Question 1

Using the hprice1 table, find the best model for the housing price that you can using the AIC and BIC.

Soln: I will choose model6 with AIC 915.0514 and BIC 937.3474 (considering AIC)
Model 2 if choosing BIC
$AIC
[1] 921.5689 917.9835 918.4286 920.4277 915.0514
$BIC
[1] 938.9103 927.8929 930.8153 935.2917 937.3474
R-code:

```
hpprice <- wpull('hprice1')
model1 <- lm(price~assess+bdrms+lotsize+sqrft+colonial,data=hpprice)
model2 <- lm(price~assess+bdrms,data=hpprice)
model3 <- lm(price~assess+bdrms+lotsize,data=hpprice)
model4 <- lm(price~assess+bdrms+lotsize+sqrft,data=hpprice)
model6 <- lm(price~assess+bdrms+I(bdrms^2)+lotsize+sqrft+I(sqrft^2)+colonial,data=hpprice)
c(AIC(model1,model2,model3,model4,model6),BIC(model1,model2,model3,model4,model6))
```

Question 2

Using the gpa2 table, find the best model for the college GPA that you can using the AIC and BIC.

Soln: I would choose model 3 with AIC 6763.834 considering AIC
Considering BIC I would choose model2 with BIC 6915.278
$AIC
[1] 6868.659 6870.984 6763.834 6846.387
$BIC
[1] 6919.281 6915.278 6827.111 6903.336
R-code:

```
model1 <- lm(colgpa~hsize+I(hsize^2)+hsperc+sat+female+athlete,data=gpa2)
summary(model1)
model2 <- lm(colgpa~hsize+hsperc+sat+female+athlete,data=gpa2)
model3 <- lm(colgpa~hsize+I(hsize^2)+hsperc+I(hsperc^2)+sat+I(sat^2)+female+athlete,data=gpa2)
model4 <- lm(colgpa~hsize+I(hsize^2)+hsperc+I(sat^2)+sat+female+athlete,data=gpa2)
c(AIC(model1,model2,model3,model4),BIC(model1,model2,model3,model4))
```

Question 3

Using the mlb1 table, find the best model that you can for salary using the AIC and BIC.

Soln: Model log(salary) ~ teamsal + nl + years + games + atbats + runs + hits + doubles + triples + hruns + rbis + bavg + so + sbases + fldperc + frstbase + scndbase + shrtstop + thrdbase + outfield + yrsallst + hispan + black + whitepop + blackpop + hisppop + pcinc + gamesyr + hrunsyr + atbatsyr + allstar + slugavg + rbisyr + sbasesyr + runsyr + percwhte + percblck + perchisp + blckpb + hispph + whtepw + blckph + hisppb, Has a AIC 210.18

Model log(salary) ~ teamsal + nl + years + games + atbats + runs + hits + doubles + triples + hruns + rbis + bavg + so + sbases + fldperc + frstbase + scndbase + shrtstop + thrdbase + yrsallst + hispan + black + whitepop + blackpop + hisppop + pcinc + gamesyr + allstar + slugavg + rbisyr + sbasesyr + percblck + perchisp + blckpb + hispph + whtepw + blckph + hisppb has lower BIC with -65

R-code:

```
model1 <- lm( log(salary) ~ teamsal+nl+years +games +atbats+runs +hits +doubles +
triples +hruns +rbis +bavg +bb+so+sbases+fldperc +frstbase +scndbase +shrtstop +
thrdbase +outfield +catcher +yrsallst +hispan+black +whitepop +blackpop +hisppop +pcinc
+gamesyr +hrunsyr +atbatsyr +allstar + slugavg +rbisyr+sbasesyr +runsyr+
percwhte +percblck +perchisp +blckpb+ hispph+whtepw+blckph+hisppb, data = mlb1)
model1_AIC <- step(model1)
model1_BIC <- step(model1,k=log(nrow(mlb1)))
```

Question 4

Use the rental table for this exercise. The data on rental prices and other variables for college towns are for the years 1980 and 1990. The idea is to see whether a stronger presence of students affects rental rates. The unobserved effects model is

$$\ln(rentit) = \alpha i + \beta 0 y90\ t + \beta 1\ \ln(popit) + \beta 2\ \ln(avgincit) + \beta 3 pctstuit + eit$$

where pop is city population, avginc is average income, and pctstu is student population as a percentage of city population (during the school year).

