Lab\_Assignment\_3.R

Owner

Fri Oct 21 09:11:30 2016

# Lab Assignment #3  
# Student Name: KEY   
#   
# Instructions: Complete the tasks outlined in this document.   
# You should upload your completed assignment online as a .pdf,   
# along with your R file.  
#   
# Problem 1 : Cool Problems  
#   
# 1A. Load the dataset “2cool4school.csv” into R.   
# These data are from Berkeley and Stanford students   
# who answered questions about their love for coffee,   
# and were then rated on how cool they were by an unbiased observer.   
#library(data.table)  
cool <- read.csv('2cool4school.csv')  
  
  
# 1A. Data Cleaning (again)  
# Last time, we did some data cleaning: (A) converted variable SEX   
# to be a factor variable with three levels (1 = female; 2 = male; 3 = other),   
# (B) changed the names of the levels of the factor variable SCH so Cal =   
# Public and Stanf = Private, and (C) created a scale score for the variable   
# coolness called coolness2. Copy and paste your code from the last lab to repeat   
# these three steps (look at the key for the lab assignment if you want to make   
# sure you are cleaning the data correctly). Use the write.csv() function to save   
# this cleaned data object as “2cool4school\_clean.csv” so you don’t have to repeat   
# these steps again.  
  
#view some of the records  
head(cool)

## cool1 cool2r cool3 cool4 cool5r cool6 cool7r cool8 coolness SCH  
## 1 4 4 5 4 3 2 4 4 3.250 Stanf  
## 2 4 4 4 4 4 4 4 4 3.250 Stanf  
## 3 4 3 4 3 3 3 4 4 3.250 Stanf  
## 4 4 4 4 4 4 2 4 4 3.000 Stanf  
## 5 3 4 4 4 3 4 3 4 3.375 Stanf  
## 6 4 1 4 4 2 4 2 5 4.250 Stanf  
## coffee.love SEX AGE  
## 1 6 2 23  
## 2 7 1 21  
## 3 1 2 22  
## 4 7 3 20  
## 5 5 2 18  
## 6 5 2 19

#number of rows   
nrow(cool)

## [1] 5738

#summary of sex variable  
summary(cool$SEX)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1.000 1.000 1.000 1.503 2.000 3.000

#create a copy of SEX  
cool$sex <- cool$SEX  
  
#as factor  
cool$sex <- as.factor(cool$SEX)  
  
#levels  
levels(cool$sex)

## [1] "1" "2" "3"

#summary  
summary(cool$sex)

## 1 2 3   
## 2879 2830 29

#summary  
summary(cool$SCH)

## Cal Stanf   
## 2870 2868

cool$SchoolF <- cool$SCH  
  
levels(cool$SchoolF)

## [1] "Cal" "Stanf"

levels(cool$SchoolF) <- c('public','private')  
  
#summary  
summary(cool$SchoolF)

## public private   
## 2870 2868

#describe using psych package  
library(psych)  
describe(cool)

## vars n mean sd median trimmed mad min max range skew  
## cool1 1 5616 3.70 1.15 4.00 3.81 1.48 1 5.0 4.0 -0.76  
## cool2r 2 5598 3.33 1.10 4.00 3.36 1.48 1 5.0 4.0 -0.45  
## cool3 3 5610 3.56 1.01 4.00 3.60 1.48 1 5.0 4.0 -0.43  
## cool4 4 5594 3.58 1.00 4.00 3.63 1.48 1 5.0 4.0 -0.47  
## cool5r 5 5602 3.17 1.16 3.00 3.19 1.48 1 5.0 4.0 -0.23  
## cool6 6 5598 3.31 1.10 4.00 3.33 1.48 1 5.0 4.0 -0.34  
## cool7r 7 5602 3.51 1.06 4.00 3.57 1.48 1 5.0 4.0 -0.66  
## cool8 8 5602 3.69 0.99 4.00 3.77 1.48 1 5.0 4.0 -0.68  
## coolness 9 5518 3.23 0.75 3.25 3.23 0.74 1 5.0 4.0 -0.06  
## SCH\* 10 5738 1.50 0.50 1.00 1.50 0.00 1 2.0 1.0 0.00  
## coffee.love 11 5738 5.45 2.64 5.00 5.44 2.97 1 10.0 9.0 0.02  
## SEX 12 5738 1.50 0.51 1.00 1.50 0.00 1 3.0 2.0 0.10  
## AGE 13 5738 21.12 4.28 20.00 20.49 1.48 18 64.8 46.8 6.08  
## sex\* 14 5738 1.50 0.51 1.00 1.50 0.00 1 3.0 2.0 0.10  
## SchoolF\* 15 5738 1.50 0.50 1.00 1.50 0.00 1 2.0 1.0 0.00  
## kurtosis se  
## cool1 -0.36 0.02  
## cool2r -0.71 0.01  
## cool3 -0.44 0.01  
## cool4 -0.40 0.01  
## cool5r -0.95 0.02  
## cool6 -0.74 0.01  
## cool7r -0.37 0.01  
## cool8 -0.05 0.01  
## coolness -0.20 0.01  
## SCH\* -2.00 0.01  
## coffee.love -1.15 0.03  
## SEX -1.71 0.01  
## AGE 46.60 0.06  
## sex\* -1.71 0.01  
## SchoolF\* -2.00 0.01

