infert\_Datset.R

Innova

Tue Jan 17 11:39:32 2017

#infert Datset  
  
library(ggplot2)  
library(reshape2)  
library(plotly)

##   
## Attaching package: 'plotly'

## The following object is masked from 'package:ggplot2':  
##   
## last\_plot

## The following object is masked from 'package:stats':  
##   
## filter

## The following object is masked from 'package:graphics':  
##   
## layout

library(MASS)

##   
## Attaching package: 'MASS'

## The following object is masked from 'package:plotly':  
##   
## select

library(caret)

## Loading required package: lattice

head(infert)

## education age parity induced case spontaneous stratum pooled.stratum  
## 1 0-5yrs 26 6 1 1 2 1 3  
## 2 0-5yrs 42 1 1 1 0 2 1  
## 3 0-5yrs 39 6 2 1 0 3 4  
## 4 0-5yrs 34 4 2 1 0 4 2  
## 5 6-11yrs 35 3 1 1 1 5 32  
## 6 6-11yrs 36 4 2 1 1 6 36

data("infert")  
names(infert)

## [1] "education" "age" "parity" "induced"   
## [5] "case" "spontaneous" "stratum" "pooled.stratum"

ls.str(infert) #Overview of the data

## age : num [1:248] 26 42 39 34 35 36 23 32 21 28 ...  
## case : num [1:248] 1 1 1 1 1 1 1 1 1 1 ...  
## education : Factor w/ 3 levels "0-5yrs","6-11yrs",..: 1 1 1 1 2 2 2 2 2 2 ...  
## induced : num [1:248] 1 1 2 2 1 2 0 0 0 0 ...  
## parity : num [1:248] 6 1 6 4 3 4 1 2 1 2 ...  
## pooled.stratum : num [1:248] 3 1 4 2 32 36 6 22 5 19 ...  
## spontaneous : num [1:248] 2 0 0 0 1 1 0 0 1 0 ...  
## stratum : int [1:248] 1 2 3 4 5 6 7 8 9 10 ...

summary(infert) #summary of the dta

## education age parity induced   
## 0-5yrs : 12 Min. :21.00 Min. :1.000 Min. :0.0000   
## 6-11yrs:120 1st Qu.:28.00 1st Qu.:1.000 1st Qu.:0.0000   
## 12+ yrs:116 Median :31.00 Median :2.000 Median :0.0000   
## Mean :31.50 Mean :2.093 Mean :0.5726   
## 3rd Qu.:35.25 3rd Qu.:3.000 3rd Qu.:1.0000   
## Max. :44.00 Max. :6.000 Max. :2.0000   
## case spontaneous stratum pooled.stratum   
## Min. :0.0000 Min. :0.0000 Min. : 1.00 Min. : 1.00   
## 1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.:21.00 1st Qu.:19.00   
## Median :0.0000 Median :0.0000 Median :42.00 Median :36.00   
## Mean :0.3347 Mean :0.5766 Mean :41.87 Mean :33.58   
## 3rd Qu.:1.0000 3rd Qu.:1.0000 3rd Qu.:62.25 3rd Qu.:48.25   
## Max. :1.0000 Max. :2.0000 Max. :83.00 Max. :63.00

table(infert$case)

##   
## 0 1   
## 165 83

#Checking out the missing values  
sapply(infert, function(x) sum(is.na(x))) #No missing values

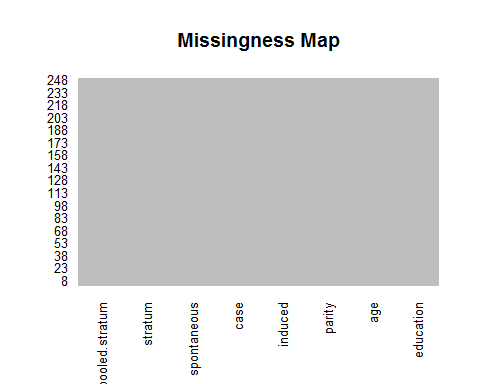
## education age parity induced case   
## 0 0 0 0 0   
## spontaneous stratum pooled.stratum   
## 0 0 0

# load libraries  
library(Amelia)

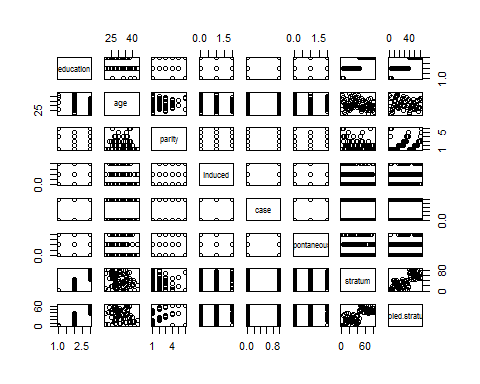
## Loading required package: Rcpp

## ##   
## ## Amelia II: Multiple Imputation  
## ## (Version 1.7.4, built: 2015-12-05)  
## ## Copyright (C) 2005-2017 James Honaker, Gary King and Matthew Blackwell  
## ## Refer to http://gking.harvard.edu/amelia/ for more information  
## ##