1)Estimate the equation by pooled OLS and report the results in standard form. What do you make of the estimate on the 1990 dummy variable? What do you get for $\beta\hat{}\ 3$?

Soln: ln(rentit) = -0.569 + 0.262 y90+ 0.0407 ln(popit)+0.571 ln(avgincit)+0.00504pctstuit+eit
N=128, r-squared=0.861
The dummy variable y90 is having a coefficient of 0.262 which means the rent has growth rate of 26.2% over 10 years and with p-value 8.78e-12 which is statistically significant at 0.1% level.
β3 which is the coefficient of student population as percentage is 0.00504 with p-value 2.4e06 which is statistically significant at 0.1% level. With every 1% increase in pctstu results in 0.5% increase in rent when other factors are remained fixed.

R-code:

```
rent$pctstu <- (rent$enroll/rent$pop)*100
model1 <- plm(log(rent)~as.factor(year)+log(pop)+log(avginc)+pctstu,model="pooling",data=rent)
summary(model1)
tidy(model1)
```

2)Are the standard errors you report in part 1 valid? Explain.

Soln: The explanatory variable ai is not included in pooled OLS estimation, which means ai will be in error term. This makes error term across the two time periods for each city to exhibit positive correlation. Hence the standard errors are not valid in part 1.

3) Now, difference the equation and estimate by OLS. Compare your estimate of β3 with that from part 1. Does the relative size of the student population appear to affect rental prices?

Soln: ln(rentit) = 0.3855+0.0722 ln(popit)+0.3099 log(avgincit)+0.0112pctstuit +eit

The coefficient of pctstu is 0.0112 with p-value 0.0087 <0.01 hence statistically significant at 1% level. Here when there is an 1% increase in pctstu results in 1.12% increase in rent whereas in part 1 we could see an increase in 0.504%. Hence the relative size of student population affect rental price.

R-code:

```
model2  <-  plm(log(rent)~log(pop)+log(avginc)+(pctstu),model='fd',data=rent)
summary(model2)
```

4)Estimate the model by fixed effects to verify that you get identical estimates and standard errors to those in part 3.

Soln: ln(rentit) = 0.3855y90+0.07224ln(popit)+0.309log(avginc)+0.0112pctstuit +eit

The coefficients in part 3 and from the above equation are same for popit, avginc and pctstu are same expect the intercept term in the part 3 is the coefficient of y90 in part 4 and the standard errors are same for both part 3 and part 4.

R-code:

```
model3 <-
plm(log(rent)~factor(year)+log(pop)+log(avginc)+(pctstu),model="within",effect="individual",data=rent)
summary(model3)
```

Question 5

Use the state-level data on murder rates and executions in murder for the following exercise.

1)Consider the unobserved effects model

$$mrdrteit = \alpha i + \theta t + \beta 1 execit + \beta 2 unemit + eit$$

where θt simply denotes different year intercepts and αi is the unobserved state effect. If past executions of convicted murderers have a deterrent effect, what should be the sign of β1? What sign do you think β2 should have? Explain.

Soln: β1 should be negative as the fear of murders punishment will follow they are executed post conviction. This would reduce the murder rate.

β2 will be +ve as if there is increase in unemployment the murder rate would increase.

2. Using just the years 1990 and 1993, estimate the equation from part 1 by pooled OLS. Ignore the serial correlation problem in the composite errors. Do you find any evidence for a deterrent effect?

Soln: Part 1 equation for years 1990 and 1993 is

mrdrteit = -4.88906+ 0.1149execit + 2.287unemit + eit where n= 102

the exec is having a positive coefficient indicating no deterrent effect

R-code:

```
murder <- murder %>% filter(year==93 | year ==90)
murder <- pdata.frame(murder,index=c('state','year'))
model1 <- plm(mrdrte~exec+unem,model="pooling",data=murder)
summary(model1)
```

3. Now, using 1990 and 1993, estimate the equation by fixed effects. You may use first differencing since you are only using two years of data. Is there evidence of a deterrent effect? How strong?

