Case Study: Biomass Estimation from Beeches

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

Introduction

- Data for this case study come from *Gestión Ambiental de Viveros y Repoblaciones de Navarra* and *Gobierno de Navarra*, 2006.

- To estimate **the amount of carbon dioxide retained in a tree**, **its biomass needs to be known and multiplied by an expansion factor** (there are several alternatives in the literature). To calculate the biomass, specific regression equations by species are frequently used. These regression equations, called allometric equations, estimate the biomass of the tree by means of some known characteristics, typically diameter and/or height of the stem and branches.

Introduction

File Data

The biomass file contains data of 42 beeches (Fagus Sylvatica) from a forest of Navarra (Spain) in 2006, where

- Dn: diameter of the stem in centimeters
- H: height of the tree in meters
- PST: weight of the stem in kilograms
- PSA: aboveground weight in kilograms

Regression Analysis with R

- Make a scatterplot of PSA and Dn. Do you think is it possible to fit a regression line to explain the weight of the beech in terms of the diameter of the stem?
- Make a scatterplot of log(PSA) and log(Dn). Do you think is it possible to fit a regression line to explain the weight of the beech in terms of the diameter of the stem?

Regression Analysis with R

```
library(PASWR)
attach(biomass)
par(mfrow=c(1,2))
plot(Dn, PSA) # a clear non-linear relationship
abline(lsfit(Dn,PSA), col=2, lwd=2)
plot(log(Dn), log(PSA))   #the relation seems to be linear
abline(lsfit(log(Dn),log(PSA) ), col=2, lwd=2)
```

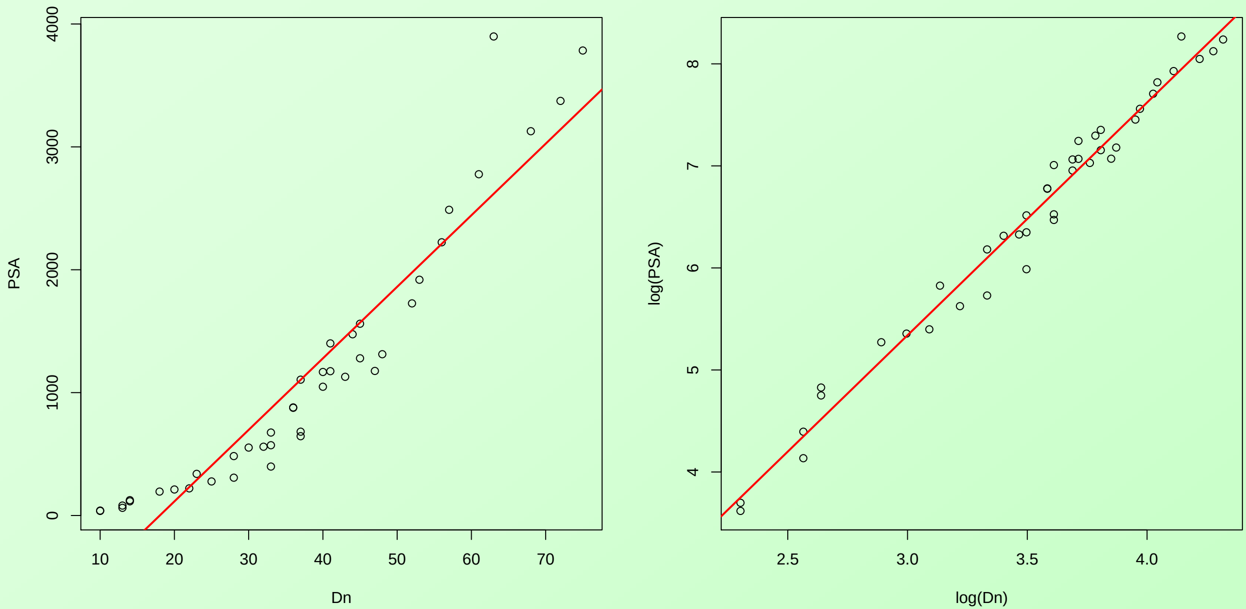

Figura 1: Scatterplots

Regression Analysis with R

- Fit the regression model

$$\log(PSA) = \beta_0 + \beta_1 \log(Dn)$$

- Compute R^2, R_a^2, and the residual variance

Solution in R

```
model.DN<-lm(log(PSA)~log(Dn))
> summary(model.DN)
        ...                   ...
Residuals:
     Min       1Q   Median       3Q      Max
-0.48510 -0.12682  0.02701  0.10766  0.32104
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.5015     0.1920  -7.822 1.38e-09 ***
log(Dn)       2.2806     0.0542  42.076  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 0.1842 on 40 degrees of freedom
Multiple R-Squared: 0.9779,     Adjusted R-squared: 0.9774
F-statistic:  1770 on 1 and 40 DF,  p-value: < 2.2e-16
```

Regression Analysis with R

- Introduce H as an explanatory variable and fit the model

$$\log(PSA) = \beta_0 + \beta_1 \log(Dn) + \beta_2 H$$

- Is H statistically significant?

Solution in R

```
> model.DNH<-lm(log(PSA)~log(Dn)+ H )

> summary(model.DNH)

...          ...        ...

Residuals:

     Min        1Q    Median         3Q        Max

-0.29861 -0.11093 -0.01903  0.07141  0.38130

Coefficients:

             Estimate Std. Error t value Pr(>|t|)

(Intercept) -1.691859   0.167001 -10.131 1.77e-12 ***

log(Dn)      2.185244   0.050682  43.117  < 2e-16 ***

H            0.023259   0.005487   4.239 0.000133 ***

---

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

Residual standard error: 0.1544 on 39 degrees of freedom

Multiple R-Squared: 0.9849,     Adjusted R-squared: 0.9841

F-statistic:  1270 on 2 and 39 DF,  p-value: < 2.2e-16
```

Regression Analysis with R

- Estimate the model's parameters and their standard errors. (See again summary(model.DHN))
- Provide an interpretation for the model's parameters.

Interpretation of the model's parameters

The fitted model is:

$$\log\left(PSA\right) = -1{,}691859 + 2{,}185244 \log\left(Dn\right) + 0{,}023259\, H$$

$\hat{\beta}_1$ can be interpreted as:

$$\hat{\beta}_1 = \frac{\%\Delta y}{\%\Delta x}$$

For a given H, if the diameter of the stem increases by $1\,\%$, the weight of the stem increases approximately $2{,}19\,\%$.

For a given diameter, each meter of increase in height produces an increase of the weight of the stem of approximately $2{,}33\,\%$

Regression Analysis with R

Compute the variance-covariance matrix of the $\hat{\beta}_s$

Solution in R

```
> vcov(model.DNH)
              (Intercept)       log(Dn)             H
(Intercept)  0.0278894670 -0.0062149387 -0.0002463658
log(Dn)     -0.0062149387  0.0025686777 -0.0001234109
H           -0.0002463658 -0.0001234109  0.0000301036
```

Regression Analysis with R

Provide 95 % confidence intervals for β_1 and β_2

Solution in R

```
> confint(model.DNH)
                  2.5 %      97.5 %
(Intercept) -2.02965082 -1.35406639
log(Dn)      2.08272951  2.28775805
H            0.01216104  0.03435673
```

Regression Analysis with R

Compute R^2, R_a^2, and the residual variance

Solution in R

```
summary(model.DNH)$r.squared
[1] 0.9848744
summary(model.DNH)$adj.r.squared
[1] 0.9840988
summary(model.DNH)$sigma^2
[1] 0.02383166
```

