288	738	1.720e-05 ***

```
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
> drop1(model.D,test="F")
```

```
Single term deletions
```

```
Model:
```

```
OBS ~ PS + AL + FR + SArea
```

	Df	Sum of Sq	RSS	AIC	F value	Pr (F)
<none>		162398404		720		
PS	1	13487261	175885665	721	3.4051	0.07223 .
AL	1	10392921	172791326	721	2.6239	0.11294
FR	1	379805229	542203634	774	95.8877	2.708e-12 ***
SArea	2	96615883	259014288	738	12.1961	6.979e-05 ***

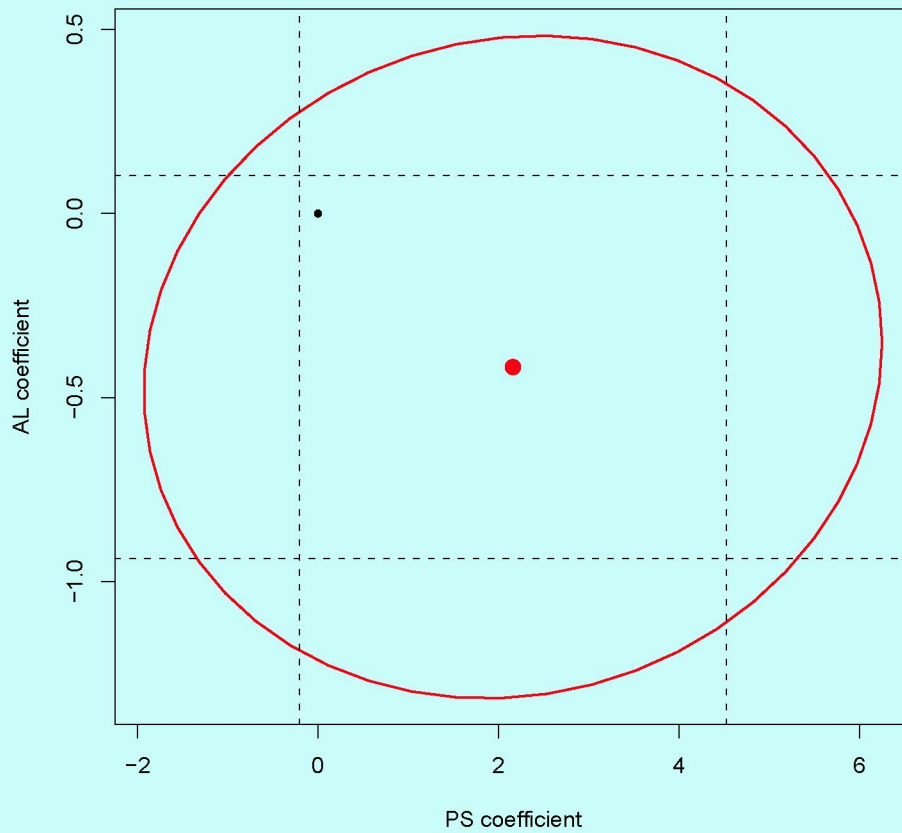
```
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```


Use `confidence.ellipse()` of package `car` to test that PS and AL are jointly equal to zero

```
library(car)
confidence.ellipse(model.D, Scheffe=TRUE)
points(0,0,pch=20)
abline(v=confint(model.D)[2,],lty=2)
abline(h=confint(model.D)[3,],lty=2)
```

Yes, the origen (0,0) is located inside the ellipse



Drop out the variables PS and AL of Model (D). Called the new model Model (E).

```
model.E<-lm(OBS~FR+SArea)
```

```
summary.aov(model.E)
```

```
> summary.aov(model.E)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
FR	1	577357111	577357111	131.945	1.091e-14	***
SArea	2	95886674	47943337	10.957	0.0001427	***
Residuals	43	188156636	4375736			

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Check normality, and homoscedasticity for Model (E) using graphics and hypotheses tests.