Soln: the equation using first differencing we get

Mrdrte= 0.413 – 0.1038exec – 0.06659 unem+eit

The coefficient of exec is -0.1038 with -ve sign with p-value 0.0207 and significant at 5% level. This explains the deterrent effect.

R-code:

```
model2 <- plm(mrdrte~exec+unem,model="fd",data=murder)
summary(model2)
```

4. Compute the heteroskedasticity-robust standard error for the estimation in part 2

Soln:

The heteroskedasticity-robust standard error for part 2 is as below:

term	estimate	std.error	statistic	p.value
(Intercept)	-4.89	6.23	-0.785	0.433
exec	0.115	0.17	0.675	0.499
unem	2.29	1.32	1.74	0.0826

R-code:

```
tidyw(model1)
```

5. Find the state that has the largest number for the execution variable in 1993. (The variable exec is total executions in 1991, 1992, and 1993.) How much bigger is this value than the next highest value?

Soln: Texas has largest number of execution in 1993 with 34 and the next highest value is Virginia VA with 11 in 1993.

R-code:

```
tidyw(model1)
max(murder$exec)
murder$exec
```

6.Estimate the equation using first differencing, dropping Texas from the analysis. Compute the usual and heteroskedasticity-robust standard errors. Now, what do you find? What is going on?

Soln: After dropping Texas we get Mrdrte = 0.413-0.0675exec-0.07 unem + eit n= 100

heteroskedasticity-robust standard errors

term	estimate	std.error	statistic	p.value
(Intercept)	0.413	0.194	2.12	0.0338
exec	-0.0675	0.0767	-0.88	0.379
unem	-0.07	0.142	-0.494	0.621

Regular OLS standard errors

term	estimate	std.error	statistic	p.value
(Intercept)	0.413	0.211	1.95	0.0569
exec	-0.0675	0.105	-0.643	0.523
unem	-0.07	0.16	-0.437	0.664

The heteroskedasticity-robust standard errors are lower than the regular OLS standard errors

R-code:

```
murder <- murder %>% filter(state!='TX')
model3 <- plm(mrdrte~exec+unem,model="fd",data=murder)
tidyw(model3)
tidy(model3)
```

7.Use all three years of data and estimate the model by fixed effects. Include Texas in the analysis. Discuss the size and statistical significance of the deterrent effect compared with only using 1990 and 1993.

Soln: the model with all three years data is model1 and model 2 is with only 1990 and 1993

Mode1 = Mrdrte = 1.556y90+1.73y93-0.138exec+0.22132 unem + eit

Model2 = Mrdrte = 0.413y93-0.1038exec-0.06659 unem + eit

The coefficient of exec is -0.138 in 3 period effect where as it is -0.1038 in two period effect, the p-value is 0.436 in three period effect which is insignificant even at 10%level.

The p-value in two period effect is 0.0207 which is significant at 5% level

R-code:

```
murder1 <- wpull('murder')
murder1 <- pdata.frame(murder1,index=c('state','year'))
model4 <- plm(mrdrte~year+exec+unem,model="within",data=murder1)
model5 <- plm(mrdrte~year+exec+unem,model="within",data=murder)
summary(model4)
summary(model5)
```

Question 6

Use the data in airfare for this exercise. We are interested in estimating the model

$\ln(fareit) = \alpha i + \theta t + \beta 1 bmktshrit + \beta 2 \ln(distit) + \beta 3[\ln(distit)]2 + eit$

where θt means that we allow for different year intercepts.

1.Estimate the above equation by pooled OLS with the year dummies included. If Δbmktshr = 0.10, what is the estimated percentage increase in fare?

