Statistical learning

Data visualization (DataViz) with R

V. Lefieux

1 Libraries

```
library(GGally) #For ggpairs
library(ggplot2) #"Grammar of graphics" library
library(latex2exp) #For LaTex formula
library(MASS) #For simulation of gaussian vectors
library(vcd) #For mosaic plots
library(vioplot) #For violin plots
```

2 Data

2.1 Ozone

Source: Laboratoire de mathématiques appliquées de l'Agrocampus Ouest http://math.agrocampus-ouest.fr/infoglueDeliverLive/enseignement/support2cours/donnees

The file *data_ozone.txt* contains 112 records with 14 attributes of ozone concentration in Rennes during summer 2001.

Name	Description
obs	Observation number
MaxO3	Daily maximum of ozone concentration ($\mu gr/m^3$)
T9, T12, T15	Temperature (at 09:00, 12:00 and 15:00)
Ne9, Ne12, Ne15	Cloud cover (at 09:00, 12:00 and 15:00)
Vx9, Vx12, Vx15	East-west component of the wind (at 09:00, 12:00 and 15:00)
MaxO3v	Daily maximum of ozone concentration for the day before
wind	Wind direction at 12:00
rain	Rainy or dry day

```
setwd("/Users/vincentlefieux/Documents/Data/")
ozone=read.table("data_ozone.txt",header=TRUE,sep=";",dec=",")
```

2.2 Bridges

Source: Yoram Reich & Steven J. Fenves, Department of Civil Engineering and Engineering Design Research Center, Carnegie Mellon University, Pittsburgh http://archive.ics.uci.edu/ml/datasets/Pittsburgh+Bridges

The file *data_bridges.txt* contains 108 records with 13 attributes of bridges in Pittsburgh.

Name	Type	Possible values	Comments
IDENTIF			Identifier of the examples
RIVER	n	A, M, O	
LOCATION	n	1 to 52	
ERECTED	c	1818-1986	
PURPOSE	n	WALK, AQUEDUCT, RR, HIGHWAY	
LENGTH	c	804-4558	
LANES	c	1, 2, 4, 6	
CLEAR-G	n	N, G	
T-OR-D	n	THROUGH, DECK	
MATERIAL	n	WOOD, IRON, STEEL	
SPAN	n	SHORT, MEDIUM, LONG	
REL-L	n	S, S-F, F	
TYPE	n	WOOD, SUSPEN, SIMPLE-T, ARCH, CANTILEV, CONT-T	

```
setwd("/Users/vincentlefieux/Documents/Data/")
bridges=read.table("data_bridges.txt",header=FALSE,sep=",",dec=".")
colnames(bridges)=c("IDENTIF","RIVER","LOCATION","ERECTED","PURPOSE","LENGTH","LANES",
                    "CLEARG","TORD","MATERIAL","SPAN","RELL","TYPE")
```

2.3 Airline passengers

airpass contains the monthly number of international airline passengers (in millions) from January 1949 to December 1960.

```
is.ts(AirPassengers)
```

```
## [1] TRUE
```

To know available colors use `colors()`.

3 Distribution plots

3.1 Categorical case

```
barplot(table(ozone$wind),main="Count")
```

```
barplot(table(ozone$wind)/sum(table(ozone$wind)),main="Frequency")
```

```
dir=names(table(ozone$wind))
freq=as.numeric(table(ozone$wind)/sum(table(ozone$wind)))
df=data.frame(dir=dir,freq=freq)

ggplot(data=df,aes(x=dir,y=freq))+
  geom_bar(stat="identity",fill="steelblue")+
  ggtitle("With ggplot2")+
  xlab("Wind direction")+
```

```
  ylab("Frequency")
```

```
ggplot(data=df,aes(x=dir,y=freq))+
  geom_bar(stat="identity",fill="steelblue")+
  theme_minimal()+
  ggtitle("With ggplot2")+
  xlab("Wind direction")+
  ylab("Frequency")
```

```
ggplot(data=df,aes(x="",y=freq,fill=dir))+
  geom_bar(stat="identity")+
  ggtitle("With ggplot2")+
  scale_fill_discrete(name="Wind direction")+
  ylab("Frequency")
```

Sometimes pie charts are used but they aren't so readable and clear:

```
pie(table(ozone$wind))
```

```
pie(table(ozone$wind),col=gray(seq(0.4,1,length=4)))
```

```
ggplot(data=ozone,aes(x=factor(1),fill=wind))+
  geom_bar(width=1)+
  coord_polar("y",start=0)+
  ggtitle("With ggplot2")+
  scale_fill_discrete(name="Wind direction")+
  theme(axis.text=element_blank(),
        axis.ticks=element_blank(),
        panel.grid=element_blank())+
  xlab("")+
  ylab("Frequency")
```

It's really better to use barplot than a pie chart...

