Fish Weight & Length & Width

R Practice 5: correlation & regression

ALY6010

Prepared by: Heejae Roh
Presented to: Professor Behzad Ahmadi
Date: Dec 11th, 2022

INTRODUCTION

In this project I will examine and test the relationships between several variables using Fish data. Correlation and regression are a bit confusing concept for me. I will try to get a clearer idea of what each of these concepts means. I hope to clarify what the difference between the two is.

PART 0. SIMPLE EDA

Headtail of Data

	Species	Weight(g)	Length.V(cm)	Length.D(cm)	Length.C(cm)	Height(cm)	Width(cm)
1	Bream	242	23.2	25.4	30	11.52	4.02
2	Bream	290	24	26.3	31.2	12.48	4.31
3	Bream	340	23.9	26.5	31.1	12.38	4.7
…				…			
157	Smelt	12.2	12.1	13	13.8	2.28	1.26
158	Smelt	19.7	13.2	14.3	15.2	2.87	2.07
159	Smelt	19.9	13.8	15	16.2	2.93	1.88

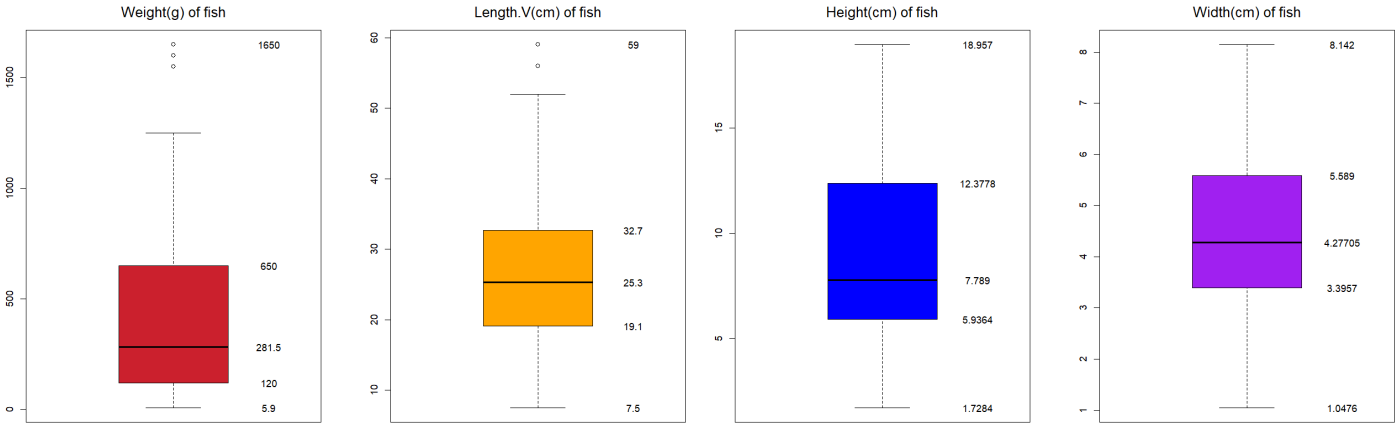
Descriptive Analysis

	Mean	Sd	Median	Trimmed	Mad	Min	Max	Range	Skew	Kurtosis	se
Weight(g)	**398.33**	**357.98**	**273.00**	**356.95**	**302.45**	**0.00**	**1650**	**1650**	**1.08**	**0.77**	**28.39**
Length.V(cm)	**26.25**	**10.00**	**25.20**	**25.79**	**9.79**	**7.50**	**59.00**	**51.50**	**0.58**	**0.35**	**0.79**
Length.D(cm)	**28.42**	**10.72**	**27.30**	**28.01**	**10.82**	**8.40**	**63.40**	**55.00**	**0.53**	**0.31**	**0.85**
Length.C(cm)	**31.23**	**11.61**	**29.40**	**30.97**	**12.75**	**8.80**	**68.00**	**59.20**	**0.38**	**0.00**	**0.92**
Height(cm)	**8.97**	**4.29**	**7.79**	**8.82**	**4.01**	**1.73**	**18.96**	**17.23**	**0.39**	**-0.66**	**0.34**
Width(cm)	**4..42**	**1.69**	**4.25**	**4.44**	**1.53**	**1.05**	**8.14**	**7.09**	**0.00**	**-0.59**	**0.13**