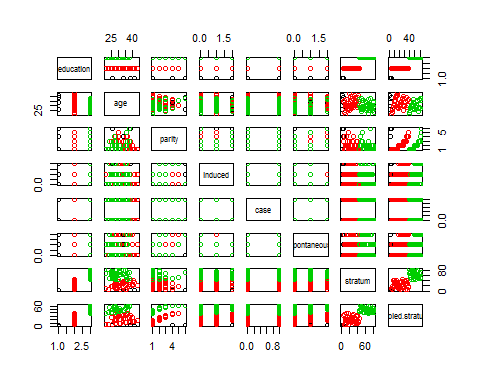
library(mlbench)  
# create a missing map  
missmap(infert, col=c("black", "grey"), legend=FALSE)



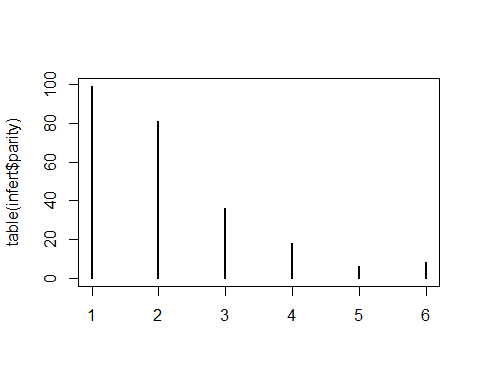
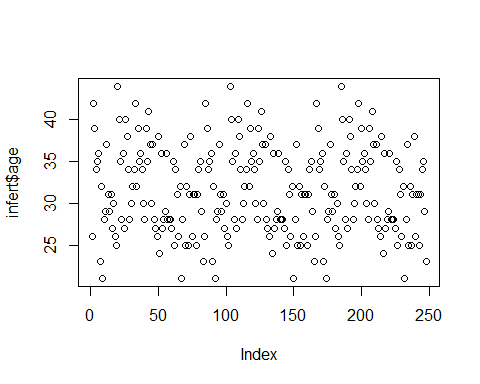
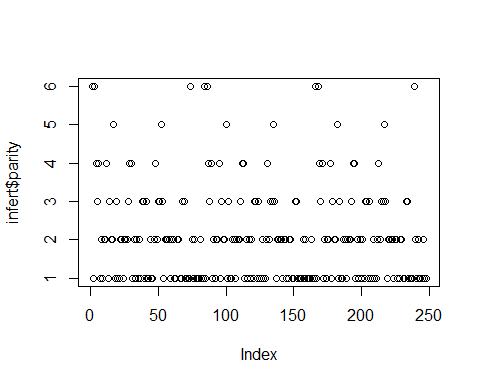
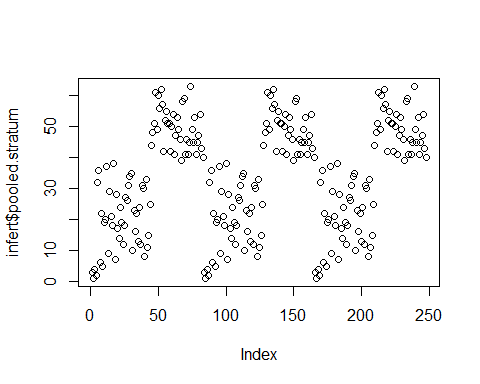
pairs(infert)



pairs(education~., data = infert, col=infert$education)



par(mfrow=c(2,2),  
plot(infert$pooled.stratum),  
plot(infert$parity),  
plot(infert$age),  
plot(table(infert$parity)))



dev.off()

## null device   
## 1

hist(infert$age)  
  
require(stats)  
model1 <- glm(case ~ spontaneous+induced, data = infert, family = binomial())  
summary(model1)

##   
## Call:  
## glm(formula = case ~ spontaneous + induced, family = binomial(),   
## data = infert)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.6678 -0.8360 -0.5772 0.9030 1.9362   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -1.7079 0.2677 -6.380 1.78e-10 \*\*\*  
## spontaneous 1.1972 0.2116 5.657 1.54e-08 \*\*\*  
## induced 0.4181 0.2056 2.033 0.042 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 316.17 on 247 degrees of freedom  
## Residual deviance: 279.61 on 245 degrees of freedom  
## AIC: 285.61  
##   
## Number of Fisher Scoring iterations: 4

model1

##   
## Call: glm(formula = case ~ spontaneous + induced, family = binomial(),   
## data = infert)  
##   
## Coefficients:  
## (Intercept) spontaneous induced   
## -1.7079 1.1972 0.4181   
##   
## Degrees of Freedom: 247 Total (i.e. Null); 245 Residual  
## Null Deviance: 316.2   
## Residual Deviance: 279.6 AIC: 285.6

library(survival)

##   
## Attaching package: 'survival'

## The following object is masked from 'package:caret':  
##   
## cluster

data(infert)  
model2 <- glm(case ~ age+parity+education+spontaneous+induced,  
 data = infert, family = binomial())  
summary(model2)

