Problem Set 4 Solutions

Question 1-3

```
model1 <-
lm(price~(bdrms+lotsize+sqrft+I(bdrms^2)+I(lotsize^2)+I(sqrft^2))^2,data=hpr
ice1) %>% step(k=log(nrow(hprice1)))
```

The best model I found had these variables: bdrms, lotsize, sqrft, I(bdrms^2), I(lotsize^2), bdrms:lotsize, bdrms:I(lotsize^2), lotsize:I(bdrms^2), lotsize:I(lotsize^2), and I(bdrms^2):I(lotsize^2). The AIC and BIC were 934.4101 and 964.1382. I chose to not include assessment value because assessment is just another measure of the price. Your answer should vary.

```
model2 <- lm(colgpa~(sat+tothrs+athlete+hsrank+female+black)^2,data=gpa2)
%>% step(k=log(nrow(gpa2)))
```

The best model I found had: sat, tothrs, hsrank, female, black, sat:tothrs, and sat:shrank. Your answer should vary. The AIC and BIC were 6892.205 and 6949.154.

```
model3 <-
lm(log(salary)~(teamsal+years+games+atbats+runs+hits+hruns+bavg+bb+sbases+fl
dperc+so+frstbase+scndbase+shrtstop+thrdbase+outfield+catcher+hispan+black+a
llstar)^2,data=mlb1) %>% step(k=log(nrow(mlb1)))
```

The best model I found had 64 variables (too many to list). The AIC and BIC were 592.7275 and 847.9144. Your answer should vary.

Question 4

```
rental <- rental %>% pdata.frame(index=c('city','year'))
plm(log(rent)~y90+log(pop)+log(avginc)+pctstu,model='pooling',data=rental)
%>% summary
```

i) Rents are higher in 1990. Inflation was 26% over the 10 years. That's about normal. Rents are 0.5% higher for every percent increase in the student population.

ii) No. You have to use the Arellano correction or differencing.

```
plm(log(rent)~y90+log(pop)+log(avginc)+pctstu,model='fd',data=rental) %>%
summary
```

iii) Here the student population going up by 1% => 1.1% increase in rental rates.

```
plm(log(rent)~y90+log(pop)+log(avginc)+pctstu,model='within',data=rental)
%>% summary
```

Question 5

i) It should be negative. I'm not sure what beta2 should be.

```
murder <- pdata.frame(murder,index=c('id','year'))
model51 <- plm(mrdrte~exec+unem,data=murder %>% subset(year==90 |
year==93),model='pooling')
```

ii) We don't see a deterrent effect here.

```
model52 <- plm(mrdrte~exec+unem,data=murder %>% subset(year==90 |
year==93),model='fd')
```

iii) Now there is a deterrent effect, but it vanishes in (iv).

```
dt <- data.table(murder)
dt[year==93][order(exec),.(state,exec),]
```

v) Texas had 34. Virginia only had 11. That means Texas is going to have a massive effect on these results. Outliers in the y direction are not serious. Outliers in the x direction spell doom.

```
model53 <- plm(mrdrte~exec+unem,data=murder %>%
subset(state!="TX"),model='within',effect='twoway')
```

vi) The effect is almost twice as big, but it's definitely insignificant.

```
model54 <- plm(mrdrte~exec+unem,data=murder,model='within',effect='twoway')
```

vii) The effect is slightly bigger, but it's insignificant at the 5% level.

Question 6

```
airfare <- pdata.frame(airfare,index=c('id','year'))
model61 <-
plm(log(fare)~concen+log(dist)+I(log(dist)^2),model='within',effect='time',d
ata=airfare)
```

i) A 0.1 increase in concentration leads to a 3.6% increase in fare.

```
library(broom)
tmp <- model61 %>% summary(vcov=plm::vcovHC(model61,method='arellano'))
tmp <- tmp$coefficients %>% tidy
rep(tmp[1,2],2)+tmp[1,3]*c(-1.96,1.96)
```