First we may have a look to the default diagnostics typing in R

```
par(mfrow=c(2,2))  
plot(model.E)
```

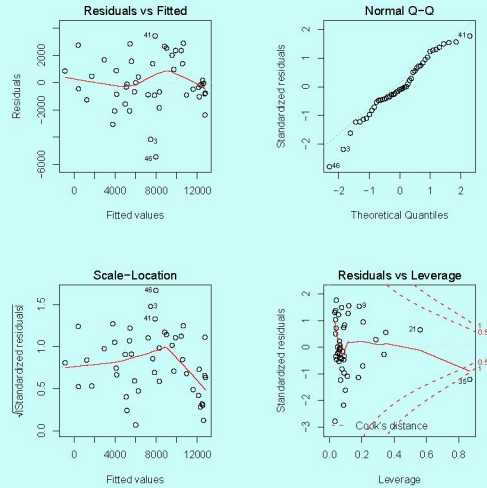


Figura 13: Default Diagnostics Model E

Check normality and homoscedasticity for Model (E) using the corresponding tests gives

```
> shapiro.test(residuals(model.E))
```

```
Shapiro-Wilk normality test
```

```
data: residuals(model.E)
```

```
W = 0.9568, p-value = 0.08035
```

```
> library(lmtest)
```

```
> bptest(model.E)
```

```
studentized Breusch-Pagan test
```

```
data: model.E
```

```
BP = 4.1865, df = 3, p-value = 0.242
```

Plot the standardized residuals, the student residuals, the Cook distances, the diagonal elements of the hat matrix, the DFFITS, and DFBETAS of Model (E). Are there any outliers and/or leverage points?

```
a<-model.E

par(mfrow=c(2,2),pty="s")
plot(hatvalues(a),type="h",xlab="",ylim=c(0,1),
ylab="diagonales de la matriz hat")
X<-model.matrix(a)
abline(h=2*(ncol(X))/nrow(X))
iden(hatvalues(a))
title("a) Elementos
diagonales \n de la matriz hat")

plot(rstandard(a),type="n",xlab="",ylab="r_i")
text(rstandard(a))
title("b) Residuales estandarizados \n internamente")
```

```
plot(rstudent(a),type="n",xlab="",ylab="r_i^*")
text(rstudent(a))
abline(h=qt(0.025, a$df.residual-1))
abline(h=qt(0.975,a$df.residual-1))
title("c) Residuales estandarizados \n externamente")

prediccion<-predict.lm(a)
plot(prediccion, satfruit$OBS,xlab="v. ajustados",ylab="y",type="n",
main="d) Observados vs. ajustados")
abline(0,1)
text(prediccion, satfruit$OBS)
```



```
# Cook distance and Dffits Model E

par(mfrow=c(2,2))
cd.E<-cooks.distance(a)
plot(cd.E, ylab="Cooks Distance", ylim=c(0,12))
iden(cd.E, a=1)
crit.value<-qf(0.5, ncol(X), nrow(X)-ncol(X))
abline(h=crit.value, lty=2)

dffits.modelE<-dffits(a)
plot(dffits.modelE, ylab="Dffits", ylim=c(-1,1))
iden(dffits.modelE, a=3)
crit.value<-2*sqrt(ncol(X)/nrow(X))
abline(h=c(-crit.value, crit.value), lty=2)
```

```
#DFbetas Model E
par(mfrow=c(2,2))
dfbetas.modelE<-dfbetas(a)
plot(dfbetas.modelE[,2], ylab="dfbetas[,2]", ylim=c(-1,1))
iden(dfbetas.modelE[,2], a=3)
crit.value<-2/sqrt(nrow(X))
abline(h=c(-crit.value, crit.value), lty=2)

plot(dfbetas.modelE[,3], ylab="dfbetas[,3]", ylim=c(-1,1))
iden(dfbetas.modelE[,3], a=3)
crit.value<-2/sqrt(nrow(X))
abline(h=c(-crit.value, crit.value), lty=2)

plot(dfbetas.modelE[,4], ylab="dfbetas[,4]", ylim=c(-1,1))
iden(dfbetas.modelE[,4], a=3)
crit.value<-2/sqrt(nrow(X))
abline(h=c(-crit.value, crit.value), lty=2)
```

