install.packages("openxlsx")

install.packages("psych")

install.packages("ggplot2")

install.packages("ztable")

install.packages("reshape")

library("psych")

library("openxlsx")

library('ggplot2') # plotting data

library("ztable")

library("reshape")

ps1=read.csv("ps1.csv")

#Q1:

ps1.agelimit=ps1[ps1$age>24 & ps1$age<67 & !is.na(ps1$schooling),]

ps1.agelimit$emp=as.numeric(ps1.agelimit$lf\_char<=3)

#Q2:

options(scipen = 3, digits = 4)

ps1.agelimit$age.sq=(ps1.agelimit$age)^2

ps1.agelimit$married=as.numeric(ps1.agelimit$marital==1)

ps1.agelimit$divorced=as.numeric(ps1.agelimit$marital==2)

ps1.agelimit$widowed=as.numeric(ps1.agelimit$marital==3)

ps1.agelimit$single=as.numeric(ps1.agelimit$marital==4)

ps1.agelimit$seperated=as.numeric(ps1.agelimit$marital==5)

lpm = lm(emp ~ age+age.sq+schooling+married+divorced+seperated+widowed,

data = ps1.agelimit)

summary(lpm)

#Q3:

ps1.agelimit$lpm=lpm$fitted.values

ps1.agelimit$lpm.resid=lpm$residuals

ggplot(data = ps1.agelimit,

aes(x=ps1.agelimit$lpm,y=ps1.agelimit$lpm.resid))+

geom\_point(aes(colour = factor(emp)))+

scale\_colour\_manual(values = c("firebrick","darkgreen"),

labels = c("Unemployed","Employed"),

name = "Employment status")+

theme(legend.position = "top")+

geom\_vline(xintercept = 0:1,

linetype = "dotted",

color = "royalblue",

size = 1.5)+

ggtitle("LPM expected probabilities vs. residauls")+

xlab("Expected probability")+

ylab("Residuals")

#Q4:

logit=glm(emp ~

age + age.sq + schooling + married + divorced + seperated + widowed,

ps1.agelimit,

family = "binomial"(link = "logit"))

probit=glm(emp ~

age + age.sq + schooling + married + divorced + seperated + widowed,

ps1.agelimit,

family = "binomial"(link = "probit"))

emp.coeffs=data.frame(LPM=lpm$coefficients,

Logit=logit$coefficients,

Probit=probit$coefficients)

ztable(head(emp.coeffs,9),

zebra = 1,

digits = 4,

size = 7)

#Q5:

ps1.agelimit$emp.logit=logit$fitted.values

ps1.agelimit$emp.probit=probit$fitted.values

emp.age=aggregate(ps1.agelimit[c("emp","lpm","emp.logit","emp.probit")],

by = list(age = ps1.agelimit$age),

FUN = mean)

colnames(emp.age)=c("Age","Actual Employment","LPM","Logit","Probit")

emp.age.melted=melt(emp.age,id="Age")

ggplot(data = emp.age.melted,

aes(x=Age, y=value,color=variable))+

geom\_line()+

ggtitle("Employment probability by age")+

theme(legend.position = "top")+

labs(color="Model:")+

scale\_y\_continuous(name = "Employment probability",

breaks = seq(0,1,by=0.05))+

scale\_x\_continuous(name = "Age",

breaks = seq(25,66, by=2))

#Q6:

emp.30.data=data.frame("Schooling" = 0:59)

lpm.xb.no.schooling = (emp.coeffs[1,1] + 30\*emp.coeffs[2,1] + 900\*emp.coeffs[3,1] + emp.coeffs[5,1])

logit.xb.no.schooling = (emp.coeffs[1,2] + 30\*emp.coeffs[2,2] + 900\*emp.coeffs[3,2] + emp.coeffs[5,2])

probit.xb.no.schooling = (emp.coeffs[1,3] + 30\*emp.coeffs[2,3] + 900\*emp.coeffs[3,3] + emp.coeffs[5,3])

emp.30.data$xb.lpm = (lpm.xb.no.schooling + emp.30.data$Schooling\*emp.coeffs[4,1])

emp.30.data$xb.logit = (logit.xb.no.schooling + emp.30.data$Schooling\*emp.coeffs[4,2])

emp.30.data$xb.probit = (probit.xb.no.schooling + emp.30.data$Schooling\*emp.coeffs[4,3])

emp.30.data$prob.logit = (exp(emp.30.data$xb.logit)/(1+exp(emp.30.data$xb.logit)))

emp.30.data$prob.probit = pnorm(emp.30.data$xb.probit,mean = 0, sd = 1)

emp.30 = emp.30.data[,c(1,2,5,6)]

emp.30.melted = melt(emp.30,id="Schooling")

ggplot(data = emp.30.melted, aes(x=Schooling,y=value,color=variable))+

geom\_line()

#Q7: