Coding Assignment 1

Team 11

Due: 2021-09-29 23:59

Table of Contents

# Put any packages you want here  
library(readxl)  
library(dplyr)

##   
## 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

library(gt)   
library(gtsummary)   
library(corrplot)

## corrplot 0.92 loaded

library(car)

## Loading required package: carData

##   
## Attaching package: 'car'

## The following object is masked from 'package:dplyr':  
##   
## recode

library(jtools)  
  
sessionInfo()

## R version 4.2.1 (2022-06-23)  
## Platform: x86\_64-apple-darwin17.0 (64-bit)  
## Running under: macOS Big Sur ... 10.16  
##   
## Matrix products: default  
## BLAS: /Library/Frameworks/R.framework/Versions/4.2/Resources/lib/libRblas.0.dylib  
## LAPACK: /Library/Frameworks/R.framework/Versions/4.2/Resources/lib/libRlapack.dylib  
##   
## locale:  
## [1] en\_US.UTF-8/en\_US.UTF-8/en\_US.UTF-8/C/en\_US.UTF-8/en\_US.UTF-8  
##   
## attached base packages:  
## [1] stats graphics grDevices utils datasets methods base   
##   
## other attached packages:  
## [1] jtools\_2.2.0 car\_3.1-0 carData\_3.0-5 corrplot\_0.92   
## [5] gtsummary\_1.6.1 gt\_0.7.0 dplyr\_1.0.9 readxl\_1.4.1   
##   
## loaded via a namespace (and not attached):  
## [1] Rcpp\_1.0.9 cellranger\_1.1.0 pillar\_1.8.1   
## [4] compiler\_4.2.1 tools\_4.2.1 digest\_0.6.29   
## [7] evaluate\_0.16 lifecycle\_1.0.1 tibble\_3.1.8   
## [10] gtable\_0.3.0 pkgconfig\_2.0.3 rlang\_1.0.4   
## [13] cli\_3.3.0 DBI\_1.1.3 rstudioapi\_0.14   
## [16] yaml\_2.3.5 xfun\_0.32 fastmap\_1.1.0   
## [19] stringr\_1.4.1 knitr\_1.39 generics\_0.1.3   
## [22] vctrs\_0.4.1 grid\_4.2.1 tidyselect\_1.1.2   
## [25] glue\_1.6.2 R6\_2.5.1 fansi\_1.0.3   
## [28] rmarkdown\_2.15 pander\_0.6.5 tidyr\_1.2.0   
## [31] purrr\_0.3.4 ggplot2\_3.3.6 magrittr\_2.0.3   
## [34] scales\_1.2.1 htmltools\_0.5.3 broom.helpers\_1.8.0  
## [37] abind\_1.4-5 assertthat\_0.2.1 colorspace\_2.0-3   
## [40] utf8\_1.2.2 stringi\_1.7.8 munsell\_0.5.0   
## [43] crayon\_1.5.1

A Florida health insurance company wants to predict annual claims for individual clients. The company pulls a random sample of 50 customers. The owner wishes to charge an actuarially fair premium to ensure a normal rate of return. The owner collects all of their current customer’s health care expenses from the last year and compares them with what is known about each customer’s plan.

The data on the 50 customers in the sample is as follows:

* Charges: Total medical expenses for a particular insurance plan (in dollars)
* Age: Age of the primary beneficiary
* BMI: Primary beneficiary’s body mass index (kg/m2)
* Female: Primary beneficiary’s birth sex (0 = Male, 1 = Female)
* Children: Number of children covered by health insurance plan (includes other dependents as well)
* Smoker: Indicator if primary beneficiary is a smoker (0 = non-smoker, 1 = smoker)
* Cities: Dummy variables for each city with the default being Sanford

Answer the following questions using complete sentences and attach all output, plots, etc. within this report.

