#Q1

hprice1 <- wpull('hprice1')

hprice\_model <- lm(price ~ sqrft + I(bdrms^2) + lotsize + I(sqrft^2) + colonial + assess + assess:lotsize, data = hprice1)

summary(hprice\_model)

AIC(hprice\_model)

BIC(hprice\_model)

# Equation: price = (1.511\*e+02) - (1.046\*e-1)\*sqrft - (9.394\*e-1)\*bdrms^2 - (2.771\*e-3)\*lotsize + (2.427\*e-5)\*sqrft^2 + (2.006\*e+1)colonial + (6.692\*e-1)\*assess + (1.721\*e-3)\*lotsize:assess

# AIC = 910.5686, BIC = 932.8646

# There are possibilities for other models but this model had the lowest AIC and BIC

#Q2

gpa2 <- wpull('gpa2')

model\_gpa <- lm(colgpa ~ sat + I(sat^2) + tothrs + athlete + verbmath + I(hsize^2) + hsize + hsrank + I(hsrank^2) + hsperc + female + black + sat:hsrank, data = gpa2)

AIC(model\_gpa)

BIC(model\_gpa)

#The euqation is of the form colgpa = 2.844 - (1.677\*e^-3)\*sat + (1.566\*e^-6)\*sat^2 + (1.707\*e^-3)\*tothrs + (1.336\*e^-1)\*athlete - (8.645\*e^-2)\*verbmath + (2.846\*e^-4)\*hsize^2 + (2.260\*e^-2)\*hsize - (5.627\*e^-4)\*hsrank + (5.196\*e^-6)\*hsrank^2 - (7.993\*e^-3)\*hsperc + (1.526\*e^-1)\*female - (3.416\*e^-1)black - (2.755\*e^-6)\*sat\*hsrank

#AIC = 6675.001, BIC = 6769.917

#Q3

mlb1 <- wpull('mlb1')

model\_mlb1 <- lm(log(salary) ~ teamsal + nl + years + games + atbats + runs + hits + doubles + triples + hruns + rbis + bavg + bb + fldperc + frstbase + scndbase + shrtstop + whitepop + gamesyr + hrunsyr + allstar + hispph, data = mlb1)

AIC(model\_mlb1)

BIC(model\_mlb1)

#The equation can't be stated as there are too many factors to list

#R-Squared = 67.53%, AIC = 712.7723, BIC = 803.9505

#Out of many possibilities, one of the models had the lowest AIC and BIC as stated above

#Q4

rental <- wpull('rental')

#i

pctstu = (rental$enroll/rental$pop)\*100

pctstu

y90 = ifelse(rental$year == '90',1,0)

y90

model\_rental1 <- plm(log(rent) ~ y90 + log(pop) + log(avginc) + pctstu, model = "pooling", data = rental)

summary(model\_rental1)

#The rents were higher in 1990. The dummy variable 'y90' indicates that there is a 26.22% increase in rent for a 10 year period holding the other factors constant

#The co-efficient of 'pctstu' states that there would be 0.5% increase in rent for every 1-point increase in 'pctstu'

#ii

#The standard errors in part 1 are not valid because we did not use differencing

#iii

#The OLS pooled equation we get after first differencing is:

#log(rent) = -0.5688064 + 0.2622267\*y90 + 0.0406863\*log(pop) + 0.5714460\*log(avginc) + 0.0050436\*pctstu

#The co-efficient of pctstu = 0.0050436, This implies that 1-point increase in pctstu increases the rent by 0.5%

#The t-value = 2.401e-6 which is statistically significant, hence, we conclude that the relative size of student population impacts the rent.

#iv

#The estimated fixed effects model's equation is

#log(rent) = 0.3855 + 0.0722\*log(pop) + 0.3099\*log(avginc) + 0.0112\*pctstu

#S.e. 0.0368245 0.0883426 0.0664771 0.0041319

#Hence, we have verified that the estimates, co-efficients and standard errors are similar for both the first differencing model and fixed effects model

#Q5

murder <- wpull ('murder')

#i

pmurder <- pdata.frame(murder,index = c('state','year'))

pmurder

# For the given model, the sign of b1 should be negative if the convicted murders have a deterring effect

# It is not possible to determine the sign of b2

#ii

murder90 <- subset(murder[year == 90|year == 93])

murder90

modelmurder123 <- plm((mrdrte) ~ exec + unem, model = "pooling", data = murder90)

summary(modelmurder123)

#Estimated pooled OLS equation by ignoring serial correlation is:

# mrdrte = -4.88906 + 0.11491(exec) + 2.28750(unem)

#Co-efficient of exec is B1 = 0.11491, co-efficient of 'unem' is b2 = 2.2875

#Since the co-efficient of 'unem' is positive, there is no deterring effect

#iii

pmurder111 <- pdata.frame(murder90,index = c('state','year'))

modelmurder456 <- plm(mrdrte ~ exec + unem, model = "fd", data = pmurder111)

summary(modelmurder456)

#Estimated the equation by using first differences is:mrdrte = 0.413267 - 0.103840\*exec - 0.066591\*unem

#We can see the deterring effect but it's not strong enough to affect the model

#iv

coeftest(modelmurder456,vcovHC)

#The Hetero-Skedasticity robust- standard errors for the pooled OLS model are:

#Intercept exec unem

#0.413267 -0.103840 -0.066591

#v

murder\_1993 <- subset(murder[year == 93])

murder\_1993[order(exec)]

#The state with highest number of executions in 1993 is Texas with exec = 34

#The state with the second highest is Virginia with 11 executions

#vi

murdernotTX <- subset(murder90[state != 'TX'])

murdernotTX

pmurder20 <- pdata.frame(murdernotTX,index = c('state','year'))

pmurder20

modelmurder4 <- plm(mrdrte ~ exec + unem, model = "fd", data = pmurder20)

summary(modelmurder4)

coeftest(modelmurder4,vcovHC)

#Estimated the equation using first differences by dropping Texas from the model and taking years 1990 and 1993

#mrdrte = 0.412523 + -0.067471\*exec + -0.070032\*unem

#0.211283 0.104913 0.160371

#The heteroskedasticity-robust standard errors are

#mrdrte = 0.412523 + -0.067471\*exec + -0.070032\*unem

#0.194331 0.076690 0.141755

#The effect is nearly twice as big, but it is insignificant

#vii

modelmurder444 <- plm(mrdrte ~ as.factor(year) + exec + unem, model = "within", data = pmurder)

summary(modelmurder444)

#Estimated the model by fixed effects using all 3 years of data and including Texas in the model

#mrdrte = 1.55621\*year90 + 1.73324\*year93 - 0.13832\*exec + 0.22132\*unem

#The co-efficient of 'exec' when all 3 years are included is -0.138. But the co-efficient when

#only 2 years is included is 0.1038 based on the fixed effects model.

#The deterring effect is statistically significant in a 3-year model, but it is statistically insignificant in case of a 2-year model

#Q6

airfare <- wpull('airfare')

pairfare <- pdata.frame(airfare,index=c('id','year'))

#i

year97 = ifelse(airfare$year == 1997,1,0)

year97

year98 = ifelse(airfare$year == 1998,1,0)

year99 = ifelse(airfare$year == 1999,1,0)

year00 = ifelse(airfare$year == 2000,1,0)

year00

concen <- airfare$bmktshr

concen

model\_airfarea<- plm(log(fare) ~ year97 + year98 + year99 + year00 + bmktshr + ldist + I(ldist^2), model = "pooling", data= pairfare)

summary(model\_airfarea)

#For a 0.1 increase in 'concen' there would be a 3.52% increase in fare

#ii

ldistsq <- log(airfare$dist^2)

ldistsq

model\_airfare2 <- lm(log(fare) ~ year97 + year98 + year99 + year00 + bmktshr + log(dist) + ldistsq, data = airfare)

summary(model\_airfare2)

confint(model\_airfare2)

model\_airfare3 <- lm(log(fare) ~ year97 + year98 + year99 + year00 + bmktshr + log(dist) + ldistsq, data = pairfare)

summary(model\_airfare3)

confint(model\_airfare3)

coeftest(model\_airfare2,vcov. = vcovHC )

confint1 <- 0.31497 + (1.96\*0.0323)

confint1

confint2 <- 0.31497 - (1.96\*0.0323)

confint2

#The robust 95% confidence interval = [0.251662,0.378278] whereas, the normal confidence interval = [0.26462,0.38316]. This shows that both models have almost the same confindence interval with robust interval being slightly bigger