Soln: **$\ln(fareit) = \alpha i + \theta t + \beta 1 bmktshrit + \beta 2 \ln(distit) + \beta 3[\ln(distit)]2 + eit$**

ln(fareit)= 6.209+0.0211y98+0.03784y99+0.0998y20+0.3601 bmktshrit-0.9016ldist+0.103ldist^2+eit

when Δbmktshr = 0.10, estimate percentage increase in fare is 0.3601*0.1*100=3.601%

R-code:

```
fare <- pdata.frame(fare,index=c('id','year'))
model1 <- plm(log(fare)~year+bmktshr+ldist+I(ldist^2),model='pooling',data=fare)
summary(model1)
tidyw(model1)
```

2. What is the usual OLS 95% confidence interval for β1? Why is it probably not reliable? If you have access to a statistical package that computes fully robust standard errors, find the fully robust 95% CI for β1. Compare it to the usual CI and comment.

Soln:

Lower limit of β1 = 0.3601-1.96*0.030069=0.3011

Upper Limit of β1 = 0.3601+1.96*0.030069=0.41905524

This is not reliable because the standard errors are not heteroskedasticity robust
The fully robust standard errors are listed as below:
ln(fareit)= 6.21+0.0211y98+0.0378y99+0.0999y20+0.36 bmktshrit-0.902ldist+0.103ldist^2+eit
Lower limit of β1 = 0.36-1.96*0.058537=0.2452
Upper Limit of β1 = 0.36+1.96*0.058537 =0.4747
The fully robust 95% CI is larger than the usual CI
This means the coefficient is statistically not different from zero when robust standard error is used than the usual OLS standard error.

3. Describe what is happening with the quadratic in ln(dist). In particular, for what value of dist does the relationship between ln(fare) and dist become positive? [Hint: First figure out the turning point value for ln(dist), and then exponentiate.] Is the turning point outside the range of the data?
Soln: The percentage change in log(fare) w.r.t dist initially decreases and then increases after a given value of dist
Change (ln(fare))/change(ln(dist)) = β2 + 2 β3(ln(dist)) = 0.901+2*0.103*ln(dist)
➔ln(dist) = 0.901/2*0.103 = 4.3737 ➔ dist =79.33
The value of dist is less than the min value of dist in dataset which is 95
R-code:
min(fare$dist)

4. Now estimate the equation using fixed effects. What is the FE estimate of β1?
Soln: ln(fare)=0.0228y98+0.036y99+0.097y20+0.168+eit
The FE estimate of β1 is 0.168.
R-code:
model2 <- plm(log(fare)~year+bmktshr+ldist+I(ldist^2),model='within',data=fare)
summary(model2)

5. Name two characteristics of a route (other than distance between stops) that are captured by αi. Might these be correlated with bmktshrit?
Soln: The weight of luggage and number of passengers and their weights, the air traffic during the Travel and the working condition fuel can be captured by αi.

6. Are you convinced that higher concentration on a route increases airfares? What is your best estimate?
Soln: Since the bmktshrit is positive related to fare which is the biggest carrier has the characteristics of route. Best estimate would be 0.168 with fixed effects.

Question 7
Use the data in loanapp for this exercise; see also question 3 in problem set 3.

1. Estimate a logit model of approve on white. Find the estimated probability of loan approval for both whites and nonwhites. How do these compare with the linear probability estimates?
Soln: From the equation in PS3 with logit model the equation for approval rate for the race will be as below:

Approve= 0.8847+1.4094 white+e , n=1989
When white =1(Whites) the approve= 0.8847+1.4094*1=0.8847+1.4094=2.2941
When white=0 (nonwhites) the approve = 0.8847+1.4094*0=0.8847
Linear Probability estimates ➔ approve = 0.70779 + 0.20060 white +e
When white = 1 (white) approve = 0.70779+0.2006*1=0.70779+0.2006=0.908
When white=0 (non whites) approve = 0.70779+0.2006*0=0.70779
With logit model we can see the higher approval rate for whites with approx. 200% and for non-whites its 88%. With Linear probability estimates the approval rate for whites are 90% and non-whites are 70% which is lower than the logit model estimates.
R-code:

```
model1 <- glm(approve~white,data=loanapp,family=binomial())
summary(model1)
model <- lm(approve~white,data=loanapp)
summary(model)
```