Regression Analysis with R

Construct a graph with the default diagnostics plots of R

Solution in R

```
win.graph()
par(mfrow=c(2,2),pty="s")
plot(model.DNH)
```

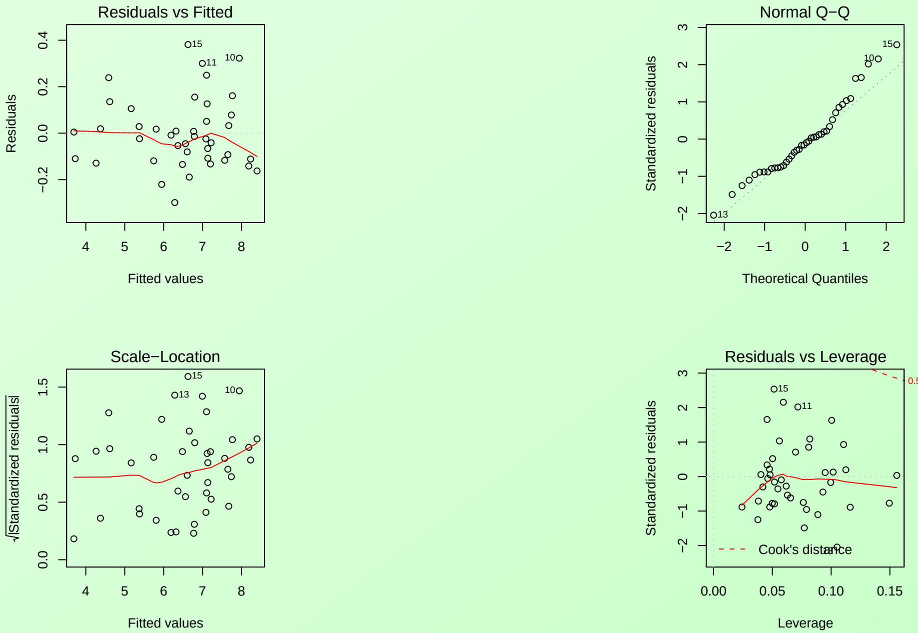

Figura 2: Diagnostics Plots

Regression Analysis with R

Can homogeneity of variance be assumed?

Solution in R

```
library(lmtest)
bptest(model.DNH)

    studentized Breusch-Pagan test

data:  model.DNH
BP = 7.626, df = 2, p-value = 0.02208
```

Regression Analysis with R

Do the residuals appear to follow a normal distribution?

Solution in R

```
> shapiro.test(rstandard(model.DNH))

        Shapiro-Wilk normality test

data:  rstandard(model.DNH)
W = 0.9569, p-value = 0.1146
```

Regression Analysis with R

Are there any outliers in the data?

Solution in R

```
a=model.DNH
win.graph()
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")
```

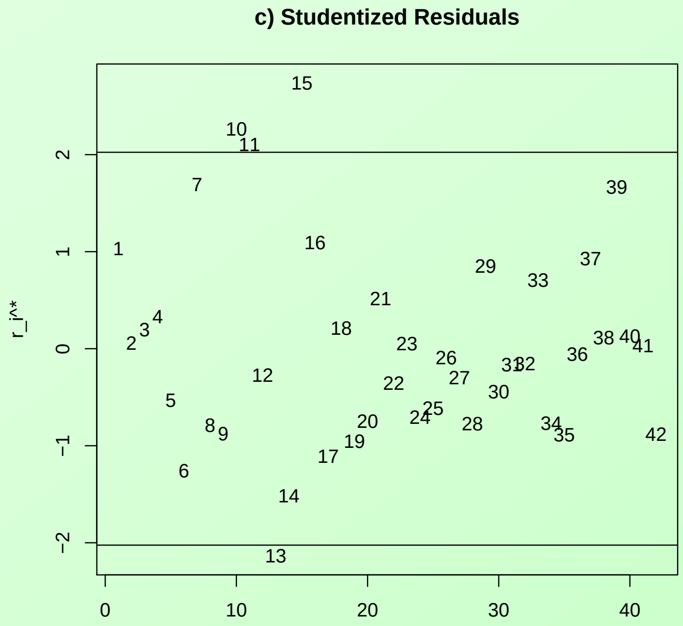

Figura 3: Diagnostics Plots

Regression Analysis with R

Are there any influential observations in the data?

Solution in R

```
 # Cook distance
a=model.DNH
win.graph()
par(mfrow=c(2,2))
cd.F<-cooks.distance(a)
plot(cd.F, ylab="Cooks Distance", ylim=c(0,0.8))
iden(cd.F, a=3)
crit.value<-qf(0.5, ncol(X), nrow(X)-ncol(X))
abline(h=crit.value, lty=2)
```

```
#  Dffits
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)
```

