install.packages("alr4")

install.packages("car")

library(car)

library(alr4)

library(fbasics)

library(tidyverse)

library(caret)

library(leaps)

library(olsrr)

load("/Users/arpitadeb/Library/CloudStorage/OneDrive-DrexelUniversity/Biostatistics Books/569-linear-regression contents/FINALS/CHR2022PA.RData")

data(CHR2022PA)

attach(CHR2022PA)

head(CHR2022PA)

str(CHR2022PA)

#summary of the dataset

summary(CHR2022PA)

summary(X..Severe.Housing.Problems)

#standard deviation calculation

sd(X..Severe.Housing.Problems)

sd(Average.Daily.PM2.5)

sd(X..Drive.Alone.to.Work)

sd(X..Long.Commute...Drives.Alone)

#Descriptive statistics of demographic measures: mean & standard deviation

mean(CHR2022PA$Population)

sd(CHR2022PA$Population)

mean(CHR2022PA$X..Less.Than.18.Years.of.Age)

sd(CHR2022PA$X..Less.Than.18.Years.of.Age)

mean(CHR2022PA$X..65.and.Over)

sd(CHR2022PA$X..65.and.Over)

mean(CHR2022PA$X..Black)

sd(CHR2022PA$X..Black)

mean(CHR2022PA$X..American.Indian...Alaska.Native)

sd(CHR2022PA$X..American.Indian...Alaska.Native)

mean(CHR2022PA$X..Asian)

sd(CHR2022PA$X..Asian)

mean(CHR2022PA$X..Native.Hawaiian.Other.Pacific.Islander)

sd(CHR2022PA$X..Native.Hawaiian.Other.Pacific.Islander)

mean(CHR2022PA$X..Hispanic)

sd(CHR2022PA$X..Hispanic)

mean(CHR2022PA$X..Non.Hispanic.white)

sd(CHR2022PA$X..Non.Hispanic.white)

mean(CHR2022PA$X..Not.Proficient.in.English)

sd(CHR2022PA$X..Not.Proficient.in.English)

mean(CHR2022PA$X..female)

sd(CHR2022PA$X..female)

mean(CHR2022PA$X..rural)

sd(CHR2022PA$X..rural)

mean(CHR2022PA$Years.of.Potential.Life.Lost.Rate)

sd(CHR2022PA$Years.of.Potential.Life.Lost.Rate)

summary(X..rural)

hist(CHR2022PA$X..Drive.Alone.to.Work)

describe(CHR2022PA$ Presence.of.Water.Violation)

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

#Sub-Section 1.2: Simple Linear Regression

m2 <- lm(Years.of.Potential.Life.Lost.Rate~ Average.Daily.PM2.5, data = CHR2022PA)

summary(m2)

m2 <- lm(Years.of.Potential.Life.Lost.Rate~ X..Severe.Housing.Problems, data = CHR2022PA)

summary(m2)

m2 <- lm(Years.of.Potential.Life.Lost.Rate~ X..Drive.Alone.to.Work, data = CHR2022PA)

summary(m2)

m2 <- lm(Years.of.Potential.Life.Lost.Rate~ X..Long.Commute...Drives.Alone, data = CHR2022PA)

summary(m2)

m2 <- lm(Years.of.Potential.Life.Lost.Rate~ Presence.of.Water.Violation, data = CHR2022PA)

summary(m2)

m2 <- lm(Years.of.Potential.Life.Lost.Rate~ Population, data = CHR2022PA)

summary(m2)

m2 <- lm(Years.of.Potential.Life.Lost.Rate~ X..Less.Than.18.Years.of.Age, data = CHR2022PA)

summary(m2)

m2 <- lm(Years.of.Potential.Life.Lost.Rate~ X..65.and.Over, data = CHR2022PA)

summary(m2)

m2 <- lm(Years.of.Potential.Life.Lost.Rate~ X..Black, data = CHR2022PA)

summary(m2)

m2 <- lm(Years.of.Potential.Life.Lost.Rate~ X..American.Indian...Alaska.Native, data = CHR2022PA)

summary(m2)

m2 <- lm(Years.of.Potential.Life.Lost.Rate~ X..Asian, data = CHR2022PA)

summary(m2)

m2 <- lm(Years.of.Potential.Life.Lost.Rate~ X..Native.Hawaiian.Other.Pacific.Islander, data = CHR2022PA)

summary(m2)

m2 <- lm(Years.of.Potential.Life.Lost.Rate~ X..Hispanic, data = CHR2022PA)

summary(m2)

m2 <- lm(Years.of.Potential.Life.Lost.Rate~ X..Non.Hispanic.white, data = CHR2022PA)

summary(m2)

m2 <- lm(Years.of.Potential.Life.Lost.Rate~ X..Not.Proficient.in.English, data = CHR2022PA)

summary(m2)

m2 <- lm(Years.of.Potential.Life.Lost.Rate~ X..female, data = CHR2022PA)

summary(m2)

m2 <- lm(Years.of.Potential.Life.Lost.Rate~ X..rural, data = CHR2022PA)

summary(m2)

##Sub-Section 1.3: Multiple linear regression of physical environment

m5 <- lm(Years.of.Potential.Life.Lost.Rate~ Average.Daily.PM2.5 + X..Severe.Housing.Problems + X..Drive.Alone.to.Work + X..Long.Commute...Drives.Alone + Presence.of.Water.Violation, data = CHR2022PA)

summary(m5)