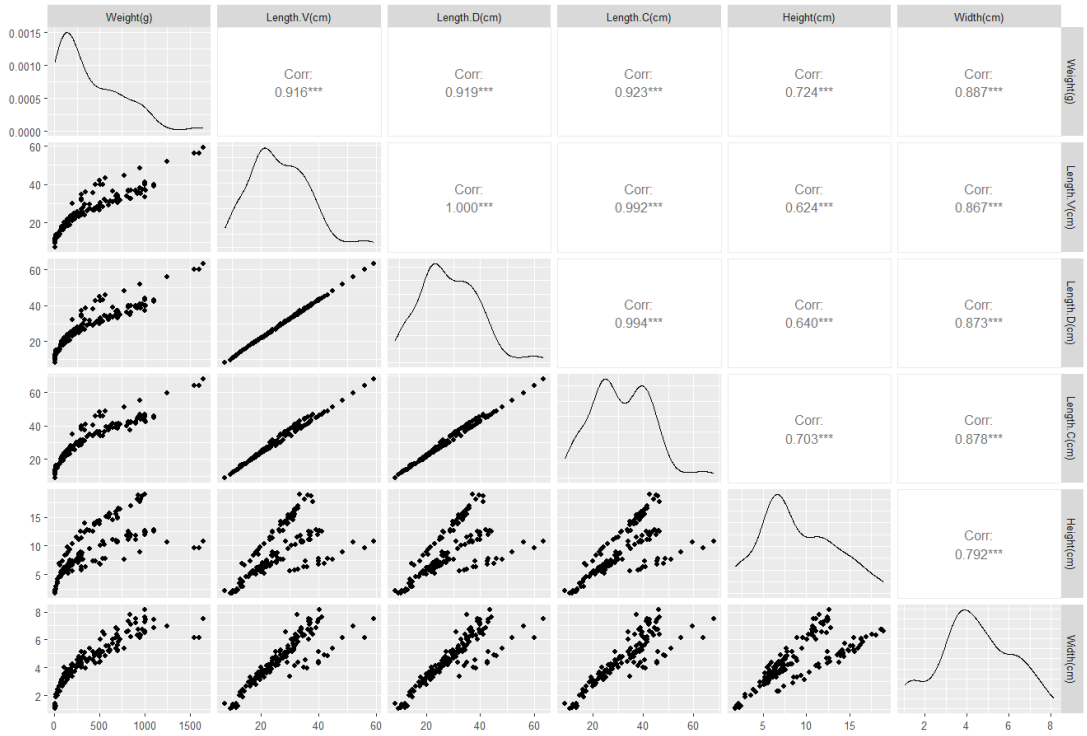
Data Cleaning:

1. Changing Colnames with unit
2. Checking Is there NA value in data
3. After descriptive analysis delete weight=0 row as outliers

4. Check boxplot with variables and leave every data (No outlier except weight=0)



Scatter plot & density plot & Pearson correlation



Finding:

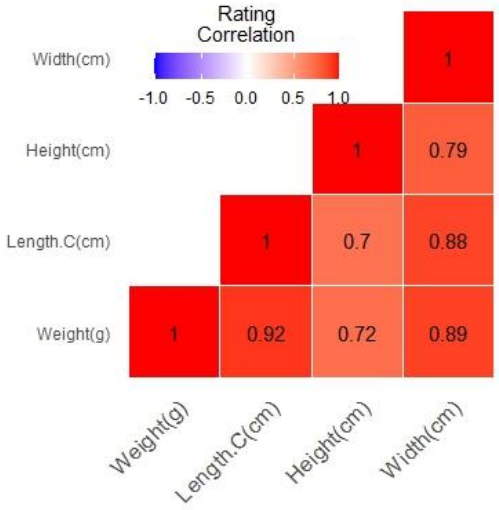
1. Correlation between Length.V and Length.D is 1.00
2. Correlation between Length.V and Length.C is 0.99
   ➔ Decide to analyze just one of Length data which is Length.C
3. It looks like Correlation with other is Weight > Width > Length.C > Height
4. It looks like the relationship between Weight and other variables isn't linear.
5. It looks like the relationship between Height and other variables isn't quite linear.

