Mars Attacks!

Dr. Joshua Reynolds

2024-01-03

knitr::opts\_chunk$set(echo = TRUE)

#data simulation
set.seed(1908)

id<-1:1000

age<-round(rnorm(1000, mean=35, sd=5))

sex<-rbinom(n=1000, size = 1, prob = .5)
library(dplyr)

## Warning: package 'dplyr' was built under R version 4.2.3

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
## filter, lag

## The following objects are masked from 'package:base':
##
## intersect, setdiff, setequal, union

sex<-recode(sex, "0"="Female", "1"= "Male")
sex<-factor(sex)

height <-rnorm(n = 1000, mean = 66, sd = 5)
height<- height + ifelse(sex=="Male", abs(rnorm(1000, mean=3, sd=.5)) , 0)

job<-round(runif(1000, min = 0, max=2))
job<-recode(job, "0"="Cook", "1"= "Miner", "2"="Cleaner")

iq<-round(rnorm(n = 1000, mean = 100, sd = 15))

library(faux)

## Warning: package 'faux' was built under R version 4.2.3

##
## \*\*\*\*\*\*\*\*\*\*\*\*
## Welcome to faux. For support and examples visit:
## https://debruine.github.io/faux/
## - Get and set global package options with: faux\_options()
## \*\*\*\*\*\*\*\*\*\*\*\*

brain <- rnorm\_pre(iq, mu = 1200, sd = 80, r = 0.5)

mk<-round(rnorm(n = 1000, mean = 50, sd = 5))
mk<- mk + ifelse(sex=="Male",round(rnorm(1000, mean=8, sd=1.5)) , 0)
mk<-mk + age/2 + round(rnorm(1000, mean=5, sd=2))

skin<-rnorm(n = 1000, mean = 100, sd = 5)

blood<-round(runif(1000, min = 0, max=3))

blood<-recode(blood, "0"="A", "1"= "B", "2"="AB", "3"= "O")

continent<-round(runif(1000, min = 0, max=6))

continent <-recode(continent, "0"="NorthA", "1"= "SouthA", "2"="Europe", "3"= "Asia", "4"= "Africa",
 "5"= "Australia", "6"= "Antarctica")

thalamus<-rnorm\_pre(brain, mu = 8.5, sd = .5, r = 0.4)
thalamus<-thalamus + ifelse(sex=="Male",rnorm(1000, mean=.7, sd=.3) , 0)

mars<-data.frame(id, age, sex, height, job, iq, brain, mk, skin, blood, continent, thalamus)

library(missMethods)

## Warning: package 'missMethods' was built under R version 4.2.3

mars<-delete\_MAR\_censoring(mars, 0.4, "continent", cols\_ctrl = "height")

library("writexl")
write\_xlsx(mars,"C:\\Users\\JoJo\\Desktop\\mars.xlsx")

#################################
#data prep
d<-mars
d$sex<-factor(d$sex)
d$sex<-relevel(d$sex, ref = "Female")
d$blood<-factor(d$blood)
d$job<-factor(d$job)
d$continent<-factor(d$continent)

###########Lower difficulty#######################
#1. What is the most frequent type for blood?
table(d$blood)

##
## A AB B O
## 152 304 362 182

#blood type B

#2. Calculate the average skin score per blood group.

library(dplyr)
d %>% group\_by(blood) %>% summarize(mean(skin))

## # A tibble: 4 × 2
## blood `mean(skin)`
## <fct> <dbl>
## 1 A 100.
## 2 AB 100.
## 3 B 100.
## 4 O 99.2

#3. Create a histogram of skin, but we want the density on the y-axis.
#We also want to overly a curve that represent a normal distribution.
#Create this plot and comment on skin’s distribution.

x1 <- seq(min(d$skin), max(d$skin), length = 1000)
nc <- dnorm(x1, mean = mean(d$skin), sd = sd(d$skin))
hist(d$skin, prob = TRUE,
 main="Histogram with Density for Skin Puncture Pressure", xlab="",
 ylab="Density")
lines(x1, nc, col = 2, lwd = 2)



#4. Are brain size and iq related?
cor.test(iq, brain, data=d) #correlation coefficient

##
## Pearson's product-moment correlation
##
## data: iq and brain
## t = 18.759, df = 998, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
## 0.4632458 0.5550022
## sample estimates:
## cor
## 0.5105761

#5. Are brain size and thalamus related?
cor.test(brain, thalamus, data=d) #correlation coefficient

##
## Pearson's product-moment correlation
##
## data: brain and thalamus
## t = 8.834, df = 998, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
## 0.2108300 0.3258562
## sample estimates:
## cor
## 0.2693033

#6. We need to know how many standard deviations apart are males and females (sex) on height.
library(effsize)
cohen.d(d$height~d$sex)

##
## Cohen's d
##
## d estimate: -0.6074726 (medium)
## 95 percent confidence interval:
## lower upper
## -0.7345896 -0.4803555

#they are .6 standard deviations apart

#7. Build a linear model. The outcome will be mk and the predictors will be age and sex.
#Create a table of the estimates and the squared semi-partials.

mod1<-lm(formula= mk~age + sex, data=d)
library(jtools)
summ(mod1, digits=3, part.corr = TRUE, vifs = TRUE)

## MODEL INFO:
## Observations: 1000
## Dependent Variable: mk
## Type: OLS linear regression
##
## MODEL FIT:
## F(2,997) = 418.659, p = 0.000
## R² = 0.456
## Adj. R² = 0.455
##
## Standard errors: OLS
## --------------------------------------------------------------------------------
## Est. S.E. t val. p VIF partial.r part.r
## ----------------- -------- ------- -------- ------- ------- ----------- --------
## (Intercept) 53.105 1.257 42.256 0.000
## age 0.542 0.035 15.375 0.000 1.000 0.438 0.359
## sexMale 8.582 0.351 24.481 0.000 1.000 0.613 0.572
## --------------------------------------------------------------------------------