ii) The original is [0.3011850,0.4190557]. The Arellano is [0.2454754,0.4747653].

```
exp(-tmp[2,2]/2/tmp[3,2])
mean(airfare$dist<exp(-tmp[2,2]/2/tmp[3,2]))
```

iii) It's quadratically increasing when the dist is greater than 79 miles. All of the flight paths are longer than the 79 miles.

```
model62 <-
plm(log(fare)~concen+log(dist)+I(log(dist)^2),model='within',effect='twoways
',data=airfare)
model62 %>% summary
model62 %>% summary(vcov=plm::vcovHC(model62,method='arellano'))
```

iv) A 0.1 increase in concentration => a 1.7% increase in fare

v) Population in the cities. Hub status. Yes, these are highly related to concentration.

vi) Yes. The pooled model is probably best for this purpose.

Question 7

```
glm(approve~white,family=binomial(),data=loanapp) %>% summary
glm(approve~white,family=binomial(),data=loanapp) %>%
predict(data.table(white=c(1,0)),type='response')
lm(approve~white,data=loanapp) %>% predict(data.table(white=c(1,0)))
```

i) The predictions are exactly the same.

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

ii) There's still a discrimination effect. This effect vanishes if we include white:obrat as discussed in class.

Question 8

```
mean(alcohol$employ)
mean(alcohol$abuse)
lm (employ~abuse,data=alcohol)                    %>% coeftest(vcov.=vcovHC)
```

i) is obvious, ii) everything is as expected

```
glm(employ~abuse,data=alcohol,family=binomial()) %>% coeftest(vcov.=vcovHC)
glm(employ~abuse,data=alcohol,family=binomial()) %>% margins
```

iii) These are basically the same. Slightly more significant.

```
lm (employ~abuse,data=alcohol)                    %>
predict(data.table(abuse=c(0,1)))
glm(employ~abuse,data=alcohol,family=binomial()) %>%
predict(data.table(abuse=c(0,1)),type="response")
```

iv) These are exactly the same.

```
lm
(employ~abuse+age+agesq+educ+educsq+married+famsize+white+northeast+midwest+
south+centcity+outercity+qrt1+qrt2+qrt3,data=alcohol) %>%
coeftest(vcov.=vcovHC)
```

v) Now it's insignificant at the 5% level.

```
glm(employ~abuse+age+agesq+educ+educsq+married+famsize+white+northeast+midwe
st+south+centcity+outercity+qrt1+qrt2+qrt3,data=alcohol,family=binomial())
%>% coeftest(vcov.=vcovHC)
```

vi) Here we get the significant result.

```
glm(employ~abuse+age+agesq+educ+educsq+married+famsize+white+northeast+midwe
st+south+centcity+outercity+qrt1+qrt2+qrt3,data=alcohol,family=binomial())
%>% margins
```

vii) Alcohol is a behavioral health problem. A heavier drinker who is healthy in other ways is a complete anomaly. We care about the average drinker.

viii) Alcohol => unemployment but unemployment => alcohol. That's a terrible instrument. Fetal alcohol syndrome. Look it up.

Question 9

```
glm(kids~educ+age+agesq+black+east+northcen+west+farm+othrural+town+smcity+y
74+y76+y78+y80+y82+y84,family=poisson(),data=fertil1) %>% summary
```

i) Women in Botswana had on average 19% fewer living children in 1982 than a woman in 1972. This decline in fertility was due to improved economic conditions.

ii) Black women had on average 36% more living children.

```
r <-
glm(kids~educ+age+agesq+black+east+northcen+west+farm+othrural+town+smcity+y
74+y76+y78+y80+y82+y84,family=poisson(),data=fertil1) %>%
predict(type="response") %>% cor(fertil1$kids)
r^2
lm(kids~educ+age+agesq+black+east+northcen+west+farm+othrural+town+smcity+y7
4+y76+y78+y80+y82+y84,data=fertil1) %>% summary
```

iii) The linear model R-squared is only slightly better than the one from the poisson. This is fairly standard and indicates that the poisson is a good fit. The linear model will always have a higher R-squared than any other generalized linear model.