PART 1. CORRELATION

Definition

1. The correlation coefficient is a statistical measure of the strength of a linear relationship between two variables. Correlation coefficients are used in science and in finance to assess the degree of association between two variables (Jason, 2021).
2. Values always range from -1 for a perfectly inverse, or negative, relationship to 1 for a perfectly positive correlation. Values at, or close to, zero indicate no linear relationship or a very weak correlation (Jason, 2021).
3. There are three correlation method in R which are pearson, kendall, spearman
4. Pearson r correlation is the most widely used correlation statistic to measure the degree of the relationship between linearly related variables (Complete Dissertation [CD], n.d.).
5. Kendall rank correlation is a non-parametric test that measures the strength of dependence between two variables (CD, n.d).
6. Spearman rank correlation is a non-parametric test that is used to measure the degree of association between two variables (CD, n.d).

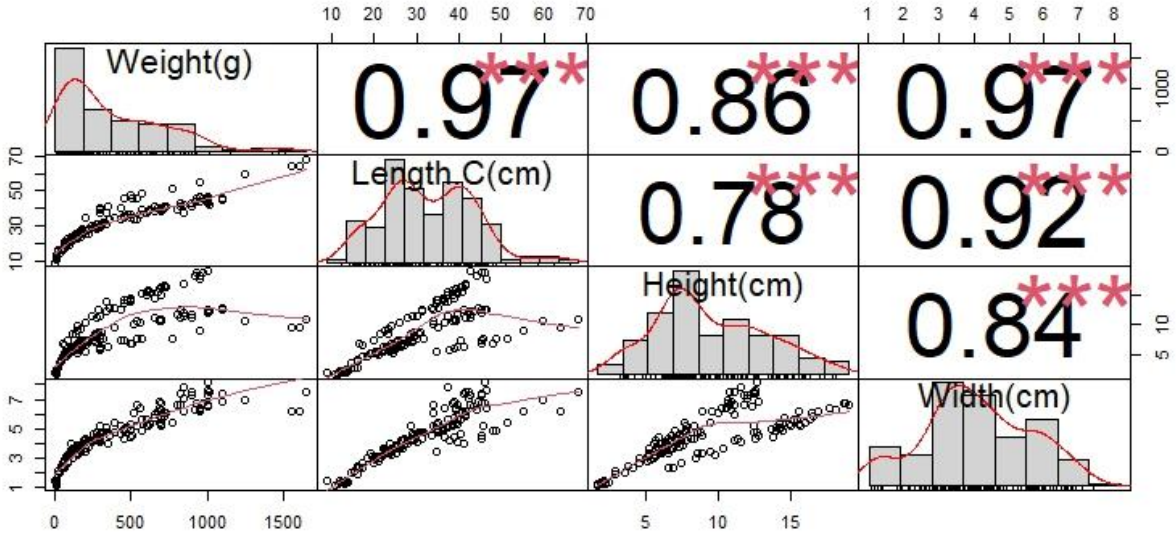
Correlation Heatmap from cor() function with no assigned 'method'



Interpretation:

1. I use cor() with default method which calculate pearson correlation cofficient
2. The Correlation between Weight and Length.C is 0.92 > Width 0.89 > Height 0.72
3. The Correlation between Length.C and Width is 0.88 > Height 0.7
4. The Correlation between Width and Height 0.79