3.2 Discrete case

A bad plot:

```
freq_lanes=table(bridges$LANES)
barplot(freq_lanes/sum(freq_lanes),main="Frequency")
```

A better one:

```
freq_lanes2=freq_lanes[-1]
plot(freq_lanes2/sum(freq_lanes2),type="h",lwd=20,lend="butt",
     xlab="Value",ylab="Frequency",main="Bridges lanes")
```

Another better one:

```
df=data.frame(x=as.numeric(names(freq_lanes2)),freq=as.numeric(freq_lanes2/sum(freq_lanes2)))

ggplot(data=df,aes(x=x,y=freq))+
  geom_bar(stat="identity",fill="steelblue",width=0.5)+
  scale_x_continuous(name="Number of bridges lanes",limits=c(0,7),breaks=c(0:7),labels=c(0:7))+
  scale_y_continuous(name="Frequency",limits=c(0,0.7))+
  ggtitle("With ggplot2")
```

3.3 Continuous case

Some histograms and kernel density estimations

```
hist(ozone$T12,freq=FALSE,xlab="T12",ylab="",main="Histogram")
```

```r
hist(ozone$T12,freq=FALSE,breaks=20,xlab="T12",ylab="",main="Histogram")
```

```r
hist(ozone$T12,freq=TRUE,xlab="T12",ylab="",main="Histogram")
```

```r
plot(density(ozone$T12))
```

```r
plot(density(ozone$T12),col="red",type="l",xlab="T12",ylab="",main="Kernel density estimation")
```

```r
ggplot(ozone,aes(x=T12))+
  geom_histogram()+
  ggtitle("With ggplot2")+
  xlab("T12")+
  ylab("Count")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
ggplot(ozone,aes(x=T12))+
  geom_histogram(aes(y=..density..))+ #For frequencies
  ggtitle("With ggplot2")+
  xlab("T12")+
  ylab("Frequency")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
ggplot(ozone,aes(x=T12,fill=rain))+
  geom_histogram(aes(y=..density..),alpha=0.5,position="identity")+
  ggtitle("With ggplot2")+
  scale_fill_discrete(name="Rain/Dry")+
  xlab("T12")+
  ylab("Frequency")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
ggplot(ozone,aes(x=T12,fill=rain))+
  geom_histogram(aes(y=..density..),bins=15,alpha=0.5,position="dodge")+
  ggtitle("With ggplot2")+
  scale_fill_discrete(name="Rain/Dry")+
  xlab("T12")+
  ylab("Frequency")
```

```
ggplot(ozone,aes(x=T12,fill=rain))+
  geom_density(alpha=0.3)+
  ggtitle("With ggplot2")+
  scale_fill_discrete(name="Rain/Dry")+
  xlab("T12")+
  ylab("Density")
```

```
ggplot(ozone,aes(x=T12))+
  geom_histogram(aes(y=..density..),colour="black",fill="white")+
  geom_density(alpha=0.1,fill="red")+
  ggtitle("With ggplot2")+
  xlab("T12")+
  ylab("")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

The empirical cumulative distribution fuction

```
plot(ecdf(ozone$T12),main="T12")
```

```r
ggplot(ozone,aes(x=T12))+
  stat_ecdf(geom="step")+
  ggtitle("With ggplot2")+
  xlab("T12")+
  ylab("ECDF")
```

Box (and whiskers) plot

```
boxplot(ozone$T12,main="Boxplot T12")
```

```
boxplot(ozone[,c("T9","T12","T15")],main="Boxplots")
```

```
ggplot(ozone,aes(x=rain,y=T12))+
  geom_boxplot()+
  ggtitle("With ggplot2")+
  xlab("")+
  ylab("T12")
```

```r
ggplot(ozone,aes(x=rain,y=T12))+
  geom_boxplot()+
  ggtitle("With ggplot2")+
  xlab("")+
  ylab("T12")+
  coord_flip()
```

Violin plots