#iii

#At 79.50 miles, the relationship between 'log(fare)' and 'dist' become positive

#iv

model\_airfare4 <- plm(log(fare) ~ bmktshr + log(dist) + ldistsq, model = "within",effect = 'twoways', data = pairfare)

summary(model\_airfare4)

model\_airfare5 <- plm(log(fare) ~ year97 + year98 + year99 + year00 + bmktshr + log(dist)+ ldistsq, model = "within", data= pairfare)

summary(model\_airfare5)

#The fixed effects model is:

# log(fare) = 0.0173 + 0.00273\*log(dist) + 0.00068\*log(dist^2)

#The co-efficient b1 = 0.017. This indicates that a 0.1% increase in concentration leads to a 1.7% increase in fare

#v

#Average passengers per day and air traffic might be the other 2 characteristics that might affect and highly correlate with concentration

#vi

#Based on the above model, we can say that as concentration increases, the fares aslo increase

#We can also say that there is a correlation between concentration and characteristics like route, distance etc

#The co-efficient of b1 from the fixed affects model is a strong estimate to support the statement

#Q7

loanapp <- wpull('loanapp')

#i

model\_loanapp <- glm(approve ~ white,family = binomial,data = loanapp)

summary(model\_loanapp)

fitted.values(model\_loanapp)

mean(fitted.values(model\_loanapp))

model\_loanapp2 <- lm(approve ~ white,data = loanapp)

summary(model\_loanapp2)

fitted.values(model\_loanapp2)

mean(fitted.values(model\_loanapp2))

#The estimated equation is of the form approve = 0.8847 + 1.4094\*white

#The estimated probabilities for whites = 0.9083879 and non-whites = 0.7077922

#The estimated probabilities are same as the linear probability estimates

#ii

model\_loanapp3 <- glm(approve ~ white + hrat + obrat + loanprc + unem + male + married + dep + sch + cosign + chist + pubrec + mortlat1 + mortlat2 + vr,family=binomial(link=probit),data = loanapp)

summary(model\_loanapp3)

#Adding all the given variables gives the co-efficient of white as 0.520254

#t-value of white = 5.371 which is statistically significant

#But we can still find evidence of discrimination against the non-whites

#Q8

#i

rowalcohol <- nrow(alcohol)

rowalcohol

alcoholempl <- subset(alcohol[status==3])

alcoholempl

rowalcoholempl <- nrow(alcoholempl)

rowalcoholempl

alcoholabuse <- subset(alcohol[abuse==1])

alcoholabuse

rowalcoholabuse <- nrow(alcoholabuse)

rowalcoholabuse

fractionempl <- (rowalcoholempl/rowalcohol)\*100

fractionempl

fractionabuse <- (rowalcoholabuse/rowalcohol)\*100

fractionabuse

#Fraction of people who are employed at time of interview is 89.82%

#Fraction of people who had alcohol abuse at time of interview is 9.92%

#ii

model\_alcohol<- lm(employ ~ abuse, data = alcohol)

model\_alcohol

summary(model\_alcohol)

coeftest(model\_alcohol,vcov.=vcovHC)

margins(model\_alcohol)

#The estimated OLS model is model\_alcohol<- lm(employ ~ abuse, data = alcohol)

#Equation is: employ = 0.9010 - (0.0283) \* abuse, standard error= 0.010206

#Normal Standard errors are:

# intercept abuse

#0.003214 0.010206

#Heteroscedasticity standard errors are:

# intercept abuse

#0.0031755 0.0111529

#There is very little difference between the normal standard errors and robust standard errors. Yes, it is statistically significant

#iii

model\_alcohol2 <- glm(employ ~ abuse, family = binomial, data = alcohol)

summary(model\_alcohol2)

margins(model\_alcohol2)

#The estimated logit model equation is employ= 2.20832 - 0.28337\*abuse,

#Standard error= 0.010206. There is no change in sign, and this is statistically significant

#The margins of linear model = -0.0283 and logit model is -0.02589. There is no much difference in average marginal effects

#iv

fitted.values(model\_alcohol)

fitted.values(model\_alcohol2)