2. Now, add the variables hrat, obrat, loanprc, unem, male, married, dep, sch, cosign, chist, pubrec, mortlat1, mortlat2, and vr to the logit model. Is there statistically significant evidence of discrimination against nonwhites?
Soln: 3.80+0.938 white+0.0133hrat-0.0530obrat-1.9loanprc-0.0666unem-0.0664male+0.503 married-0.0907 dep+0.0412sch+0.132cosign+1.07chist-1.34pubrec-0.310mortlat1-0.895mortlat2-0.35vr+e
The co-efficient of white is 0.938 which has positive impact on approval rate with p-value 5.84e-8 which is statistically significant at 0.1% level hence showing the evidence for discrimination against non-whites.
R-code:

```
model2 <- glm(approve~white + hrat + obrat + loanprc + unem + male + married + dep + sch + cosign + chist + pubrec + mortlat1+mortlat2+vr,
        data=loanapp,family=binomial())
tidy(model2)
```

Question 8
Use the data set in alcohol, obtained from Terza (2002), to answer this question. The data, on 9 822 men, includes labor market information, whether the man abuses alcohol, and demographic and background variables. In this question you will study the effects of alcohol abuse on employ, which is a binary variable equal to one if the man has a job. If employ = 0 the man is either unemployed or not in the workforce.
1. What fraction of the sample is employed at the time of the interview? What fraction of the sample has abused alcohol?
Soln: Approximately 89% of sample are employed at the time of the interview and approx 9.9 % of sample are abused alcohol
R-code:

```
sum(alco$employ==1)
8822/9822
sum(alco$abuse==1)
974/9822
```

2. Run the simple regression of employ on abuse and report the results in the usual form, obtaining the heteroskedasticity-robust standard errors. Interpret the estimated equation. Is the relationship as you expected? Is it statistically significant?

Soln: employ= b0+b1 abuse ➔ employ = 0.901-0.0283 abuse+e , n= 1989, R-squared= 0.0007826
When there is no abuse the employ is 0.901 if there is an abuse then chance of employment is reduced by 0.028 units.
The coefficient of abuse is -ve which implies with increase in abuse there is a decrease chance of getting employment. The p-value is 0.0056 which is statistically significant at 1% level.

Below is the heteroskedasticity-robust standard errors:

term	estimate	std.error	statistic	p.value
(Intercept)	0.901	0.00318	284	0
abuse	-0.0283	0.0112	-2.54	0.0112

3. Run a glm-logit of employ on abuse. Do you get the same sign and statistical significance as in part 2? How does the average marginal effect for the logit compare with that for the linear probability model?

Soln: employ=2.20832-0.28337 abuse+e, n= 9822. We can see the same -ve sign as part 2 on abuse with p-value 0.0057 which is statistically significant at 1% level this is same as in part 2.
Since the variable is binary in nature the avg marginal effect will be similar to that of linear probability model.
R-code:

```
model2 <- glm(employ~abuse,data=alco,family=binomial())
summary(model2)
```

4. Obtain the fitted values for the LPM estimated in part 2 and report what they are when abuse = 0 and when abuse = 1. How do these compare to the logit fitted values, and why?

Soln: The fitted values are found using the function fitted(), the fitted value are 0.901 when abuse =1 and 0.873 when abuse=0. since these are binary variables there will be the same fitted values for both types of models.
R-code:

```
fitted(model1)
fitted(model2)
```

5. To the LPM in part 2 add the variables age, age2 , educ, educ2 , married, famsize, white, northeast, midwest, south, centcity, outercity, qrt1, qrt2, and qrt3. What happens to the coefficient on abuse and its statistical significance?