#8. Build a linear model. The outcome will be mk and the predictors will be age, job, sex, and iq.
#Create a table of the estimates and the squared semi-partials.
mod2<-lm(formula= mk~age + job + sex + iq, data=d)
library(jtools)
summ(mod2, digits=3, part.corr = TRUE, vifs = TRUE)

## MODEL INFO:
## Observations: 1000
## Dependent Variable: mk
## Type: OLS linear regression
##
## MODEL FIT:
## F(5,994) = 168.273, p = 0.000
## R² = 0.458
## Adj. R² = 0.456
##
## Standard errors: OLS
## --------------------------------------------------------------------------------
## Est. S.E. t val. p VIF partial.r part.r
## ----------------- -------- ------- -------- ------- ------- ----------- --------
## (Intercept) 51.670 1.752 29.492 0.000
## age 0.542 0.035 15.382 0.000 1.002 0.438 0.359
## jobCook 0.815 0.510 1.599 0.110 1.009 0.051 0.037
## jobMiner 0.312 0.424 0.735 0.462 1.009 0.023 0.017
## sexMale 8.573 0.351 24.414 0.000 1.004 0.612 0.570
## iq 0.011 0.012 0.904 0.366 1.007 0.029 0.021
## --------------------------------------------------------------------------------

#9. What is the predicted military knowledge for a person who is 20 years old,
#has indicated a job preference of miner, is a female, and has an IQ of 128?
predict(mod2, data.frame(age=20, job="Miner", sex="Female", iq=128), interval="confidence", level=.95)

## fit lwr upr
## 1 64.20223 62.85736 65.54709

###########Higher difficulty#######################
#10. Compare the two linear models that you made.
#Which one makes better out of sample predictions given the number of effective parameters?
library(AICcmodavg)
modlist<- list(mod1, mod2)

mod.names <- c("mod1", "mod2")

aictab(cand.set = modlist, modnames = mod.names)

##
## Model selection based on AICc:
##
## K AICc Delta\_AICc AICcWt Cum.Wt LL
## mod1 4 6264.98 0.00 0.78 0.78 -3128.47
## mod2 7 6267.46 2.49 0.22 1.00 -3126.67

bictab(cand.set = modlist, modnames = mod.names)

##
## Model selection based on BIC:
##
## K BIC Delta\_BIC BICWt Cum.Wt LL
## mod1 4 6284.57 0.00 1 1 -3128.47
## mod2 7 6301.70 17.14 0 1 -3126.67

#11. Create a scatterplot with mk on the y-axis and age on the x-axis. Also incorporate iq, and sex.
#Facet that by job. Also add a regression line for each faceted plot. Describe what you see.
library(ggplot2)

## Warning: package 'ggplot2' was built under R version 4.2.3

s1 <- ggplot(aes(y = mk, x = age), data = d) +
 geom\_point(mapping = aes(color=sex, size = iq), alpha=.5) +
 theme\_minimal() + geom\_smooth(method = "lm") +
labs(y="Military Knowledge", x="Age (years)")

s1 <- s1 +
 theme(axis.text.x = element\_text(face="bold", size=14),
 axis.text.y = element\_text(face="bold", size=14)) +
 theme(axis.title = element\_text(size = 15))

s1 <- s1 + guides(color = guide\_legend(title = "Sex")) + guides(size = guide\_legend(title = "Intelligence"))

s1 + facet\_grid(. ~ job)

## `geom\_smooth()` using formula = 'y ~ x'



#12. Your Martian superiors are interested in if humans of different blood groups are more intelligent.
#Specifically, they want an index of the relative strength of evidence from the data about the hypotheses (i.e., the change from prior to posterior odds brought about by the data).
#The numerator would be the hypothesis that blood groups do have an effect on intelligence and the denominator would be that they don’t. Calculate that and summarize what you found.

library(BayesFactor)

## Loading required package: coda
## Loading required package: Matrix
## \*\*\*\*\*\*\*\*\*\*\*\*
## Welcome to BayesFactor 0.9.12-4.4. If you have questions, please contact Richard Morey (richarddmorey@gmail.com).
##
## Type BFManual() to open the manual.
## \*\*\*\*\*\*\*\*\*\*\*\*

anovaBF(iq ~ blood, data=d)

## Bayes factor analysis
## --------------
## [1] blood : 0.008539719 ±0.04%
##
## Against denominator:
## Intercept only
## ---
## Bayes factor type: BFlinearModel, JZS

#13. You no doubt noticed that there is some missingness for the variable, continent.
#Can you figure out the cause of the missingness? Include a plot if you can.
table(d$continent, useNA="ifany")

##
## Africa Antarctica Asia Australia Europe NorthA SouthA
## 81 56 105 97 105 46 110
## <NA>
## 400

library(ggplot2)
library(visdat)
vis\_miss(d, show\_perc = F) + coord\_flip()

## Warning: `gather\_()` was deprecated in tidyr 1.2.0.
## ℹ Please use `gather()` instead.
## ℹ The deprecated feature was likely used in the visdat package.
## Please report the issue at <]8;;https://github.com/ropensci/visdat/issueshttps://github.com/ropensci/visdat/issues]8;;>.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last\_lifecycle\_warnings()` to see where this warning was
## generated.



ggplot(d,
 aes(x = continent,
 y = height)) +
 geom\_point() +
 labs(y="Height", x="Continents")