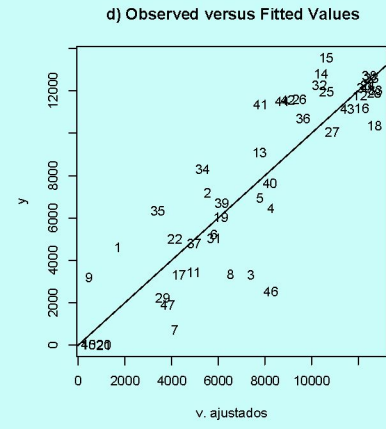
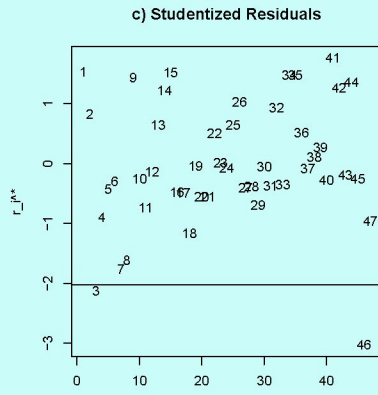
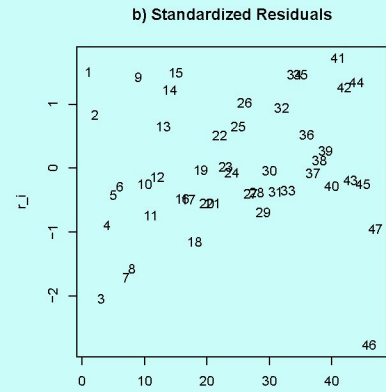
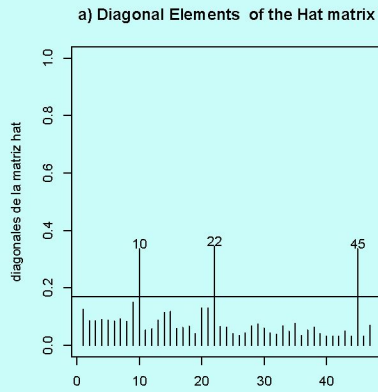


Figura 14: Diagnostics Model E

Drop out the 46 record of Model (E). Fit the new model and called Model (F)

```
satfruit1<-satfruit[-46,]  
dim(satfruit1)  
detach(satfruit)  
attach(satfruit1)  
model.F<-lm(OBS ~ FR + SArea)
```

Do the default diagnostic regression plots of Model (F).

```
par(mfrow=c(2,2))  
plot(model.F)
```

Plot the standardized residuals, the student residuals, the Cook distances, the diagonal elements of the hat matrix, the DFFITS, and DFBETAS of Model (F). Are there any leverage points and/or any outliers?

```
a<-model.F
par(mfrow=c(2,2),pty="s")
plot(hatvalues(a),type="h",xlab="",ylim=c(0,1),
ylab="diagonales de la matriz hat")
X<-model.matrix(a)
abline(h=2*(ncol(X))/nrow(X))
iden(hatvalues(a))
title("a) Diagonal Elements of the Hat matrix")

plot(rstandard(a),type="n",xlab="",ylab="r_i")
text(rstandard(a))
title("b) Standardized Residuals")

plot(rstudent(a),type="n",xlab="",ylab="r_i^*")
text(rstudent(a))
abline(h=qt(0.025, a$df.residual-1))
abline(h=qt(0.975,a$df.residual-1))
```

```
title("c) Studentized Residuals")

prediccion<-predict.lm(a)
plot(prediccion, satfruit1$OBS, xlab="v. ajustados", ylab="y",type="n",
main="d) Observed versus Fitted Values")
abline(0,1)
text(prediccion, satfruit1$OBS)
```

```

# Cook distance and Dffits Model F

par(mfrow=c(2,2))
cd.F<-cooks.distance(a)
plot(cd.F, ylab="Cooks Distance", ylim=c(0,12))
iden(cd.F, a=1)
crit.value<-qf(0.5, ncol(X), nrow(X)-ncol(X))
abline(h=crit.value, lty=2)

dffits.modelF<-dffits(a)
plot(dffits.modelF, ylab="Dffits", ylim=c(-1,1))
iden(dffits.modelF, a=3)
crit.value<-2*sqrt(ncol(X)/nrow(X))
abline(h=c(-crit.value, crit.value), lty=2)

#DFbetas Model F
par(mfrow=c(2,2))
dfbetas.modelF<-dfbetas(a)
plot(dfbetas.modelF[,2], ylab="dfbetas[,2]", ylim=c(-1,1))