Correlation chart with ‘Spearman’ method

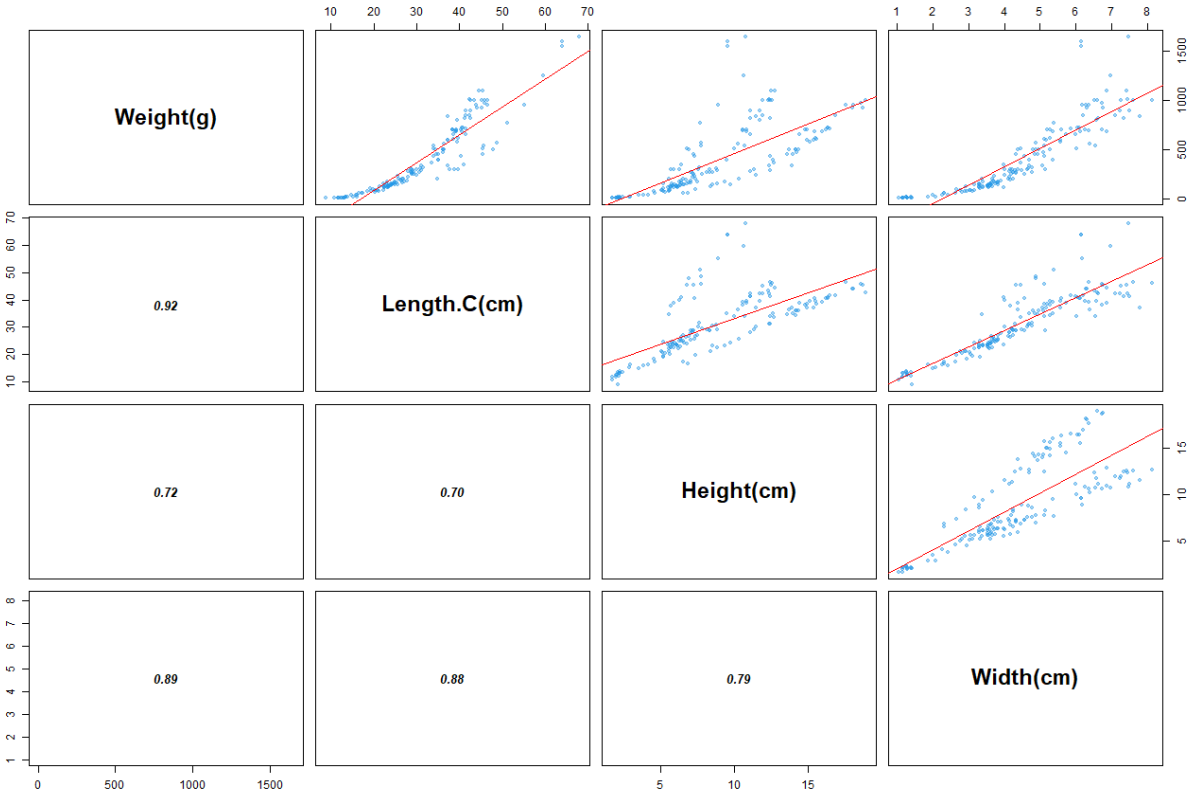


Interpretation:

1. Spearman rank correlation uses ranking order
2. Make no assumptions about the distribution of data in the population.
3. It is effective when the probability distribution of the population is not normally distributed.
4. It can be applied even when the sample size is small.

PART 2. REGRESSION

Correlation chart with Pearson method



Interpretation:

1. In this case I use lm (linear model), so the numbers are same as heatmap. Because both use pearson correlation coefficient to calculate correlation and regression slope.
2. The Correlation between Weight and Length.C is 0.92, it is the strongest linear relationship in this dataset.
3. The Correlation between Length.C and Height is 0.7, it is the weakest linear relationship in this dataset.

Question & Answer: How does regression analysis differ from correlation analysis?

1. Correlation quantifies the strength of the linear relationship between a pair of variables equation (Viv, Liz, and Jonathan [VLJ], 2003).
2. Whereas regression expresses the relationship in the form of an equation (VLJ, 2003)
3. Correlation determines the association and represents a linear relationship. There is no difference between dependent and independent variables. It just analyzes the correlation. Correlation tells us how two variables behave.

4. Regression analysis describes with an equation how the independent variable is numerically related to the dependent variable. It seeks to find the best equation and explain the dependent variable through the independent variable. There is a clear difference between dependent and independent variables. Regression analyzes the effect of x on the estimated variable y.

CONCLUSION

In this chapter, I study about the difference between correlation and regression. The most confusing part was to tell the difference between the two. In correlation there is no difference between the independent variable and the dependent variable. Regression is to find an equation that has an independent variable and a dependent variable and predicts the dependent variable y according to the independent variable.

The reason I was confused is that both correlation and regression learn based on the Pearson method. This is because linear correlation is found with the pearson method when searching for correlation, and linear regression also. In other words, the basic regression, comes from the pearson method. However, I need to know the difference between the two methods and apply the. I also learn when to use other methods like spearman, and learned a linear relationship is not the only one in regression.

REFERENCE

Bluman, Allan. (2017). Elementary statistics: a step by step approach 10th edition. McGraw-Hill.

AUNG PYAE. (2019). Fish market. kaggle. Retrieved from https://www.kaggle.com/datasets/aungpyaeap/fish-market

GAETANO CHIRIACO. (2019). Weight Prediction - LogLog Linear Regression. kaggle. Retrieved from https://www.kaggle.com/code/gaetanochiriaco/weight-prediction-loglog-linear-regression

codersgram9. (2021, March 16). Change column name of a given dataframe in R. geeksforgeeks. Retrieved from https://www.geeksforgeeks.org/change-column-name-of-a-given-dataframe-in-r/

Statistics Globe. (n.d.). R find missing values (6 examples for data frame, column & vector). Retrieved from https://statisticsglobe.com/r-find-missing-values/