cool.df <- with(cool,#with() adds the data set to all the following objects (its a shortcut)  
 data.frame(cool1, 6-cool2r, cool3, cool4, (6-cool5r), cool6, (6-cool7r), cool8))  
  
cool$coolness2 <- rowMeans(cool.df)  
  
head(cbind(cool$coolness,cool$coolness2))

## [,1] [,2]  
## [1,] 3.250 3.250  
## [2,] 3.250 3.250  
## [3,] 3.250 3.250  
## [4,] 3.000 3.000  
## [5,] 3.375 3.375  
## [6,] 4.250 4.250

alpha(cool.df)

##   
## Reliability analysis   
## Call: alpha(x = cool.df)  
##   
## raw\_alpha std.alpha G6(smc) average\_r S/N ase mean sd  
## 0.85 0.85 0.86 0.42 5.8 0.003 3.2 0.75  
##   
## lower alpha upper 95% confidence boundaries  
## 0.85 0.85 0.86   
##   
## Reliability if an item is dropped:  
## raw\_alpha std.alpha G6(smc) average\_r S/N alpha se  
## cool1 0.83 0.83 0.83 0.41 4.9 0.0034  
## X6...cool2r 0.84 0.84 0.84 0.43 5.2 0.0033  
## cool3 0.84 0.84 0.83 0.42 5.1 0.0033  
## cool4 0.83 0.83 0.83 0.42 5.0 0.0033  
## X.6...cool5r. 0.82 0.82 0.82 0.40 4.7 0.0036  
## cool6 0.85 0.85 0.85 0.45 5.6 0.0031  
## X.6...cool7r. 0.84 0.84 0.84 0.43 5.2 0.0033  
## cool8 0.83 0.83 0.83 0.40 4.8 0.0035  
##   
## Item statistics   
## n raw.r std.r r.cor r.drop mean sd  
## cool1 5616 0.74 0.73 0.68 0.62 3.7 1.15  
## X6...cool2r 5598 0.68 0.68 0.62 0.57 2.7 1.10  
## cool3 5610 0.68 0.69 0.64 0.57 3.6 1.01  
## cool4 5594 0.69 0.71 0.66 0.59 3.6 1.00  
## X.6...cool5r. 5602 0.78 0.77 0.75 0.69 2.8 1.16  
## cool6 5598 0.61 0.60 0.50 0.47 3.3 1.10  
## X.6...cool7r. 5602 0.68 0.67 0.62 0.56 2.5 1.06  
## cool8 5602 0.75 0.76 0.72 0.66 3.7 0.99  
##   
## Non missing response frequency for each item  
## 1 2 3 4 5 miss  
## cool1 0.05 0.15 0.12 0.43 0.26 0.02  
## X6...cool2r 0.11 0.43 0.19 0.21 0.06 0.02  
## cool3 0.02 0.14 0.25 0.41 0.17 0.02  
## cool4 0.02 0.14 0.24 0.42 0.17 0.03  
## X.6...cool5r. 0.11 0.36 0.21 0.24 0.08 0.02  
## cool6 0.06 0.21 0.23 0.39 0.12 0.02  
## X.6...cool7r. 0.14 0.49 0.16 0.17 0.04 0.02  
## cool8 0.02 0.12 0.19 0.48 0.19 0.02

max(scale(cool$AGE))

## [1] 10.20972

min(scale(cool$AGE))

## [1] -0.7300494

(cool$AGE[1234] - mean(cool$AGE))/sd(cool$AGE)

## [1] -0.02881427

scale(cool$AGE)[1234]

## [1] -0.02881427

#construct linear model  
mod1F <- lm(formula = coolness ~SchoolF, data = cool)  
  
#summary of model  
summary(mod1F)

##   
## Call:  
## lm(formula = coolness ~ SchoolF, data = cool)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.25063 -0.45595 -0.00063 0.49937 1.79405   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.25063 0.01432 226.967 <2e-16 \*\*\*  
## SchoolFprivate -0.04469 0.02026 -2.206 0.0274 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.7524 on 5516 degrees of freedom  
## (220 observations deleted due to missingness)  
## Multiple R-squared: 0.0008814, Adjusted R-squared: 0.0007003   
## F-statistic: 4.866 on 1 and 5516 DF, p-value: 0.02743

