###########################################

# Project: Week 12 Lab

# Purpose: Cleaning + linear Regression in R

# Author: Aileen Gonzalez

# Edit date: Nov 20, 2023

# Data: Michigan State of the State Fall 2019 Survey

###########################################

# CLEAN DATA -------------------------------------------------------

# set working directory so that R knows which folder your file is in

# read in dataset and save in object 'data'

data<- read.csv('Lab 12 Logistic Data.csv', header=TRUE)

summary(data)

# install 'car' package to get ready to recode variables

# USE THIS COMMAND FOR THE FIRST TIME:

install.packages('car') # run this only once

# then load the package using library()

library(car)

# use recode() leader.bio to change '1' and '2' to Male and Female

# first have to "create" a new variable in data (gender), just by adding it on 'data'

data$gender<- recode(data$CD1, "1='Male'; 2='Female'")

# because 'Female' and 'Male' are characters, have to also put them in single quotes

# if they are numbers (1, 2, 3), don't have to

# use table() to get counts

table(data$CD1)

table(data$gender)

# check the data type of CD1, new "gender" variable, and age using class()

class(data$CD1)

class(data$gender)

class(data$age) # Age is being read as a numeric

#Why did we have you run summary statistics for age but not for gender?

#Gender is a character type while age is a numeric type. As a result, running summary(data$age) gives us the summary statistics for age but does not make sense or work for gender.

summary(data$age) #Run summary statistics for age

# Write a sentence or two describing the characteristics of the MI 18+ population in terms of age and gender.

#The MI 18+ population in this survey are ~53.4% (499/935) Male and ~46.6% (436/935) female while the average age is 57 years with a minimum of 18 and a maximum of 94.

# Define a new variable in “data” called “labor”, and in it store a recoded version of the dependent variable (laborforce), where 1s are still 1s, and the 2s are now 0s

data$labor<- recode(data$laborforce, "1=1; 2=0")

# Check the class of labor.

class(data$labor)

# Change the data type of labor to be a factor variable using as.factor()

data$labor<- as.factor(data$labor)

#Use table() twice to check that your new labor variable has the same counts as the original laborforce variable

table(data$labor.f)

table(data$laborforce)

# Use levels() to check that the zero category is set as the dummy category. The first category listed will be the dummy category.

levels(data$labor)

#Is zero listed first in your levels check? Which category does “0” represent?

#Zero is listed first in the levels check. The category "0" is the dummy category in labor representing "not in labor force".

# glm() for binomial logistic regression (dichotomous dependent variable)

glm <- glm(data$labor ~ data$age + data$gender, family = binomial("logit"))

summary(glm)

# From the signs on the coefficients, what can you tell about how gender and age affect the odds of being in the labor force? What is the statistical significance of the independent variables?

#Intercept, age, and gender all have very small p-values indicating that they are statistically significant. For age, the coefficient is -0.08406 meaning that for each increase in age, the odds of being in the laborforce decrease by 0.08406. For gender, male is associated with an increase in odds of being in the laborforce by 0.80658 (its coefficient) compared to female.

# Install and load the package ‘odds.n.ends'

install.packages('odds.n.ends') # only have to install once

library(odds.n.ends)

# Use odds.n.ends() to find the odds ratios of model

odds.n.ends(glm)

#Is the model statistically significant overall, and what does this tell us? What is the effect of gender and age on odds of being in the labor force? Be specific about the magnitude of the effect (how much do the odds change given changes in gender and age?

# The model statistically is significant overall which tells us at least one of the predictors contributes significantly to predicting the odds of one being in the labor force. Age has a negative effect on the odds of being in the labor force while being male has a positive effect on being in the labor force. The odds ratio for age is 0.9194 from 0.9081 to 0.9301 with 95% CI implying that for each one unit increase in age, odds of being in the labor force decreases by ~8.06%. On the other hand, the odds ratio for gender male is 2.2402 from 1.6222 to 3.1069 with 95% CI implying that being male is about 2.24x greater odds of being in the labor force compared to gender female.