# StatR 101: Fall 2012

Homework 11

Rod Doe

Wedneday, December 12, 2012

# Regression

setwd("C:/Users/Rod/SkyDrive/R/101/Week11")

metrics = read.table(file="FlowerMetrics.txt", header=TRUE)

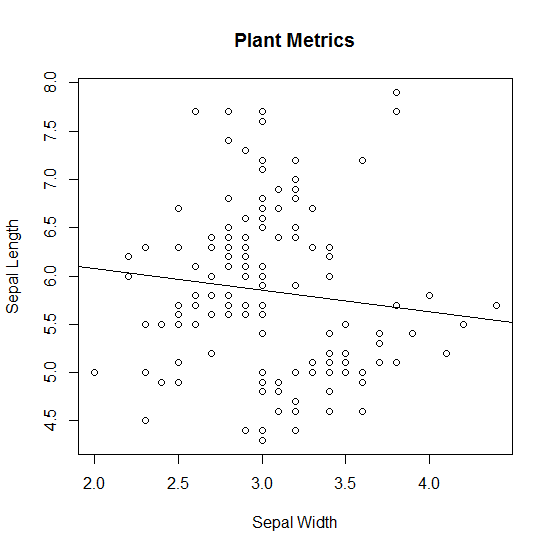
# Model the sepal length (Y) with respect to sepal width (X) using a linear regression of the form

# M1:Yi=a+ßXi+?i

M1 <- lm(metrics$Sepal.Length ~ metrics$Sepal.Width)

plot(metrics$Sepal.Width, metrics$Sepal.Length, main="Plant Metrics", xlab="Sepal Width", ylab="Sepal Length")

abline(lm(metrics$Sepal.Length ~ metrics$Sepal.Width))



summary(M1)

Call:

lm(formula = metrics$Sepal.Length ~ metrics$Sepal.Width)

Residuals:

Min 1Q Median 3Q Max

-1.5561 -0.6333 -0.1120 0.5579 2.2226

Coefficients:

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

(Intercept) 6.5262 0.4789 13.63 <2e-16 \*\*\*

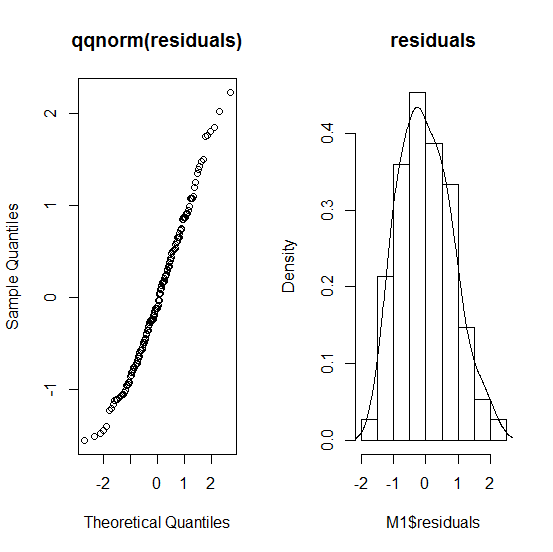
metrics$Sepal.Width -0.2234 0.1551 -1.44 0.152

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

## Regression Residuals

The residuals are normally distributed, so the assumption of independent and identically distributed residuals is validated.



The anova result shows a p-value of 0.15, so I would say that there is not a significant relationship between the variables.

anova(M1)

Analysis of Variance Table

Response: metrics$Sepal.Length

Df Sum Sq Mean Sq F value Pr(>F)

metrics$Sepal.Width 1 1.412 1.41224 2.0744 0.1519

Residuals 148 100.756 0.68078

# One-Way Anova

Get the data ready for battle:

Sepal.Length = metrics$Sepal.Length

Species = factor(metrics$Species)

lmfit = lm(Sepal.Length ~ Species)

boxplot(lmfit, main="Sepal Length by Species", col=c(2,3,4))

## Summary

## Summary Table of Data

SummaryTable <- data.frame(

n = tapply(Sepal.Length, Species, length),

mean = tapply(Sepal.Length, Species, mean),

sd = tapply(Sepal.Length, Species, sd))