##   
## Call:  
## glm(formula = case ~ age + parity + education + spontaneous +   
## induced, family = binomial(), data = infert)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.7603 -0.8162 -0.4956 0.8349 2.6536   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -1.14924 1.41220 -0.814 0.4158   
## age 0.03958 0.03120 1.269 0.2046   
## parity -0.82828 0.19649 -4.215 2.49e-05 \*\*\*  
## education6-11yrs -1.04424 0.79255 -1.318 0.1876   
## education12+ yrs -1.40321 0.83416 -1.682 0.0925 .   
## spontaneous 2.04591 0.31016 6.596 4.21e-11 \*\*\*  
## induced 1.28876 0.30146 4.275 1.91e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 316.17 on 247 degrees of freedom  
## Residual deviance: 257.80 on 241 degrees of freedom  
## AIC: 271.8  
##   
## Number of Fisher Scoring iterations: 4

spontaneous = infert$spontaneous  
age = infert$age  
stratum = infert$stratum  
  
  
plot( stratum, age )  
abline(lm( age ~ stratum)) #generate a linear regression model of the two variables  
  
library('epiDisplay')

## Loading required package: foreign

## Loading required package: nnet

##   
## Attaching package: 'epiDisplay'

## The following object is masked from 'package:lattice':  
##   
## dotplot

## The following object is masked from 'package:ggplot2':  
##   
## alpha

model1 <- glm(case ~ induced + spontaneous, data=infert, family=binomial)  
logistic.display(model1)

##   
## Logistic regression predicting case   
##   
## crude OR(95%CI) adj. OR(95%CI) P(Wald's test)  
## induced (cont. var.) 1.05 (0.74,1.5) 1.52 (1.02,2.27) 0.042   
##   
## spontaneous (cont. var.) 2.9 (1.97,4.26) 3.31 (2.19,5.01) < 0.001   
##   
## P(LR-test)  
## induced (cont. var.) 0.042   
##   
## spontaneous (cont. var.) < 0.001   
##   
## Log-likelihood = -139.806  
## No. of observations = 248  
## AIC value = 285.612

# Having two spontaneous abortions is quite close to being infertile!  
# This is actually not a causal relationship  
layout(matrix(1, ncol = 1))  
lroc(model1, title=TRUE, auc.coords=c(.5,.1))

## $model.description  
## [1] "case ~ induced + spontaneous"  
##   
## $auc  
## [1] 0.7285506  
##   
## $predicted.table  
## predicted.prob Non-diseased Diseased  
## 0.1534 60 7  
## 0.2158 33 12  
## 0.2949 20 9  
## 0.3750 25 22  
## 0.4768 11 5  
## 0.5806 4 4  
## 0.6651 11 18  
## 0.7511 1 6  
##   
## $diagnostic.table  
## 1-Specificity Sensitivity  
## 1.000000000 1.00000000  
## > 0.636363636 0.91566265  
## > 0.436363636 0.77108434  
## > 0.315151515 0.66265060  
## > 0.163636364 0.39759036  
## > 0.096969697 0.33734940  
## > 0.072727273 0.28915663  
## > 0.006060606 0.07228916  
## > 0.000000000 0.00000000

# For PowerPoint presentation, the graphic elements should be enhanced as followed   
lroc(model1, title=TRUE, cex.main=2, cex.lab=1.5, col.lab="blue", cex.axis=1.3, lwd=3)

## $model.description  
## [1] "case ~ induced + spontaneous"  
##   
## $auc  
## [1] 0.7285506  
##   
## $predicted.table  
## predicted.prob Non-diseased Diseased  
## 0.1534 60 7  
## 0.2158 33 12  
## 0.2949 20 9  
## 0.3750 25 22  
## 0.4768 11 5  
## 0.5806 4 4  
## 0.6651 11 18  
## 0.7511 1 6  
##   
## $diagnostic.table  
## 1-Specificity Sensitivity  
## 1.000000000 1.00000000  
## > 0.636363636 0.91566265  
## > 0.436363636 0.77108434  
## > 0.315151515 0.66265060  
## > 0.163636364 0.39759036  
## > 0.096969697 0.33734940  
## > 0.072727273 0.28915663  
## > 0.006060606 0.07228916  
## > 0.000000000 0.00000000

lroc1 <- lroc(model1) # The main title and auc text have disappeared  
  
model2 <- glm(case ~ spontaneous, data=infert, family=binomial)  
logistic.display(model2)

##   
## Logistic regression predicting case   
##   
## OR(95%CI) P(Wald's test) P(LR-test)  
## spontaneous (cont. var.) 2.9 (1.97,4.26) < 0.001 < 0.001   
##   
## Log-likelihood = -141.8808  
## No. of observations = 248  
## AIC value = 287.7616

lroc2 <- lroc(model2, add=TRUE, line.col="brown", lty=2)  
legend("bottomright",legend=c(lroc1$model.description, lroc2$model.description),  
 lty=1:2, col=c("red","brown"), bg="white")  
title(main="Comparison of two logistic regression models")  
lrtest(model1, model2)

## Likelihood ratio test for MLE method   
## Chi-squared 1 d.f. = 4.149652 , P value = 0.04164309