**For this assignment, ignore the categorical variables (gender, smoker, cities)**

# Question 1

Perform univariate analyses on the quantitative variables (center, shape, spread). Include descriptive statistics, and histograms. Be sure to use terms discussed in class such as bimodal, skewed left, etc.

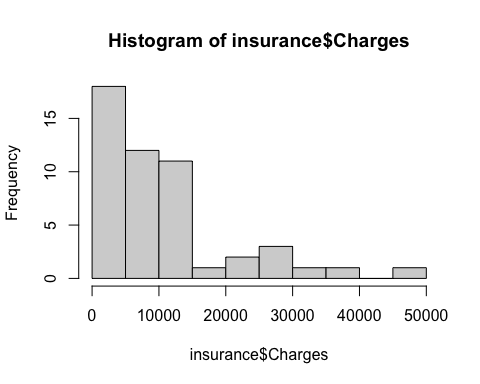
mean(insurance$Charges)

## [1] 10860.18

median(insurance$Charges)

## [1] 7814.885

hist(insurance$Charges)



var(insurance$Charges)

## [1] 97681418

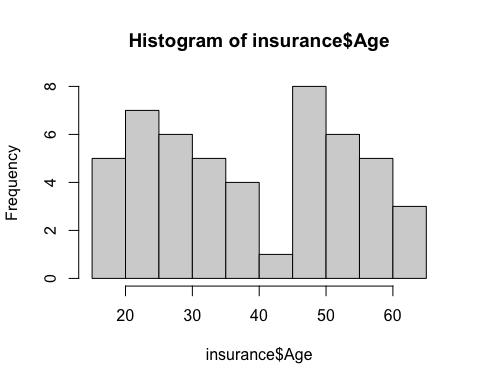
mean(insurance$Age)

## [1] 39.32

median(insurance$Age)

## [1] 38

hist(insurance$Age)



var(insurance$Age)

## [1] 205.8955

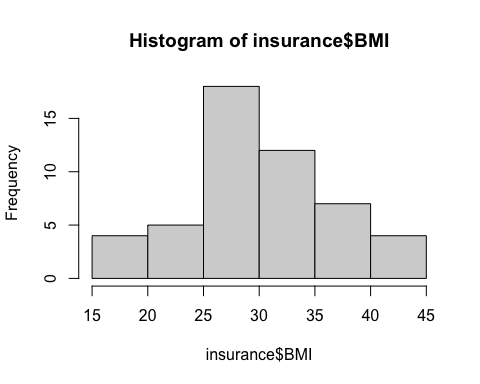
mean(insurance$BMI)

## [1] 30.0219

median(insurance$BMI)

## [1] 29.865

hist(insurance$BMI)



var(insurance$BMI)

## [1] 38.85582

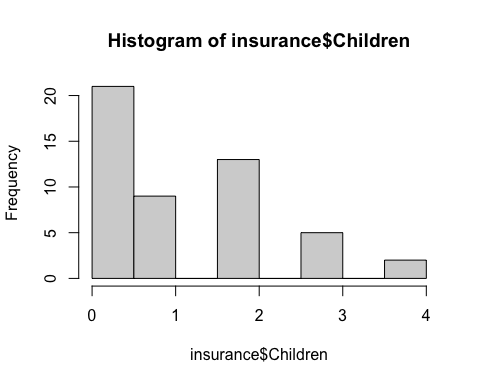
mean(insurance$Children)

## [1] 1.16

median(insurance$Children)

## [1] 1

hist(insurance$Children)



var(insurance$Children)

## [1] 1.443265

summary(insurance)