#objects  
objects(mod1F)

## [1] "assign" "call" "coefficients" "contrasts"   
## [5] "df.residual" "effects" "fitted.values" "model"   
## [9] "na.action" "qr" "rank" "residuals"   
## [13] "terms" "xlevels"

#estimate for cal students  
mod1F$coefficients[1]

## (Intercept)   
## 3.250634

#estimate for Stanford students  
mod1F$coefficients[1] + mod1F$coefficients[2]

## (Intercept)   
## 3.205946

confint(mod1F)[1,]

## 2.5 % 97.5 %   
## 3.222557 3.278711

confint(mod1F)[2,]

## 2.5 % 97.5 %   
## -0.084401636 -0.004973804

confint(mod1F)[1,] + confint(mod1F)[2,]

## 2.5 % 97.5 %   
## 3.138156 3.273737

actual <- mod1F$model$coolness  
  
predicted <- mod1F$fitted.values  
  
error <- mod1F$residuals  
  
#R^2 by hand  
SSmod <- sum(error^2) #error in the model  
SStotal <- sum((actual - mean(actual))^2) #errors in the baseline model  
(SStotal - SSmod)/SStotal

## [1] 0.0008813949

#F-test by hand  
dfmod <- length(coef(mod1F)) - 1  
dfres <- mod1F$df.residual  
F <- ((SStotal - SSmod)/dfmod)/(SSmod/dfres)  
F # same thing as model output!

## [1] 4.866063

#subset data  
C <- cool[cool$SCH == "Cal",]  
s <- cool[cool$SCH == "Stanf",]  
  
head(C)

## cool1 cool2r cool3 cool4 cool5r cool6 cool7r cool8 coolness SCH  
## 718 5 2 4 4 2 4 2 4 4.125 Cal  
## 719 4 3 3 4 4 3 3 4 3.250 Cal  
## 720 5 4 4 4 2 4 4 5 3.750 Cal  
## 721 4 4 3 3 2 2 3 3 3.000 Cal  
## 722 4 4 4 4 3 4 4 4 3.375 Cal  
## 723 4 2 5 4 1 5 2 5 4.500 Cal  
## coffee.love SEX AGE sex SchoolF coolness2  
## 718 8 2 20 2 public 4.125  
## 719 5 1 22 1 public 3.250  
## 720 9 2 21 2 public 3.750  
## 721 5 2 19 2 public 3.000  
## 722 1 1 18 1 public 3.375  
## 723 2 1 22 1 public 4.500

head(s)

## cool1 cool2r cool3 cool4 cool5r cool6 cool7r cool8 coolness SCH  
## 1 4 4 5 4 3 2 4 4 3.250 Stanf  
## 2 4 4 4 4 4 4 4 4 3.250 Stanf  
## 3 4 3 4 3 3 3 4 4 3.250 Stanf  
## 4 4 4 4 4 4 2 4 4 3.000 Stanf  
## 5 3 4 4 4 3 4 3 4 3.375 Stanf  
## 6 4 1 4 4 2 4 2 5 4.250 Stanf  
## coffee.love SEX AGE sex SchoolF coolness2  
## 1 6 2 23 2 private 3.250  
## 2 7 1 21 1 private 3.250  
## 3 1 2 22 2 private 3.250  
## 4 7 3 20 3 private 3.000  
## 5 5 2 18 2 private 3.375  
## 6 5 2 19 2 private 4.250

t.test(s$coolness,C$coolness) #same as model output

##   
## Welch Two Sample t-test  
##   
## data: s$coolness and C$coolness  
## t = -2.2059, df = 5513.4, p-value = 0.02743  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -0.08440194 -0.00497350  
## sample estimates:  
## mean of x mean of y   
## 3.205946 3.250634

#removing NA values  
mS <- mean(s$coolness,na.rm = T)  
mC <- mean(C$coolness,na.rm = T)  
  
varS <- var(s$coolness,na.rm = T)  
varC <- var(C$coolness,na.rm = T)  
  
nC <- sum(!is.na(C$coolness))  
nS <- sum(!is.na(s$coolness))  
  
dfC <- nC - 1  
dfS <- nS - 1  
  
sPool <- (dfC\*varC + dfS\*varS)/(dfC + dfS)  
sePool <- sqrt((sPool/nC)+(sPool/nS))  
meanDiff <- mS - mC  
t <- meanDiff/sePool  
t

## [1] -2.205916