Gottumukkala Sravan Kumar. (2022, November 10). How to delete rows in R? explained with examples. SparkBy Examples. Retrieved from https://sparkbyexamples.com/r-programming/drop-dataframe-rows-in-r/

finnstats. (2021, June 24). ggpairs in R- a brief introduction to ggpairs. R-bloggers. Retrieved from https://www.r-bloggers.com/2021/06/ggpairs-in-r-a-brief-introduction-to-ggpairs/

R CODER. (n.d). Correlation plot in R. Retrieved from https://r-coder.com/correlation-plot-r/

Jason Fernando. (2021, October 5). The correlation coefficient: what it is, what it tells investors. Investopedia. Retrieved from https://www.investopedia.com/terms/c/correlationcoefficient.asp

Complete Dissertation [CD]. (n.d.). Correlation (Pearson, Kendall, Spearman). Retrieved from https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/correlation-pearson-kendall-spearman/

Viv Bewick, Liz Cheek, and Jonathan Ball. (2003). Statistics review 7: correlation and regression. National Library of Medicine. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC374386/

R-Codes

```
library("psych")
library("ggplot2")
library("reshape2")

#Setwd in file direction
setwd("C:\\Users\\14083\\Desktop")
#Import dataset using read.csv()
fdata <- read.csv("fish.csv", stringsAsFactors = T,
                  header=T)

#Checking dataset & data structure
headtail(fdata, 5)

#change colnames
colnames(fdata) <- c("Species", "Weight(g)", "Length.V(cm)", "Length.D(cm)", "Length.C(cm)",
"Height(cm)", "Width(cm)")
str(fdata)

#check NA value
which(is.na(fdata$Species))
which(is.na(fdata$`Weight(g)`))
which(is.na(fdata$`Length.V(cm)`))
which(is.na(fdata$`Length.D(cm)`))
which(is.na(fdata$`Length.C(cm)`))
which(is.na(fdata$`Height(cm)`))
which(is.na(fdata$`Width(cm)`))

# descriptive data analysis
describe(fdata)

#find min=0 weight value
t <- subset(fdata, `Weight(g)`== 0,
            select = Species:`Width(cm)`)

#delete min=0 value
fdata2 <- fdata[-41,]

#see descriptive value again
describe(fdata2)

#checking outliers with boxplot: Weight(g) 1
opar <- par(no.readonly = TRUE)
par(fig=c(0, 0.5, 0, 1))
boxplot(fdata2$`Weight(g)`, col='#cb202d')
```

```
text(y=fivenum(fdata2$`Weight(g)`), labels=fivenum(fdata2$`Weight(g)`), x=1.35, cex=1)
mtext("Weight(g) of fish", side=3, line=1, cex=1.5)

#checking outliers with boxplot: Length.V(cm) 2
par(fig=c(0.5, 1, 0, 1), new=TRUE)
boxplot(fdata2$`Length.V(cm)`, col='orange')
text(y=fivenum(fdata2$`Length.V(cm)`), labels=fivenum(fdata2$`Length.V(cm)`), x=1.35, cex=1)
mtext("Length.V(cm) of fish", side=3, line=1, cex=1.5)

#checking outliers with boxplot: Length.D(cm) 3
opar <- par(no.readonly = TRUE)
par(fig=c(0, 0.5, 0, 1))
boxplot(fdata2$`Length.D(cm)`, col='yellow')
text(y=fivenum(fdata2$`Length.D(cm)`), labels=fivenum(fdata2$`Length.D(cm)`), x=1.35, cex=1)
mtext("Length.D(cm) of fish", side=3, line=1, cex=1.5)

#checking outliers with boxplot: Length.C(cm) 4
par(fig=c(0.5, 1, 0, 1), new=TRUE)
boxplot(fdata2$`Length.C(cm)`, col='green')
text(y=fivenum(fdata2$`Length.C(cm)`), labels=fivenum(fdata2$`Length.C(cm)`), x=1.35, cex=1)
mtext("Length.C(cm) of fish", side=3, line=1, cex=1.5)

#checking outliers with boxplot: Height(cm) 5
opar <- par(no.readonly = TRUE)
par(fig=c(0, 0.5, 0, 1))
boxplot(fdata2$`Height(cm)`, col='blue')
text(y=fivenum(fdata2$`Height(cm)`), labels=fivenum(fdata2$`Height(cm)`), x=1.35, cex=1)
mtext("Height(cm) of fish", side=3, line=1, cex=1.5)

