Assignment-7.R

Zachary

2020-11-28

library(tidyverse)

## Warning: package 'tidyverse' was built under R version 4.0.3

## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.2 v purrr 0.3.4  
## v tibble 3.0.4 v dplyr 1.0.2  
## v tidyr 1.1.2 v stringr 1.4.0  
## v readr 1.4.0 v forcats 0.5.0

## Warning: package 'tibble' was built under R version 4.0.3

## Warning: package 'tidyr' was built under R version 4.0.3

## Warning: package 'readr' was built under R version 4.0.3

## Warning: package 'stringr' was built under R version 4.0.3

## Warning: package 'forcats' was built under R version 4.0.3

## -- Conflicts ------------------------------------------ tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

library(ggplot2)  
  
# load the csv  
smoking\_df <- read\_csv("smoking.csv")

##   
## -- Column specification --------------------------------------------------------  
## cols(  
## years = col\_double(),  
## smoker = col\_double(),  
## age = col\_double()  
## )

# explore the dataset  
glimpse(smoking\_df)

## Rows: 1,000  
## Columns: 3  
## $ years <dbl> 22.06873, 28.72932, 22.55478, 26.07060, 24.67653, 24.89933, ...  
## $ smoker <dbl> 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 1, 0, 0, 0, ...  
## $ age <dbl> 46.46951, 37.51216, 52.60310, 52.65284, 50.11282, 46.79830, ...

head(smoking\_df)

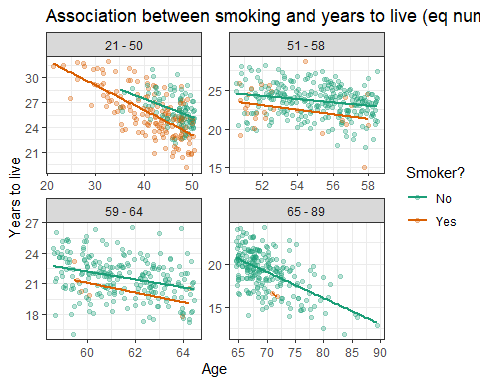
## # A tibble: 6 x 3  
## years smoker age  
## <dbl> <dbl> <dbl>  
## 1 22.1 0 46.5  
## 2 28.7 0 37.5  
## 3 22.6 1 52.6  
## 4 26.1 0 52.7  
## 5 24.7 0 50.1  
## 6 24.9 1 46.8

summary(smoking\_df)

## years smoker age   
## Min. :11.83 Min. :0.000 Min. :21.28   
## 1st Qu.:20.51 1st Qu.:0.000 1st Qu.:50.25   
## Median :22.69 Median :0.000 Median :57.61   
## Mean :22.64 Mean :0.191 Mean :56.85   
## 3rd Qu.:24.90 3rd Qu.:0.000 3rd Qu.:63.89   
## Max. :31.94 Max. :1.000 Max. :89.47

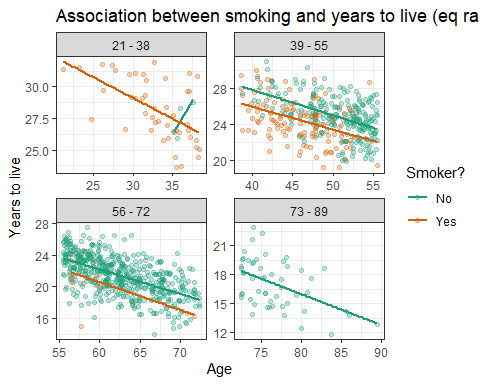
# splitting into equal numbers  
smoking\_df <- smoking\_df %>% mutate(age\_round = (round(age,0))) %>%  
 mutate(quartile = cut\_number(age\_round,4))  
  
smoking\_df$quartile <- factor(smoking\_df$quartile, labels = c("21 - 50", "51 - 58", "59 - 64", "65 - 89"))  
  
plot <- smoking\_df %>% ggplot(aes(x=age, y=years,color=as.factor(smoker))) +  
 geom\_point(alpha=0.3) +  
 geom\_smooth(method="lm", se=F) +  
 labs(title = "Association between smoking and years to live (eq numbers)") + xlab("Age") + ylab("Years to live") +  
 facet\_wrap(~quartile,scales = "free") +  
 scale\_color\_brewer(name ="Smoker?",  
 labels=c("No", "Yes"),  
 palette = "Dark2") +  
 theme\_bw()  
  