Soln: the part 2 equation becomes employ= 0.179-0.0202abuse+0.0159age-0.00023age2+0.0369educ-0.0087educ2 + 0.057married +0.002985famsize+ 0.098white+ 0.0166northeast+ 0.0047midwest+ 0.015south-0.015centcity+0.014outercity-0.0187qrt1-0.0067qrt2-0.0015qrt3+e
The coefficient on abuse is still -ve but the p-value is 0.409 which is significant at 5% level but not at 1% level, which was earlier.
R-code:

```
model3 <- lm(employ~abuse+age+I(age^2)+educ+I(educ^2)
```

+married+famsize+white+northeast+midwest+south+centcity+outercity+qrt1+qrt2+qrt3,data=alco)
summary(model3)

6. Estimate a glm-logit model using the variables in part 5. Find the marginal effect of abuse and its t-statistic. Is the estimated effect now identical to that for the linear model? Is it "close"?
Soln: The t-statistics of abuse is -0.22955/0.10699 = 2.14 the p-value here is 0.0319 which is significant at 5% level which is close to the results in as part 5

R-code:
model4 <- glm(employ~abuse+age +I(age^2)+educ+I(educ^2)+married+ famsize+white
+northeast+midwest+south+ centcity+outercity+qrt1+qrt2+qrt3,data=alco,family=binomial())
summary(model4)

7. Variables indicating the overall health of each man are also included in the data set. Is it obvious that such variables should be included as controls? Explain.
Soln: There is no indication for other health indicators because some of the health problem might be because of the alcohol abuse. This might be a reason for underestimating the effect on employment if we use as fixed variables.

8. Why might abuse be properly thought of as endogenous in the employ equation? Do you think the variables mothalc and fathalc, indicating whether a man's mother or father were alcoholics, are sensible instrumental variables for abuse?
Soln: The variable abuse might be endogenous in employ equation because it might have unobserved correlation with other factors like mothalc and fathalc. From the below equation we can see that the variables mothalc is having p-value 0.0063 significant at 1%level and fathalc with p-value approx. 0 significant at 0.1%level, hence making them significant controlling variables for abuse.
R-code:
model5 <- lm(abuse~age+I(age^2)+educ+I(educ^2)+married+famsize+white+northeast+midwest
+south+centcity+outercity+qrt1+qrt2+qrt3+mothalc+fathalc,data=alco)
summary(model5)

Question 9
Refer to the data in fertil1 to estimate a linear model for kids, the number of children ever born to a woman.
1. Estimate a Poisson regression model for kids, using educ, age, age2 , black, east, northcen, west, f arm, othrural, town, smcity, y74 , y76 , y78 , y80 , y82 , and y84 . Interpret the coefficient on y82 .
Soln: -3.0604626+ 0.0932809 y74 -0.0287888 y76 -0.0156856 y78 -0.0196524 y80
-0.1926076y82 -0.2143735y84-0.0482027 educ+ 0.2044553 age-0.0022290 I(age^2)
0.3603475 black +0.0878001 east+0.1417221 northcen
+0.0795427 west -0.0148484 farm -0.0572939 othrural +0.0306807 town+0.0741129 smcity
The coeffiecient on y82 means a decline in fertility rate by 19.2% in 1982 vs 1972 and is significant at 1% level with p-value 0.0043.
R-code:
model1 <- glm(kids~as.factor(year)+educ+age+I(age^2)+black+east+

```
        northcen+west+farm+othrural+town+smcity,data=fert,family=poisson())
summary(model1)
```

2. What is the estimated percentage difference in fertility between a black woman and a nonblack woman, holding other factors fixed?

Soln: By keeping other factors fixed and since being poisson regression model the black women fertility rate would be 143.3% with non-black women to be 100% hence difference is 43.3%

```
exp(0.360348)
```

3. Compute the fitted values from the Poisson regression and obtain the R-squared as the squared correlation between kidsi and kids di .. Compare this with the R-squared for the linear regression model.

Soln: The R-squared for Linear regression is 0.1295 and for poission regression is 0.12089 which has sminimal difference

R-Code:

```
residuals(model1)
r <- glm(kids~as.factor(year)+educ+age+I(age^2)+black+east+
        northcen+west+farm+othrural+town+smcity,data=fert,family=poisson()) %>%
  predict(type="response") %>% cor(fert$kids)
r
model2 <- lm(kids~as.factor(year)+educ+age+I(age^2)+black+east+
        northcen+west+farm+othrural+town+smcity,data=fert,family=poisson())
summary(model2)
```