```



```
iden(dfbetas.modelF[,2], a=3)
crit.value<-2/sqrt(nrow(X))
abline(h=c(-crit.value, crit.value), lty=2)
```

```
plot(dfbetas.modelF[,3], ylab="dfbetas[,3]", ylim=c(-1,1))
iden(dfbetas.modelF[,3], a=3)
crit.value<-2/sqrt(nrow(X))
abline(h=c(-crit.value, crit.value), lty=2)
```

Check the adequacy of the normality, and homoscedasticity assumptions of Model (F)

```
> shapiro.test(residuals(model.F))
```

```
Shapiro-Wilk normality test
```

```
data: residuals(model.F)
```

```
W = 0.9561, p-value = 0.08064
```

```
> bptest(model.F)
```

```
studentized Breusch-Pagan test
```

```
data: model.F
```

```
BP = 12.4314, df = 3, p-value = 0.006043
```

Compute 95 % confidence intervals for the parameters of the explanatory variables in Model (F) and comment on the results

```
> confint(model.F)
              2.5 %      97.5 %
(Intercept) -1808.7291106 2678.215761
FR           0.7671677    1.076457
SAreaR67    -1698.5805870 3397.376242
SAreaR68    -5486.4253362  90.860897
```

How many hectares of observed fruits are expected to be incremented if the classified hectares of fruit trees by the satellite are increased by 10000 m2 (1 ha)?

```
> summary(model.F)
> summary(model.F)$coef[2,1]
[1] 0.9218121
> summary(model.F)$coef[2,1]*10000
[1] 9218.121
```

Suppose the total classified fruits by the satellite in area R63 is 97044.28 m², in area R67 is 4878603.43 m², and in area R68 is 2883488.24 m², calculate the total prediction of fruit trees by small areas

```
#R63
> summary(model.F)$coef[1,1]+summary(model.F)$coef[2,1]*(97044.28)
[1] 89891.34

#R67
> summary(model.F)$coef[1,1]+summary(model.F)$coef[2,1]*4878603.43
+ summary(model.F)$coef[3,1]
[1] 4498440

#R68
> summary(model.F)$coef[1,1]+summary(model.F)$coef[2,1]*2883488.24 +
summary(model.F)$coef[4,1]
[1] 2655771
```

```
# Simpler way
FR.pob<-c(97044.28, 4878603.43, 2883488.24)
SArea.pob<-c("R63", "R67", "R68")
newdata<-data.frame(FR.pob, SArea.pob)
names(newdata)<-c("FR", "SArea")
> predict(model.F, newdata)
           1           2           3
89891.34 4498439.95 2655771.39
```

In hectares:

```
> predict(model.F, newdata)/10000
           1           2           3
 8.989134 449.843995 265.577139
```

Plot in the same graphical page FR versus OBS separately by the three areas.
Superimpose the corresponding regression lines

