# Exercises from Cookbook for R

# Correlation

# Make some data  
# X increases (noisily)  
# Z increases slowly  
# Y is constructed so it is inversely related to xvar and positively related to xvar\*zvar  
set.seed(955)  
xvar <- 1:20 + rnorm(20,sd=3)  
zvar <- 1:20/4 + rnorm(20,sd=2)  
yvar <- -2\*xvar + xvar\*zvar/5 + 3 + rnorm(20,sd=4)  
  
# Make a data frame with the variables  
dat <- data.frame(x=xvar, y=yvar, z=zvar)  
# Show first few rows  
head(dat)

## x y z  
## 1 -4.252354 4.5857688 1.89877152  
## 2 1.702318 -4.9027824 -0.82937359  
## 3 4.323054 -4.3076433 -1.31283495  
## 4 1.780628 0.2050367 -0.28479448  
## 5 11.537348 -29.7670502 -1.27303976  
## 6 6.672130 -10.1458220 -0.09459239

cor(dat$x,dat$y)

## [1] -0.7695378

cor(dat$y,dat$z)

## [1] 0.004172295

cor(dat)

## x y z  
## x 1.0000000 -0.769537849 0.491698938  
## y -0.7695378 1.000000000 0.004172295  
## z 0.4916989 0.004172295 1.000000000

# Regression Example

# These two commands will have the same outcome:  
fit <- lm(y ~ x, data=dat) # Using the columns x and y from the data frame  
fit <- lm(dat$y ~ dat$x) # Using the vectors dat$x and dat$y  
fit

##   
## Call:  
## lm(formula = dat$y ~ dat$x)  
##   
## Coefficients:  
## (Intercept) dat$x   
## -0.2278 -1.1829

summary(fit)

##   
## Call:  
## lm(formula = dat$y ~ dat$x)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -15.8922 -2.5114 0.2866 4.4646 9.3285   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -0.2278 2.6323 -0.087 0.932   
## dat$x -1.1829 0.2314 -5.113 7.28e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6.506 on 18 degrees of freedom  
## Multiple R-squared: 0.5922, Adjusted R-squared: 0.5695   
## F-statistic: 26.14 on 1 and 18 DF, p-value: 7.282e-05

# t-test Example

sleep

## extra group ID  
## 1 0.7 1 1  
## 2 -1.6 1 2  
## 3 -0.2 1 3  
## 4 -1.2 1 4  
## 5 -0.1 1 5  
## 6 3.4 1 6  
## 7 3.7 1 7  
## 8 0.8 1 8  
## 9 0.0 1 9  
## 10 2.0 1 10  
## 11 1.9 2 1  
## 12 0.8 2 2  
## 13 1.1 2 3  
## 14 0.1 2 4  
## 15 -0.1 2 5  
## 16 4.4 2 6  
## 17 5.5 2 7  
## 18 1.6 2 8  
## 19 4.6 2 9  
## 20 3.4 2 10

sleep\_wide <- data.frame(  
 ID=1:10,  
 group1=sleep$extra[1:10],  
 group2=sleep$extra[11:20]  
)  
sleep\_wide

## ID group1 group2  
## 1 1 0.7 1.9  
## 2 2 -1.6 0.8  
## 3 3 -0.2 1.1  
## 4 4 -1.2 0.1  
## 5 5 -0.1 -0.1  
## 6 6 3.4 4.4  
## 7 7 3.7 5.5  
## 8 8 0.8 1.6  
## 9 9 0.0 4.6  
## 10 10 2.0 3.4

# Welch t-test  
t.test(extra ~ group, sleep)

##   
## Welch Two Sample t-test  
##   
## data: extra by group  
## t = -1.8608, df = 17.776, p-value = 0.07939  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -3.3654832 0.2054832  
## sample estimates:  
## mean in group 1 mean in group 2   
## 0.75 2.33

t.test(sleep\_wide$group1, sleep\_wide$group2)

##   
## Welch Two Sample t-test  
##   
## data: sleep\_wide$group1 and sleep\_wide$group2  
## t = -1.8608, df = 17.776, p-value = 0.07939  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -3.3654832 0.2054832  
## sample estimates:  
## mean of x mean of y   
## 0.75 2.33

# ANOVA

data <- read.table(header=TRUE, text='  
 subject sex age before after  
 1 F old 9.5 7.1  
 2 M old 10.3 11.0  
 3 M old 7.5 5.8  
 4 F old 12.4 8.8  
 5 M old 10.2 8.6  
 6 M old 11.0 8.0  
 7 M young 9.1 3.0  
 8 F young 7.9 5.2  
 9 F old 6.6 3.4  
 10 M young 7.7 4.0  
 11 M young 9.4 5.3  
 12 M old 11.6 11.3  
 13 M young 9.9 4.6  
 14 F young 8.6 6.4  
 15 F young 14.3 13.5  
 16 F old 9.2 4.7  
 17 M young 9.8 5.1  
 18 F old 9.9 7.3  
 19 F young 13.0 9.5  
 20 M young 10.2 5.4  
 21 M young 9.0 3.7  
 22 F young 7.9 6.2  
 23 M old 10.1 10.0  
 24 M young 9.0 1.7  
 25 M young 8.6 2.9  
 26 M young 9.4 3.2  
 27 M young 9.7 4.7  
 28 M young 9.3 4.9  
 29 F young 10.7 9.8  
 30 M old 9.3 9.4  
')

# One way between:  
# IV: sex  
# DV: before  
aov1 <- aov(before ~ sex, data=data)  
summary(aov1)

## Df Sum Sq Mean Sq F value Pr(>F)  
## sex 1 1.53 1.529 0.573 0.455  
## Residuals 28 74.70 2.668

# 2x2 between:  
# IV: sex  
# IV: age  
# DV: after  
# These two calls are equivalent  
aov2 <- aov(after ~ sex\*age, data=data)  
summary(aov2)

## Df Sum Sq Mean Sq F value Pr(>F)   
## sex 1 16.08 16.08 4.038 0.0550 .   
## age 1 38.96 38.96 9.786 0.0043 \*\*   
## sex:age 1 89.61 89.61 22.509 6.6e-05 \*\*\*  
## Residuals 26 103.51 3.98   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

TukeyHSD(aov2)

## Tukey multiple comparisons of means  
## 95% family-wise confidence level  
##   
## Fit: aov(formula = after ~ sex \* age, data = data)  
##   
## $sex  
## diff lwr upr p adj  
## M-F -1.519139 -3.073025 0.03474709 0.0549625  
##   
## $age  
## diff lwr upr p adj  
## young-old -2.31785 -3.846349 -0.7893498 0.0044215  
##   
## $`sex:age`  
## diff lwr upr p adj  
## M:old-F:old 2.8971429 -0.3079526 6.1022384 0.0869856  
## F:young-F:old 2.1733333 -1.1411824 5.4878491 0.2966111  
## M:young-F:old -2.2183333 -5.1319553 0.6952887 0.1832890  
## F:young-M:old -0.7238095 -3.7691188 2.3214997 0.9138789  
## M:young-M:old -5.1154762 -7.7187601 -2.5121923 0.0000676  
## M:young-F:young -4.3916667 -7.1285380 -1.6547953 0.0008841