## Charges Age BMI Female Children   
## Min. : 1243 Min. :18.00 Min. :16.82 Min. :0.00 Min. :0.00   
## 1st Qu.: 4089 1st Qu.:26.00 1st Qu.:26.30 1st Qu.:0.00 1st Qu.:0.00   
## Median : 7815 Median :38.00 Median :29.86 Median :1.00 Median :1.00   
## Mean :10860 Mean :39.32 Mean :30.02 Mean :0.52 Mean :1.16   
## 3rd Qu.:12542 3rd Qu.:52.00 3rd Qu.:33.75 3rd Qu.:1.00 3rd Qu.:2.00   
## Max. :45702 Max. :64.00 Max. :43.34 Max. :1.00 Max. :4.00   
## Smoker WinterSprings WinterPark Oviedo   
## Min. :0.0 Min. :0.00 Min. :0.00 Min. :0.00   
## 1st Qu.:0.0 1st Qu.:0.00 1st Qu.:0.00 1st Qu.:0.00   
## Median :0.0 Median :0.00 Median :0.00 Median :0.00   
## Mean :0.1 Mean :0.26 Mean :0.22 Mean :0.16   
## 3rd Qu.:0.0 3rd Qu.:0.75 3rd Qu.:0.00 3rd Qu.:0.00   
## Max. :1.0 Max. :1.00 Max. :1.00 Max. :1.00

sd(insurance$Charges)

## [1] 9883.391

sd(insurance$Age)

## [1] 14.34906

sd(insurance$BMI)

## [1] 6.233444

sd(insurance$Children)

## [1] 1.20136

getModes <- function(x) {  
ux <- unique(x)  
tab <- tabulate(match(x, ux))  
ux[tab == max(tab)]  
}  
  
getModes(insurance$Charges)

## [1] 3972.925 1242.816 45702.022 12235.839 9630.397 33750.292 5846.918  
## [8] 21195.818 11286.539 2842.761 4779.602 27808.725 7050.021 10214.636  
## [15] 2789.057 12044.342 9910.360 6406.411 2867.120 10370.913 25992.821  
## [22] 8733.229 4718.204 3481.868 4438.263 8428.069 6640.545 3877.304  
## [29] 7201.701 17878.901 3378.910 38746.355 4561.189 2150.469 11454.022  
## [36] 12950.071 2534.394 26467.097 4894.753 2026.974 5138.257 14988.432  
## [43] 1719.436 11987.168 7133.903 20878.784 8988.159 2632.992 12644.589  
## [50] 14394.398

getModes(insurance$Age)

## [1] 55 18 25

getModes(insurance$BMI)

## [1] 29.830 33.345

getModes(insurance$Children)

## [1] 0

Univariate Anaysis

Age- Bimodal,Not Symmetrical, slightly skewed right BMI- Unimodal, Symmetrical, non skewed Children- Unimodal, Not Symmetrical, skewed to the right

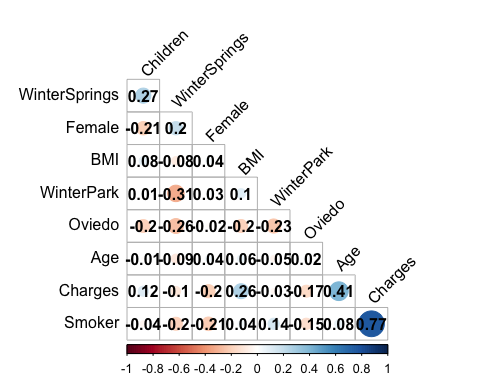
# Question 2

Perform bivariate analyses on the quantitative variables (direction, strength and form). Describe the linear association between all variables.

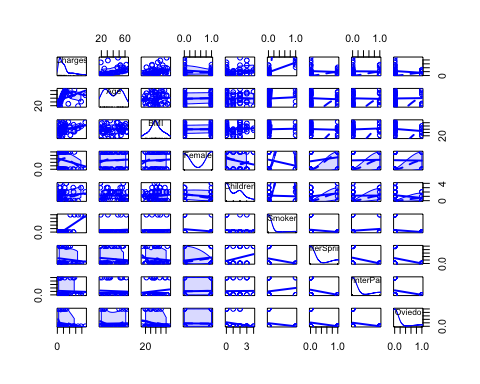
cor(insurance)