Let us compute first the coefficients of the regression lines for every area

```
> contrasts(satfruit1$SArea)
      R67 R68
R63    0    0
R67    1    0
R68    0    1

> coef(model.F) ##### general coefficients
(Intercept)          FR      SAreaR67      SAreaR68
434.7433254      0.9218121  849.3978277 -2697.7822198
```

```
#### coefficients for R63
```

```
> coef.R63<-cbind(coef(model.F) [1],coef(model.F) [2])
```

```
> coef.R63
```

```
          [,1]      [,2]
```

```
(Intercept) 434.7433 0.9218121
```

```
#### coefficients for R67
```

```
> coef.R67<-cbind(coef(model.F) [1]+coef(model.F) [3], coef(model.F) [2])
```

```
> coef.R67
```

```
          [,1]      [,2]
```

```
(Intercept) 1284.141 0.9218121
```

```
#### coefficients for R68
```

```
> coef.R68<-cbind(coef(model.F) [1]+coef(model.F) [4], coef(model.F) [2])
```

```
> coef.R68
```

```
          [,1]      [,2]
```

```
(Intercept) -2263.039 0.9218121
```

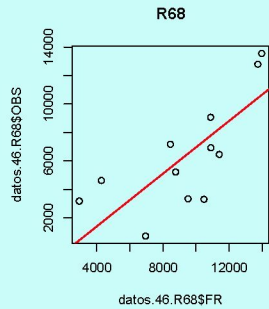
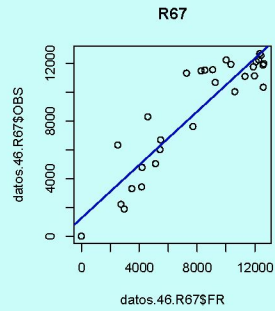
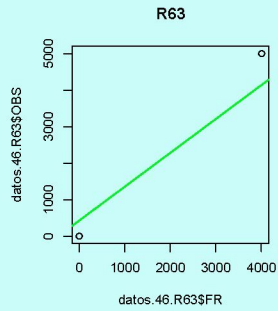


```
par(mfrow=c(2,2))
```

```
satfruit1.R63=satfruit1[satfruit1$SArea=="R63",]  
plot(satfruit1.R63$FR,satfruit1.R63$OBS,main="R63")  
abline(coef.R63,col="green",lwd=2)
```

```
satfruit1.R67=satfruit1[satfruit1$SArea=="R67",]  
plot(satfruit1.R67$FR,satfruit1.R67$OBS,main="R67")  
abline(coef.R67,col="blue",lwd=2)
```

