Q2.R

joann

2021-12-24

```
# Quiz 2
# Use the Anscombe dataset from the package carData

library(carData)
data(Anscombe)

# The research objective is to predict education based on income, young and urban

# You are free to use any lecture notes, R code examples, packages and commands.

# When done submit the R code with comments where your answers are

# Task 1: Fit the regression to the original dataset and report the estimated equation
# Your answer should look something like education = b0 + b1*income + b2*young + b3*urban
# where b0, b1, b2 and b3 are the estimated coefficients after you perform the lm function


#check data
str(Anscombe)
```

```
## 'data.frame':    51 obs. of  4 variables:
##  $ education: int  189 169 230 168 180 193 261 214 201 172 ...
##  $ income   : int  2824 3259 3072 3835 3549 4256 4151 3954 3419 3509 ...
##  $ young    : num  351 346 348 335 327 ...
##  $ urban    : int  508 564 322 846 871 774 856 889 715 753 ...
```

```
#fit regression
model1 = lm(education ~ income+ young+ urban, data = Anscombe)
summary(model1)
```

```
##
## Call:
## lm(formula = education ~ income + young + urban, data = Anscombe)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -60.240 -15.738  -1.156  15.883  51.380
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.868e+02  6.492e+01  -4.418 5.82e-05 ***
```

```
## income       8.065e-02  9.299e-03   8.674 2.56e-11 ***
## young        8.173e-01  1.598e-01   5.115 5.69e-06 ***
## urban       -1.058e-01  3.428e-02  -3.086  0.00339 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 26.69 on 47 degrees of freedom
## Multiple R-squared:  0.6896, Adjusted R-squared:  0.6698
## F-statistic: 34.81 on 3 and 47 DF,  p-value: 5.337e-12
```

```
#show intercept and coefficients of regression model
coef(model1)
```

```
##   (Intercept)        income         young         urban
## -286.83876273    0.08065325    0.81733774   -0.10580623
```

```
# Task 2. Now use ampute from mice package to create missingness wih the following commands
# which keeps education completely observed
library(mice)
```

```
## Warning: package 'mice' was built under R version 4.0.3
```

```
##
## Attaching package: 'mice'
```

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

```
## The following objects are masked from 'package:base':
##
##     cbind, rbind
```

```
set.seed(895)
ans.miss = ampute(Anscombe, prop = 0.6)$amp
ans.miss = cbind(Anscombe$education, ans.miss[, 2:4])


# Task 3. Report the percent missing from each variable in ans.miss
# Your answer should be three percentages, the averages of non-deleted income, young and urban.

#check for missing values
m=sapply(ans.miss, function(x) sum(is.na(x)))
#percentage of missingness
total = nrow(ans.miss)
(m/total)*100
```

```
## Anscombe$education            income             young             urban
##           0.000000           7.843137          21.568627           5.882353
```

```r
#missing percentage for income: 7.8%, for young: 21.57%, for urban:5.88%

#find average of non-deleted income
mean=sapply(ans.miss, function(x) mean(x,na.rm=TRUE))
mean
```

```
## Anscombe$education            income             young             urban
##          196.3137         3182.2979          359.1000          669.9375
```

```r
#the average of income is 3192.3, average of young is 359.1, and average of urban
#is 669.9



# 4. Is the resulting dataset ans.miss MCAR, MAR, or MNAR?

#the resulting dataset is MAR

# 5. Refit the regression model using lm command on the new dataset ans.miss
# Report the total change in all regression coefficients
# Your answer should be a single number obtain the following way:
# sum(abs(reg$coef - reg.miss$coef)), where reg is the original regression and reg.miss is from ans.miss

model2 = lm(Anscombe$education ~ income+ young+ urban, data = ans.miss)
#total as below:
sum(abs(model1$coef - model2$coef))
```

```
## [1] 48.31509
```

```r
# 6. What is the name of the missing data technique you applied in part 5?

#the missing data technique is regression imputation

# 7. Use ans.miss dataset and apply mean imputation to restore all variables.
# Store the restored data as ans.mean
# Report the four imputed means.
# Your answer should be four numbers.

## Mean imputation
mean.imp <- function (a){
  missing <- is.na(a)
  a.obs <- a[!missing]
  imputed <- a
  imputed[missing] <- mean(a.obs)
  return (imputed)
}

ans.mean=mean.imp(ans.miss)
mean2=sapply(ans.mean, function(x) mean(x,na.rm=TRUE))
mean2
```

```
## Anscombe$education            income             young             urban
##          196.3137         3019.6133          520.6425          695.7100
```

```r
# 8. Fit again the regression model on mean imputed data and report the total change in the regression
# using same techinque as in part 5.
# Your answer should be a single number indicating the total absolute change in coefficients.

model3 = lm(Anscombe$education ~ income+ young+ urban, data = ans.mean)
#total as below:
sum(abs(model1$coef - model3$coef))
```

```
## [1] 373.7955
```

```r
# 9. Repeat parts 7-8 but now use regression imputation based on the complete variable education
# Your answer should be a single number, the total change in coefficients
d =complete(mice(data.frame(ans.miss), method = "norm.predict", m = 1, maxit = 1))
```

```
##
##  iter imp variable
##    1   1  income  young  urban
```

```r
model4 = lm(Anscombe$education ~ income+ young+ urban, data = d)
#total as below:
sum(abs(model1$coef - model4$coef))
```

```
## [1] 179.6639
```

```r
# 10. Which method resulted in smallest change in the regression coefficients
# as compared to the full dataset?

#the regression imputation gives the samllest change
```