SummaryTable

n mean sd

setosa 50 5.006 0.3524897

versicolor 50 5.936 0.5161711

virginica 50 6.588 0.6358796

## Linear Model of Sepal Length by Species

lmfit = lm(Sepal.Length ~ Species)

lmfit

Call:

lm(formula = Sepal.Length ~ Species)

Coefficients:

(Intercept) Speciesversicolor Speciesvirginica

5.006 0.930 1.582

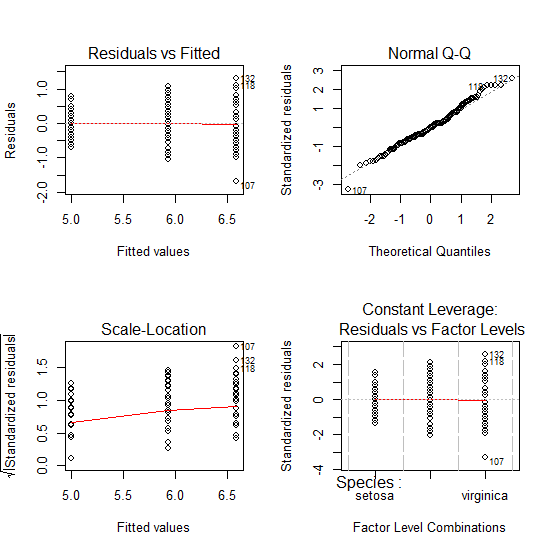
Interesting thing to note here is that the values given here are either the mean value of setosa, or offsets of the species mean from the setosa mean.

## Diagnostic Plots

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

> plot(lmfit)

The qqnorm plot shows normally distributed residuals.



## Null Hypothesis

H0: There is no relationship between sepal length and species.

> anova(lmfit)

Analysis of Variance Table

Response: Sepal.Length

Df Sum Sq Mean Sq F value Pr(>F)

Species 2 63.212 31.606 119.26 < 2.2e-16 \*\*\*

Residuals 147 38.956 0.265

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

With this p-value, it is appropriate to reject the null hypothesis that there is no relationship between the sepal length and plant species.

### Doing it the hard way

> mean(metrics$Sepal.Length)

[1] 5.843333

> sum((metrics$Sepal.Length - mean(metrics$Sepal.Length))^2)

[1] 102.1683

> sum((metrics$Sepal.Length[metrics$Species == 'setosa'] - mean(metrics$Sepal.Length[metrics$Species == 'setosa']))^2)

[1] 6.0882

> sum((metrics$Sepal.Length[metrics$Species == 'versicolor'] - mean(metrics$Sepal.Length[metrics$Species == 'versicolor']))^2)

[1] 13.0552

> sum((metrics$Sepal.Length[metrics$Species == 'virginica'] - mean(metrics$Sepal.Length[metrics$Species == 'virginica']))^2)

[1] 19.8128

6.0882 + 13.0552 + 19.8128

[1] 38.9562

> qf(0.95, 1, 147)

[1] 3.905498

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Source** | **Sum of Squares** | **Degrees of Freedom**  **=K(N-1)** | **Mean Square = Sum of Squares / Degrees of Freedom** | **F-Ratio** |
| SSY | 102.1683 | 1 | 102.1683 | 385.5288 |
| SSE | 38.9562 | 147 = 3(50 – 1) | 38.9562 / 147 = 0.27 |  |
| Total | 141.1245 | 148 |  |  |

According to my manual calculation, the calculated F-ratio of 385 far exceeds the 95% confidence limit of 3.9. I reject the null hypothesis that there is no relationship between sepal length and species.

## Ancova

I ran out of time here, but enjoyed an excellent concert at my son’s school. My [Statistics: An Introduction Using R](http://www.amazon.com/Statistics-Introduction-Michael-J-Crawley/dp/0470022981/ref=sr_1_1?s=books&ie=UTF8&qid=1355378686&sr=1-1&keywords=crawley+statistics) has an excellent chapter on this subject. I promise to deep dive into it after the project is complete.