#checking outliers with boxplot: Width(cm) 6
par(fig=c(0.5, 1, 0, 1), new=TRUE)
boxplot(fdata2$`Width(cm)`, col='purple')
text(y=fivenum(fdata2$`Width(cm)`), labels=fivenum(fdata2$`Width(cm)`), x=1.35, cex=1)
mtext("Width(cm) of fish", side=3, line=1, cex=1.5)
par(opar)

#ggpairs before regression
install.packages("GGally")
library("GGally")
str(fdata2)
ggpairs(fdata2[2:7])

#Correlation heatmap
str(fdata2)
cordata3 <- fdata2[,c(2,5:7)]
cordata3
```

```
cor(cordata3)
cor(cordata3, method="spearman")

aR <- round(cor(cordata3), 2)

get_upper_tri<-function(aR){
  aR[lower.tri(aR)] <- NA
  return(aR) }

upper_tri <- get_upper_tri(aR)
upper_tri
melted_cormat <- melt(upper_tri, na.rm = TRUE)
melted_cormat

reorder_cormat <- function(aR){
  # Use correlation between variables as distance
  dd <- as.dist((1-cormat)/2)
  hc <- hclust(dd)
  aR <-aR[hc$order, hc$order] }

aR <- reorder_cormat(aR)
aR
upper_tri <- get_upper_tri(aR)

melted_cormat <- melt(upper_tri, na.rm = TRUE)

#NEW
ggheatmap <- ggplot(melted_cormat, aes(Var2, Var1, fill = value))+
  geom_tile(color = "white")+
  scale_fill_gradient2(low = "blue", high = "red", mid = "white",
                       midpoint = 0, limit = c(-1,1), space = "Lab",
                       name="Rating\nCorrelation") +
  theme_minimal()+ # minimal theme
  theme(axis.text.x = element_text(angle = 45, vjust = 1,
                                   size = 12, hjust = 1))+
  coord_fixed()

ggheatmap +
  geom_text(aes(Var2, Var1, label = value), color = "black", size = 4) +
  theme(
    axis.title.x = element_blank(),
    axis.title.y = element_blank(),
    panel.grid.major = element_blank(),
    panel.border = element_blank(),
    panel.background = element_blank(),
```

```
    axis.ticks = element_blank(),
    legend.justification = c(1, 0),
    legend.position = c(0.6, 0.7),
    legend.direction = "horizontal")+
  guides(fill = guide_colorbar(barwidth = 7, barheight = 1,
                               title.position = "top", title.hjust = 0.5))

#correlation chart
install.packages("PerformanceAnalytics")
library(MASS)
print(.libPaths())
library("PerformanceAnalytics")
chart.Correlation(cordata3, histogram = TRUE, method = "spearman")

#linear regression
Length.C_Weight <- ggplot(fdata2,aes(x=`Length.C(cm)`,y=`Weight(g)`,col=Species))+
  geom_point(aes(size=2 ,alpha=0.6))+
  geom_smooth(col="red",method = "lm",se=F, lwd=0.9,formula="y~x")
Length.C_Weight

#regression only
panel.cor <- function(x, y, digits = 2, prefix = "", cex.cor, ...)
{
  usr <- par("usr"); on.exit(par(usr))
  par(usr = c(0, 1, 0, 1))
  r <- abs(cor(x, y))
  txt <- format(c(r, 0.123456789), digits = digits)[1]
  txt <- paste0(prefix, txt)
  text(0.5, 0.5, txt, cex = 1.1, font = 4) }

reg <- function(x, y, col) abline(lm(y~x), col=col)

panel.lm =  function (x, y, col = par("col"), bg = NA, pch = par("pch"),
                      cex = 1, col.smooth = "red", span = 2/3, iter = 3, ...)  {
  points(x, y, pch = pch, col = col, bg = bg, cex = cex)
  ok <- is.finite(x) & is.finite(y)
  if (any(ok)) reg(x[ok], y[ok], col.smooth) }

str(cordata3)
pairs(cordata3[1:4], panel = panel.lm,
      cex = 0.7, pch = 19, col = adjustcolor(4, .4), cex.labels = 2,
      font.labels = 2, lower.panel = panel.cor)

pairs(cordata3[1:4], panel = panel.lm,
      cex = 0.6, pch = 19, col = cordata3$rating, cex.labels = 2,
      font.labels = , lower.panel = panel.cor)
```