02\_Assignment\_Brown.RMD

Binta Brown

9/05/2019

## Clear Environment  
rm(list=ls())  
## Get tidyverse library  
library(tidyverse)

## ── Attaching packages ─────────────────────────────────────── tidyverse 1.2.1 ──

## ✔ ggplot2 3.2.1 ✔ purrr 0.3.2  
## ✔ tibble 2.1.3 ✔ dplyr 0.8.3  
## ✔ tidyr 0.8.3 ✔ stringr 1.4.0  
## ✔ readr 1.3.1 ✔ forcats 0.4.0

## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()

## Load in the data  
load("/Users/BBrown/Desktop/8200/LLO8200-master/county\_educ.Rdata")

load("/Users/BBrown/Desktop/8200/LLO8200-master/county\_to\_zip.Rdata")

load("pd.Rdata")

load("/Users/BBrown/Desktop/8200/LLO8200-master/pd.Rdata")

## Full explanation of data in codebook  
load("pd\_lab\_explain.Rdata")  
  
#or use View  
View(lab\_explain)

##Question 1 Calculate the mean of the outcome.

I piped the the data loaded to the summarize function to view the unconditional avergae of the homeownership rate variable, homewon\_rate, which I found using the the mean fucntion

pd%>%summarize(mean\_homeown\_rate=mean(homeown\_rate,na.rm=TRUE))

## # A tibble: 1 x 1  
## mean\_homeown\_rate  
## <dbl>  
## 1 72.7

## Question 2

Use your mean as a prediction. Create a new variable that consists of the mean of the outcome

To create a new variable, I piped the data to the mutate fucntion and added mean\_homeown\_rate as a new column in my original dataset pd

pd<-pd%>%mutate(mean\_homeown\_rate=mean(homeown\_rate,na.rm=TRUE))

## Question 3

Calculate a summary measure of the errors for each observation—the difference between your prediction and the outcome.

Every observation in my dataset now has the same value of 72.74%, I can now calculate the error for each observation, that is, how far each county’s homewonership rate is from the mean, that is, 72.74%. So now, I subtracted the mean\_homeown\_rate from each individual county’s homeown\_rate.

pd<-pd%>%mutate(error1=(homeown\_rate-mean\_homeown\_rate))

To see far how far off this unconditional prediction is, I calculated the root mean sqaured error (RMSE).

## RMSE  
  
rmse\_uncond\_mean<-rmse(pd$homeown\_rate,pd$mean\_homeown\_rate)  
  
rmse\_uncond\_mean

## [1] 7.653637

This rmse indicates that the unconditional mean is off by 7.65%

## Question 4

Calculate the mean of the outcome at levels of a predictor variable.

I randonly selected house\_unit\_multi into four levels using the ntile function and added it as a variable using mutate

pd<-pd%>%mutate(house\_unit\_multi\_level=ntile(house\_unit\_multi,4))  
table(pd$house\_unit\_multi\_level)

##   
## 1 2 3 4   
## 772 772 772 772

From here, I grouped data by house\_unit\_multi\_level, then calculated the pred\_homewon\_rate for each level and then ranked them in descending order.

pd<-pd%>%group\_by(house\_unit\_multi\_level)%>% ## Group by predictor  
 ##Calculate mean at each level of predictor  
 mutate(pred\_homeown\_rate=mean(homeown\_rate))%>%  
 ## Ungroup  
 ungroup()%>%   
 #Rank by prediction, with ties sorted randomly  
 mutate(pred\_homeown\_rate\_rank=rank(-pred\_homeown\_rate,ties.method="random"))

## Question 5

Use these conditional means as a prediction: for every county, use the conditional mean to provide a ‘’best guess” as to that county’s level of the outcome.

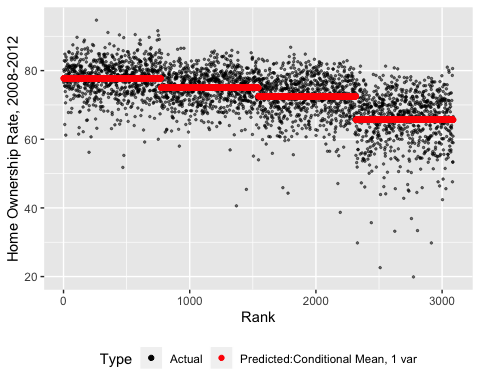
I view the counties and sort them by pred\_homeown\_rate

pd%>%select(county,house\_unit\_multi,house\_unit\_multi\_level,pred\_homeown\_rate)%>%View()

Those who live in multi-level homes, 2 or more, have a less likelihood of owning homes.

To visualize this as a plot, I used the ggplot function.

gg<-ggplot(data=pd,aes(x=pred\_homeown\_rate\_rank,y=homeown\_rate,color="Actual"))  
gg<-gg+geom\_point(alpha=.5,size=.5)  
gg<-gg+geom\_point(aes(x=pred\_homeown\_rate\_rank,y=pred\_homeown\_rate,color="Predicted:Conditional Mean, 1 var"))  
gg<-gg+ scale\_color\_manual("Type",values=c("Predicted:Conditional Mean, 1 var"="red","Actual"="black"))  
gg<-gg+theme(legend.position="bottom")  
gg<-gg+xlab("Rank")+ylab("Home Ownership Rate, 2008-2012")  
gg



# Question 6

Calculate a summary measure of the error in your predictions.

rmse\_cond\_mean<-rmse(pd$pred\_homeown\_rate,pd$mean\_homeown\_rate)  
rmse\_uncond\_mean

## [1] 7.653637

## R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

summary(cars)

## speed dist   
## Min. : 4.0 Min. : 2.00   
## 1st Qu.:12.0 1st Qu.: 26.00   
## Median :15.0 Median : 36.00   
## Mean :15.4 Mean : 42.98   
## 3rd Qu.:19.0 3rd Qu.: 56.00   
## Max. :25.0 Max. :120.00

## Including Plots

You can also embed plots, for example:



Note that the echo = FALSE parameter was added to the code chunk to prevent printing of the R code that generated the plot.