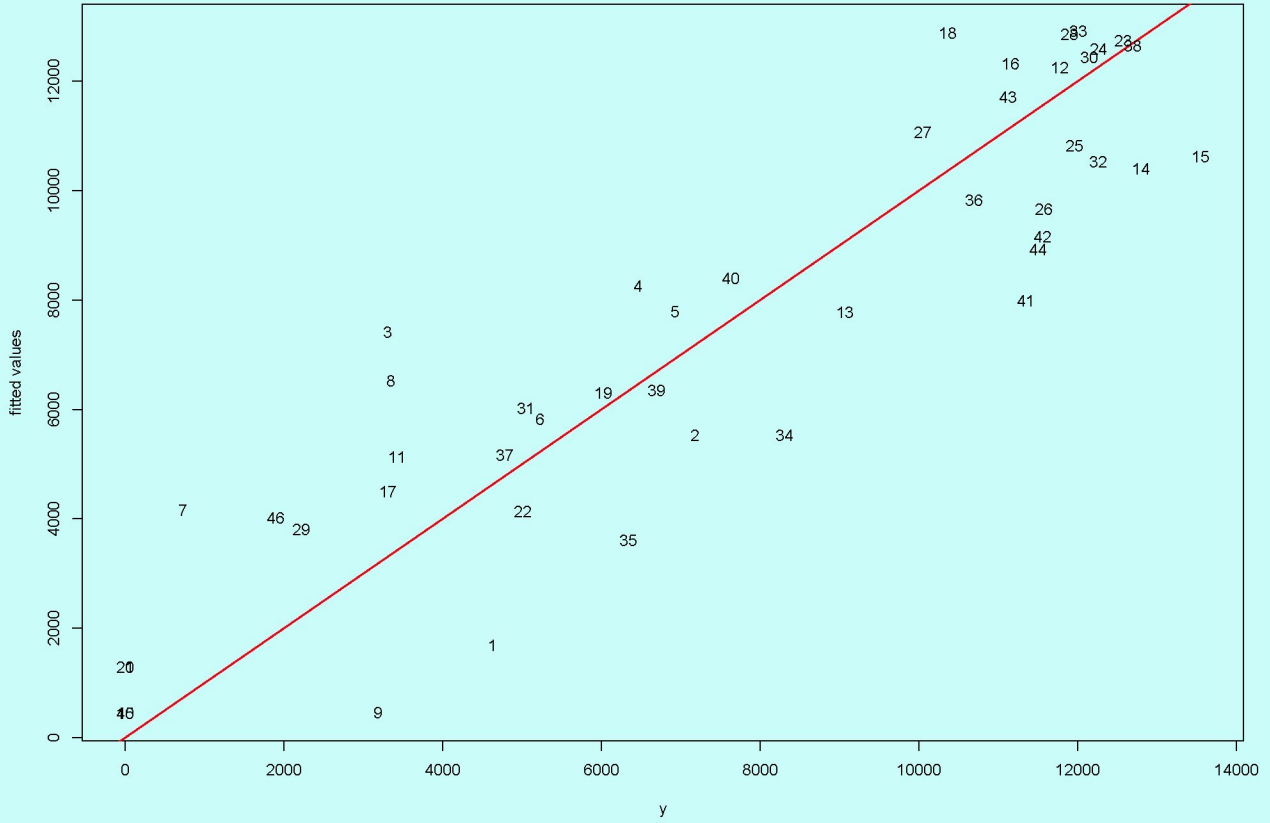
```
satfruit1.R68=satfruit1[satfruit1$SArea=="R68",]  
plot(satfruit1.R68$FR,satfruit1.R68$OBS,main="R68")  
abline(coef.R68,col="red",lwd=2)
```



Plot the individual predictions versus the observed data. Add a diagonal line to the plot.

```
par(mfrow=c(1,1))
prediccion<-predict.lm(model.F)
plot(satfruit1$OBS,prediccion,ylab="fitted values",xlab="y",
type="n", main="Fitted vs. Observed")
abline(0,1,col="red",lwd=2)
text(satfruit1$OBS,prediccion)
```

Fitted vs. Observed



Do a barplot to graph simultaneously the predicted totals (using the regression model) and the direct estimates by areas. Recall that the direct estimate by areas is calculated multiplying the observed mean by the total number of classified segments that are: 119, 703, and 564 for R63, R67, and R68, respectively

```
means.AREAS<-sapply(split(satfruit$OBS, satfruit$SArea), mean)
TOTAL.AREAS<-means.AREAS*c(119, 703, 564)
> TOTAL.AREAS
      R63      R67      R68
198466.5 5867470.0 3589159.5

> TOTAL.AREASPRED<-predict(model.F, newdata)
      1      2      3
89891.34 4498439.95 2655771.39
```

```
>resumen<-rbind(TOTAL.AREAS,TOTAL.AREASPRED)
```

```
> resumen
```

```
                R63      R67      R68
TOTAL.AREAS      198466.50 5867470 3589159
TOTAL.AREASPRED  89891.34 4498440 2655771
```

```
par(mfrow=c(1,1))
```

```
row.names(resumen)<-c("Direct Est.," "Model Prediction")
```

```
barplot(resumen/10000,legend=row.names(resumen),main="Direct estimates  
and Predicted Fruits Totals in ha.",beside=TRUE,col=c(3,4))
```

```
> sum(TOTAL.AREASPRED)
```

```
[1] 7244103
```

```
> sum(TOTAL.AREASPRED)/10000 #Total number of hectares in Comarca VII
```

```
[1] 724.4103
```

Direct estimates and Predicted Fruits Totals in ha.