##Sub-Section 1.3: Multiple linear regression of Demographic measures

m2 <- lm(Years.of.Potential.Life.Lost.Rate~ Population + X..Less.Than.18.Years.of.Age + X..65.and.Over + X..Black + X..American.Indian...Alaska.Native + X..Asian + X..Native.Hawaiian.Other.Pacific.Islander + X..Hispanic + X..Non.Hispanic.white + X..Not.Proficient.in.English + X..female + X..rural, data = CHR2022PA)

summary(m2)

#correlation matrix

cor(CHR2022PA[, c("Years.of.Potential.Life.Lost.Rate", "Population", "X..Less.Than.18.Years.of.Age", "X..65.and.Over", "X..Black", "X..American.Indian...Alaska.Native", "X..Asian", "X..Native.Hawaiian.Other.Pacific.Islander", "X..Hispanic", "X..Non.Hispanic.white", "X..Not.Proficient.in.English", "X..female", "X..rural")])

cor(CHR2022PA[, c("Years.of.Potential.Life.Lost.Rate", "Average.Daily.PM2.5", "X..Severe.Housing.Problems", "X..Drive.Alone.to.Work", "X..Long.Commute...Drives.Alone", "Presence.of.Water.Violation")])

#scatterplot to check relation between response variable with final predictors

scatterplotMatrix(~ Years.of.Potential.Life.Lost.Rate + X..Long.Commute...Drives.Alone + X..Black + X..Asian + X..Hispanic + X..Non.Hispanic.white + X..Not.Proficient.in.English, data = CHR2022PA)

smooth=FALSE, diagonal=FALSE)

scatterplotMatrix(~ Years.of.Potential.Life.Lost.Rate + Population + X..Less.Than.18.Years.of.Age + X..65.and.Over + X..Black + X..American.Indian...Alaska.Native + X..Asian + X..Native.Hawaiian.Other.Pacific.Islander + X..Hispanic + X..Non.Hispanic.white + X..Not.Proficient.in.English + X..female + X..rural, data = CHR2022PA)

smooth=FALSE, diagonal=FALSE)

##forward stepwise regression of physical environment was selected for answer.

m5 <- lm(Years.of.Potential.Life.Lost.Rate~ Average.Daily.PM2.5 + X..Severe.Housing.Problems + X..Drive.Alone.to.Work + X..Long.Commute...Drives.Alone + Presence.of.Water.Violation, data = CHR2022PA)

ols\_step\_forward\_aic(m5)

ols\_step\_forward\_p(m5, penter = 0.05)

#forward stepwise regression of demographic measures

m2 <- lm(Years.of.Potential.Life.Lost.Rate~ Population + X..Less.Than.18.Years.of.Age + X..65.and.Over + X..Black + X..American.Indian...Alaska.Native + X..Asian + X..Native.Hawaiian.Other.Pacific.Islander + X..Hispanic + X..Non.Hispanic.white + X..Not.Proficient.in.English + X..female + X..rural, data = CHR2022PA)

summary(m2)

ols\_step\_forward\_aic(m2)

ols\_step\_forward\_p(m2, penter = 0.05)

#TO find interaction between Years.of.Potential.Life.Lost.Rate ~ X..Long.Commute...Drives.Alone + variables X..Black + X..Asian + X..Hispanic + X..Non.Hispanic.white + X..Not.Proficient.in.English

#final model

m0 <- lm(Years.of.Potential.Life.Lost.Rate~ X..Long.Commute...Drives.Alone + X..Less.Than.18.Years.of.Age, data = CHR2022PA)

summary(m0)

plot(m0)

#final interaction model based on forward stepwise

m\_int <- lm(Years.of.Potential.Life.Lost.Rate~ X..Long.Commute...Drives.Alone + X..Less.Than.18.Years.of.Age + X..Long.Commute...Drives.Alone \* X..Less.Than.18.Years.of.Age, data = CHR2022PA)

summary(m\_int)

plot(m\_int)

#Anova test of the models

m0 <- lm(Years.of.Potential.Life.Lost.Rate~ 1, data = CHR2022PA)

m6 <- update(m0, ~ X..Long.Commute...Drives.Alone, data = CHR2022PA)

m7 <- update(m6, ~. + X..Less.Than.18.Years.of.Age,data = CHR2022PA)

m8 <- update(m7, ~. + X..Long.Commute...Drives.Alone + X..Less.Than.18.Years.of.Age, data = CHR2022PA)

m9 <- update(m8, ~. + X..Long.Commute...Drives.Alone : X..Less.Than.18.Years.of.Age, data = CHR2022PA)

anova(m9,m8)

Anova(m9,m8)

Anova(m8,m7)

Anova(m7,m6)

Anova(m6,m0)

#transformation independently with X..Less.Than.18.Years.of.Age

with(CHR2022PA, invTranPlot(X..Less.Than.18.Years.of.Age, Years.of.Potential.Life.Lost.Rate))

#transformation independently with X..Long.Commute...Drives.Alone

with(CHR2022PA, invTranPlot(X..Long.Commute...Drives.Alone, Years.of.Potential.Life.Lost.Rate))

#residual analysis using residual plot of final model

residualPlots(m0)

#nonconstant variance of final model

ncvTest(m0)

#outlier deterction of interaction model

outlierTest(m0)

#normality assumption q-q plot

qqPlot(m0)