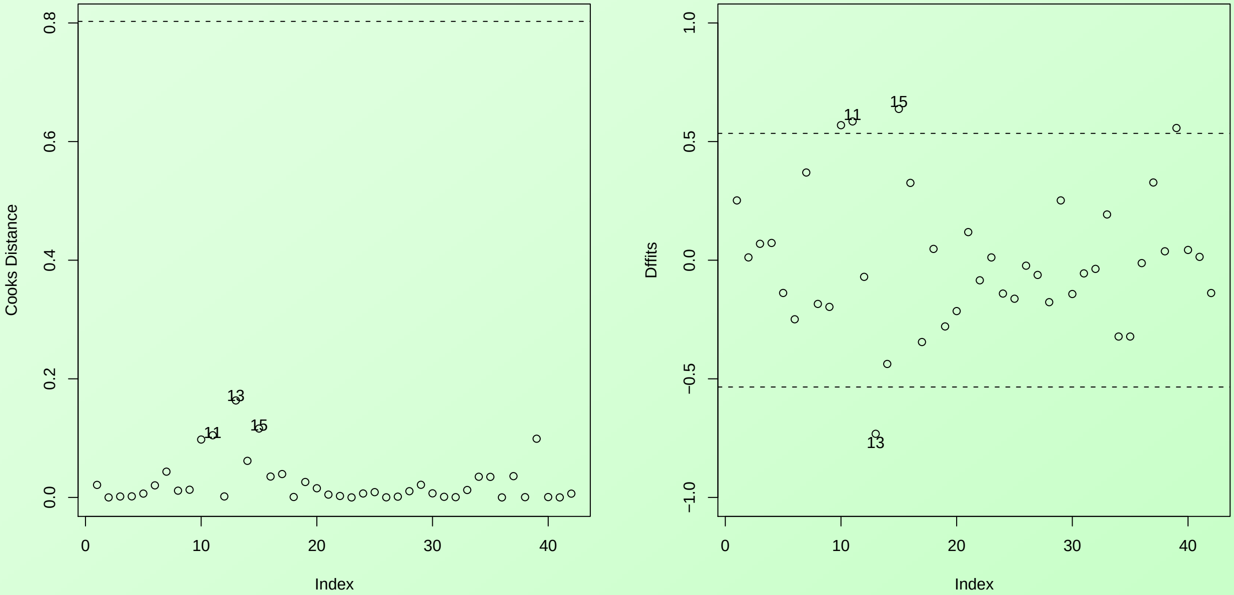

Figura 4: Dffits

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

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)
```

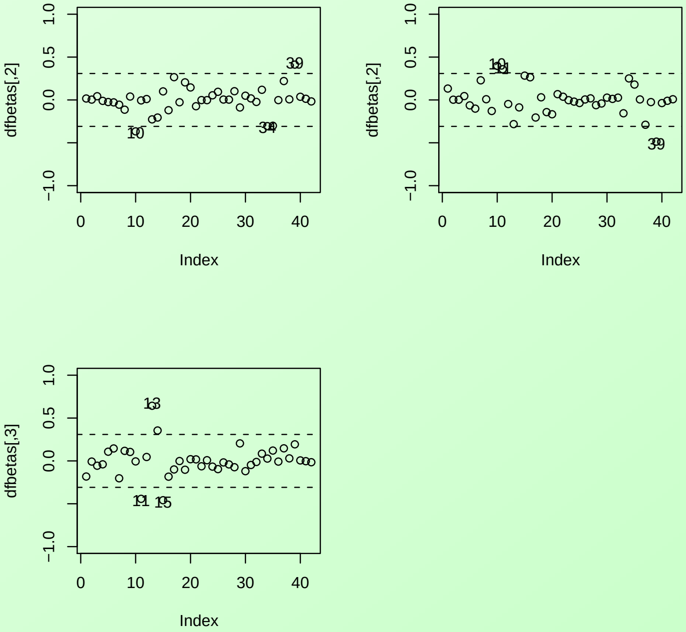

Figura 5: Dfbetas

Predictions

Obtain predictions of the aboveground biomass of trees with diameters $Dn = seq(12{,}5, 42{,}5, 5)$ and heights $H = seq(10, 40, 5)$. Note that the weight predictions are obtained from back transforming the logarithm. The bias correction is obtained by means of the log-normal distribution: if $\hat{Y}_{\text{pred}}$ is the prediction, the corrected (back-transformed) prediction $\tilde{Y}_{pred}$ is given by

$$\tilde{Y}_{\text{pred}} = \text{exp}(\hat{Y}_{\text{pred}} + \hat{\sigma}^2/2)$$

where $\hat{\sigma}^2$ is the variance of the error term.

Predictions with R

```
> Dn<-seq(12.5,42.5,5)
> H<-seq(10,40,5)
> newdata<-data.frame(Dn,H)
> predictions<-exp(predict.lm(model.DNH,newdata)+summary(model.DNH)$sigma^2/2)
> predictions
        1          2          3          4          5          6          7
 58.67295  137.48945  267.47468  465.83604  753.84143 1157.69554 1709.56309
```

A Model for the Stem

Fit the following regression model for the weight of the stem

$$PST = \beta_0 + \beta_1 Dn + \beta_2 H$$

- Display the default diagnostics plots. What does the fitted values vs. the residuals plot suggest?
- Propose a model to correct the above problem
- Does your new model correct the residuals problem detected?