#plot the graph  
plot

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



# splitting into equal age ranges  
smoking\_df\_2 <- smoking\_df %>% mutate(age\_round = (round(age,0))) %>%  
 mutate(quartile = cut\_interval(age\_round,4))  
  
smoking\_df\_2$quartile <- factor(smoking\_df\_2$quartile, labels = c("21 - 38", "39 - 55", "56 - 72", "73 - 89"))  
  
plot2 <- smoking\_df\_2 %>% ggplot(aes(x=age, y=years,color=as.factor(smoker))) +  
 geom\_point(alpha=0.3) +  
 geom\_smooth(method="lm", se=F) +  
 labs(title = "Association between smoking and years to live (eq range)") + xlab("Age") + ylab("Years to live") +  
 facet\_wrap(~quartile,scales = "free") +  
 scale\_color\_brewer(name ="Smoker?",  
 labels=c("No", "Yes"),  
 palette = "Dark2") +  
 theme\_bw()  
  
#plot the graph  
plot2

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



#create the lm model  
lm(smoking\_df$years ~ smoking\_df$smoker)

##   
## Call:  
## lm(formula = smoking\_df$years ~ smoking\_df$smoker)  
##   
## Coefficients:  
## (Intercept) smoking\_df$smoker   
## 22.090 2.896

#create lm model, controlling for age  
lm(smoking\_df$years ~ smoking\_df$smoker + smoking\_df$age)

##   
## Call:  
## lm(formula = smoking\_df$years ~ smoking\_df$smoker + smoking\_df$age)  
##   
## Coefficients:  
## (Intercept) smoking\_df$smoker smoking\_df$age   
## 39.6069 -1.7517 -0.2925

#saving model for summary  
model <-lm(smoking\_df$years ~ smoking\_df$smoker + smoking\_df$age)  
  
#view summary of model  
summary(model)

##   
## Call:  
## lm(formula = smoking\_df$years ~ smoking\_df$smoker + smoking\_df$age)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -6.1704 -1.3111 0.0578 1.3495 6.7682   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 39.606862 0.458148 86.450 <2e-16 \*\*\*  
## smoking\_df$smoker -1.751731 0.199636 -8.775 <2e-16 \*\*\*  
## smoking\_df$age -0.292493 0.007561 -38.684 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.982 on 997 degrees of freedom  
## Multiple R-squared: 0.6469, Adjusted R-squared: 0.6462   
## F-statistic: 913.3 on 2 and 997 DF, p-value: < 2.2e-16

#construct 95% confidence interval   
std.err <- sqrt(diag(vcov(model)))  
coeff <- model$coefficients  
  
#print out confidence intervals   
for (i in 1:3){  
 cat(sprintf("95%% Confidence Interval for %s: %f \u00b1 %f \n", names(coeff[i]), coeff[i], (2\*std.err[i])))  
}

## 95% Confidence Interval for (Intercept): 39.606862 ± 0.916296   
## 95% Confidence Interval for smoking\_df$smoker: -1.751731 ± 0.399273   
## 95% Confidence Interval for smoking\_df$age: -0.292493 ± 0.015122

#99% confidence intervals  
for (i in 1:3){  
 cat(sprintf("99%% Confidence Interval for %s: %f \u00b1 %f \n", names(coeff[i]), coeff[i], (3\*std.err[i])))  
}

## 99% Confidence Interval for (Intercept): 39.606862 ± 1.374444   
## 99% Confidence Interval for smoking\_df$smoker: -1.751731 ± 0.598909   
## 99% Confidence Interval for smoking\_df$age: -0.292493 ± 0.022683

#creating sampling list  
sampling = list(20,50,100,500)  
  
#print confint for all sample sizes  
for (i in sampling){  
 new\_sample <- sample\_n(smoking\_df,i)  
 model <- lm(new\_sample$smoker ~ new\_sample$years + new\_sample$age)  
 print(confint(model))  
}

## 2.5 % 97.5 %  
## (Intercept) 2.07724089 7.914885543  
## new\_sample$years -0.14937621 -0.001428515  
## new\_sample$age -0.08288586 -0.026759197  
## 2.5 % 97.5 %  
## (Intercept) 1.50735667 4.622805039  
## new\_sample$years -0.07236753 0.007524389  
## new\_sample$age -0.04976550 -0.023325053  
## 2.5 % 97.5 %  
## (Intercept) 1.82757522 4.025485061  
## new\_sample$years -0.05909308 -0.001599784  
## new\_sample$age -0.04507090 -0.026866837  
## 2.5 % 97.5 %  
## (Intercept) 2.24698875 3.22007234  
## new\_sample$years -0.04959490 -0.02425701  
## new\_sample$age -0.03414887 -0.02613678