## Charges Age BMI Female Children  
## Charges 1.0000000 0.40851021 0.26010167 -0.20062613 0.119471220  
## Age 0.4085102 1.00000000 0.06023366 0.04418932 -0.010134006  
## BMI 0.2601017 0.06023366 1.00000000 0.03980372 0.081592774  
## Female -0.2006261 0.04418932 0.03980372 1.00000000 -0.207349166  
## Children 0.1194712 -0.01013401 0.08159277 -0.20734917 1.000000000  
## Smoker 0.7692945 0.07696905 0.04273356 -0.21350421 -0.044844853  
## WinterSprings -0.1022566 -0.09360057 -0.08301334 0.20443403 0.265306220  
## WinterPark -0.0302302 -0.04595276 0.10483491 0.02705871 0.009743085  
## Oviedo -0.1672654 0.02473317 -0.19604548 -0.01747141 -0.196330606  
## Smoker WinterSprings WinterPark Oviedo  
## Charges 0.76929454 -0.10225662 -0.030230204 -0.16726535  
## Age 0.07696905 -0.09360057 -0.045952757 0.02473317  
## BMI 0.04273356 -0.08301334 0.104834907 -0.19604548  
## Female -0.21350421 0.20443403 0.027058710 -0.01747141  
## Children -0.04484485 0.26530622 0.009743085 -0.19633061  
## Smoker 1.00000000 -0.19758299 0.144841365 -0.14547859  
## WinterSprings -0.19758299 1.00000000 -0.314800094 -0.25869686  
## WinterPark 0.14484136 -0.31480009 1.000000000 -0.23178450  
## Oviedo -0.14547859 -0.25869686 -0.231784498 1.00000000

corrplot(cor(insurance),  
type = "lower",  
order = "hclust",  
tl.col = "black",  
tl.srt = 45,  
addCoef.col = "black",  
diag = FALSE)



scatterplotMatrix(insurance)

 Linear Relationship Between Variables:

Age- Strong positive linear relationship BMI- Stronger, even more so than age, positive linear relationship Children- Strongest positive linear relationship amongst all variables included in analysis

All variables included in analysis have a strong positive linear relationship which makes sense as older individuals that weigh more and have more dependents are typically going to have higher healthcare cost than those individuals that are young, slim, and with no dependents.

# Question 3

Generate a regression equation in the following form:

model <- lm(Charges ~ Age + BMI + Children, data = insurance)  
  
summ(model)

## MODEL INFO:  
## Observations: 50  
## Dependent Variable: Charges  
## Type: OLS linear regression   
##   
## MODEL FIT:  
## F(3,46) = 4.67, p = 0.01  
## R² = 0.23  
## Adj. R² = 0.18   
##   
## Standard errors: OLS  
## -------------------------------------------------------  
## Est. S.E. t val. p  
## ----------------- ----------- --------- -------- ------  
## (Intercept) -11700.62 7054.72 -1.66 0.10  
## Age 272.66 89.09 3.06 0.00  
## BMI 361.03 205.75 1.75 0.09  
## Children 863.03 1065.66 0.81 0.42  
## -------------------------------------------------------

also write out the regression cleanly in this document.

Charges = -11700.62 + 272.66Age + 361.03BMI + 863.03Children

# Question 4

An eager insurance representative comes back with a potential client. The client is 40, their BMI is 30, and they have one dependent. Using the regression equation above, predict the amount of medical expenses associated with this policy. (Provide a 95% confidence interval as well)

newPrediction <- data.frame(Age = 40,  
 BMI= 30,  
 Children = 1)  
predict(model, newdata = newPrediction)

## 1   
## 10899.59

predict(model,   
 newdata = newPrediction,   
 interval = "confidence",  
 level = .95)

## fit lwr upr  
## 1 10899.59 8331.644 13467.54