```r
dry=ozone$T12[ozone$rain=="Dry"]
rain=ozone$T12[ozone$rain=="Rain"]

vioplot(dry,rain,names=c("Dry","Rain"),col="gold")
title("Violin plots")
```

qq-plots (normality)

```
qqnorm(ozone$T9)
```

```
ggplot(data=ozone,aes(sample=T15))+
  geom_point(stat="qq")+
```

```
ggtitle("With ggplot2")
```

4 Links between variables

4.1 Scatterplots

Classic scatterplots

```
plot(ozone$T12,ozone$maxO3,xlab="T12",ylab="MaxO3",type="p",
     col="blue",main="MaxO3=f(T12)")
abline(h=mean(ozone$maxO3),v=mean(ozone$T12),col=c("red","red"))
```

```
ggplot(ozone,aes(x=T12,y=maxO3))+
  geom_point(aes(colour=rain))+
  ggtitle("With ggplot2")+
  scale_color_discrete(name="Rain/Dry")+
  xlab("T12")+
  ylab("MaxO3")
```

```
ggplot(ozone,aes(x=T12,y=maxO3))+
  geom_point(aes(colour=maxO3v))+
  ggtitle("With ggplot2")+
  scale_color_continuous(name="MaxO3 day before")+
  xlab("T12")+
  ylab("MaxO3")
```

Scatterplots with binning for large datasets

```
mu=rep(-5,3)
sigma=matrix(c(10,3,3,2),2,2)

simul1=as.data.frame(mvrnorm(100,rep(0,2),sigma))

ggplot(simul1,aes(x=V1,y=V2))+
  geom_point()+
  ggtitle("With ggplot2: no problem")+
  xlab("V1")+
  ylab("V2")
```

```
simul2=as.data.frame(mvrnorm(10000,rep(0,2),sigma))

ggplot(simul2,aes(x=V1,y=V2))+
  geom_point()+
  ggtitle("With ggplot2: impossible to see anything")+
  xlab("V1")+
  ylab("V2")
```

```
ggplot(simul2,aes(x=V1,y=V2))+
  geom_point()+
  stat_binhex()+
  ggtitle("With ggplot2: Better with binning")+
  xlab("V1")+
  ylab("V2")
```

4.2 qq-plots

```
qqplot(ozone$T9,ozone$T15)
```

4.3 Pairs

```r
pairs(ozone[,-c(1,6,7,8,13,14)])
```

```r
ggpairs(ozone,columns=c(2:5,9:12),title="With GGally")
```

```r
ggpairs(ozone,columns=c(2:5,9:12),ggplot2::aes(colour=rain),title="With GGally")
```

With GGally

4.4 Mosaics

```
mosaic(table(ozone[,c(13,14)]),legend=TRUE)
```

5 Time series

```
n=100
x=rnorm(n)
y=rnorm(n)

plot(1:n,x,xlab="t",ylab="",type="l",col="red",main="White noise")
abline(h=0,col="black")
```

```
wnoises=data.frame(t=1:n,x=x,y=y)

ggplot(wnoises,aes(x=t,y=x))+
  geom_line()+
  ggtitle("With ggplot2")+
  xlab("t")+
  ylab(TeX("$X_t$"))
```

```
par(mfrow=c(1,2))
plot(1:n,x,xlab="t",ylab=TeX("$X_t$"),type="l",
     col="blue",main="White noise 1/2")
abline(h=0)
plot(1:n,y,xlab="t",ylab=TeX("$Y_t$"),type="l",
     col="red",main="White noise 2/2")
abline(h=0)
```

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

lab=list(TeX("$X_t$"),TeX("$Y_t$"))

ggplot()+
  geom_line(data=wnoises,aes(x=t,y=x,colour="1"))+
  geom_line(data=wnoises,aes(x=t,y=y,colour="2"))+
  geom_hline(yintercept=0)+
  xlab("t")+
  ylab("")+
  scale_colour_manual(values=c("red","blue"),label=lab,name="Time series")
```

6 Example of export

```
jpeg(file="/Users/vincentlefieux/Documents/plot.jpeg")
pairs(ozone[,-c(1,6,7,8,13,14)])
dev.off()
```

```
## pdf
##   2
```