#Fitted values are the same for both LPM model and logit model

#For abuse = 1, value = 0.8726899 and for abuse = 0, value = 0.9009946

#v

model\_alcohol3 <- lm(employ ~ abuse + age + I(age^2) + educ + I(educ^2) + married + famsize + white + northeast + midwest + south + centcity + outercity + qrt1 + qrt2 + qrt3, data = alcohol)

summary(model\_alcohol3)

margins(model\_alcohol3)

#After adding all parameters and estimating the equation, the co-efficient of abuse = -0.02025 making it statistically insignificant at the 5% significance level

#vi

model\_alcohol4 <- glm(employ ~ abuse + age + I(age^2) + educ + I(educ^2) + married + famsize + white + northeast + midwest + south + centcity + outercity + qrt1 + qrt2 + qrt3, family = binomial, data = alcohol)

summary(model\_alcohol4)

margins(model\_alcohol4)

#The co-efficient of abuse = -0.2295500, t-value = -2.145

#The average marginal effect of abuse for given logit model is -0.01938

#The estimated effect of logit model is not identical to that of linear model but it is very close and is significant at the 5% significance level

#vii

model\_alcohol5 <- glm(employ ~ abuse + exhealth + vghealth + goodhealth, family=binomial, data=alcohol)

margins(model\_alcohol5)

#When we estimate the model which only has factors indicating health conditions, the estimated effect = -0.01746 which is greater than that of previous model = -0.02025. So, adding these factors which are statistically significant make the model better

#viii

model\_alcohol6 <- lm(abuse ~ age + I(age^2) + educ + I(educ^2) + married + famsize + white + northeast + midwest + south + centcity + outercity + qrt1 + qrt2 + qrt3 + mothalc + fathalc, data = alcohol)

summary(model\_alcohol6)

#According to the respective model, the abuse variable is influenced by many factors. So it's endogenous

#The t-value of mothalc = 2.733, p-value = 0.00630 and t-value of fathinc = 5.707 and p-value almost = 0

#They are statistically significant even at 1% level. So, we can say that the factors 'mothalc' and 'fathalc' are influencing abuse

#Q9

fertil1 <- wpull('fertil1')

#i

year74 = ifelse(fertil1$year == 74,1,0)

year76 = ifelse(fertil1$year == 76,1,0)

year78 = ifelse(fertil1$year == 78,1,0)

year80 = ifelse(fertil1$year == 80,1,0)

year82 = ifelse(fertil1$year == 82,1,0)

year84 = ifelse(fertil1$year == 84,1,0)

model\_fertil <- glm(kids ~ educ + age + I(age^2) + black + east + northcen + west + farm + othrural + town + smcity + year74 + year76 + year78 + year80 + year82 + year84,family = poisson,data = fertil1)

summary(model\_fertil)

#The estimated OLS model is of the equation kids = -3.0604622 - 0.0482027\*educ + 0.2044553\*age - 0.0022290\*agesq + 0.3603475\*black + 0.0878001\*east + 0.1417221\*northcen + 0.0795427\*west - 0.0148484\*farm - 0.0572939\*othrural + 0.0306807\*town + 0.0741129\*smcity + 0.0932809\*y74 - 0.0287888\*y76 - 0.0156856\*y78 - 0.0196524\*y80 - 0.1926076\*y82 - 0.2143735\*y84

#The co-efficient of year82 = -0.1926076. This implies that the fertility rate decreases by 19.26% from 1972 to 1982

#ii

estdifference <- exp(0.3603475)-1

estdifference

#Black women had on average 36% more children than non-black women

#iii

fitted.values(model\_fertil)

kids\_estimate <- model\_fertil%>%predict(type="response")

kids\_estimate

correlation\_kids\_estimate <- cor(fertil1$kids,kids\_estimate)

correlation\_kids\_estimate

rsquared1 <- correlation\_kids\_estimate^2

rsquared1

model\_fertil2<- lm(kids ~ educ + age + I(age^2) + black + east + northcen + west + farm + othrural + town + smcity + year74 + year76 + year78 + year80 + year82 + year84,data = fertil1)

summary(model\_fertil2)

#The fitted values of Poisson model computed is:

#The correlation between kids and kids estimate = 0.34769 or 34.769%

#R-Squared = square of correlation => R-squared = (0.34769)^2

#R-Squared = 0.12089 and R-Squared of linear model = 0.1295

#Hence, the R-Squared for linear model is greater than that of the Poisson distribution model