Homework 3

Following code belongs to the per capita income by state

Code:

```
#Installing the mosaic package
install.packages("mosaic")
require(mosaic)

#Reading the file
data = read.csv("countyComplete.csv")


#Grouping the mean of per capita income and state

mn = mean(data$per_capita_income ~ data$state)


mnf <- function(x){return(mn[[x]])}


mnff <- function(v){sapply(v,mnf)}



data= transform(data,fitted= mnff(state))
data= transform(data, resid=(per_capita_income - fitted))

#Calculating variance
var(data$fitted)
[1] 8438371
var(data$resid)
[1] 20815321
var(data$fitted) + var(data$resid)
[1] 29253692
var(data$per_capita_income)
[1] 29253692
(var(data$fitted)) / (var(data$per_capita_income))
[1] 0.2884549
```

As from the above model, about 29% of the variance is accounted by grouping the counties by state. Grouping is meaningful because it adds a bit more information to the data set. Grouping helps us answer the questions about the average per capita income for a particular state, or the comparison between the states for which state has more average per capita income. If we consider the mean for all the counties together, such questions would be difficult to answer as the data is becomes ambiguous if not grouped. Since the fitted values gives variance which explains or accounts for by the model, whereas the residual values are which remain unexplained or unaccounted for by the model. So, 29 is a high percentage of value that accounts to a lot of the variance value by just grouping the counties by state.

However, answering questions like “What if the “rich states” are simply those where wealthier people happen to live?” would be difficult as it would not be possible to do that as we don’t have information of the count of people in the state as well as the total income of the state. If we were to group the per capita income with respect to the population of the counties, after grouping the data by state, then we could have more information to answer the above question. If the population is very low and the per capita income for that particular population size is high, then we could say that the per capita income variable was controlled by the people. Thereby, answering the question if the rich states were those that had wealthier people.
More analysis could be conducted if we were to group the per capita income by median household income, after the grouping of the counties by state. This would help in answering more questions like, which state has wealthier people? does the state have low per capita income even if there are wealthier people? And would give us more insights regarding which states have houses with more income.

2.

**Model 1**:

Code:

```
#installing mosaic package
install.packages("mosaic")
require(mosaic)

#reading file
group1 = read.csv("countyComplete.csv")

#Taking max and min values to categorize
min(group1$age_over_65)

max(group1$age_over_65)

#Using cut() function to transform the numerical data to categorical
group1$Over65 = cut(group1$age_over_65, breaks=c(0,10,20,30,40,50),
labels=c("Senior_1","Senior_2","Senior_3","Senior_4","Senior_5"))
group1$Over65

#Calculating the mean value
mn_age = mean(group1$poverty ~ group1$Over65)
mn_age

mn_agef <- function(x){return(mn_age[[x]])}

mn_agef("Senior_2")

mn_ageff <- function(v){sapply(v,mn_agef)}
group1$Over65
```

```
group1= transform(group1,fitted= mn_ageff(Over65))
group1= transform(group1, resid=(poverty - fitted))

#Calculating variance
 var(group1$fitted)
[1] 0.542924
 var(group1$resid)
[1] 40.21076
 var(group1$fitted) + var(group1$resid)
[1] 40.75368
 var(group1$poverty)
[1] 40.75368
 (var(group1$fitted)) / (var(group1$poverty))
[1] 0.01332208
```

In this model, we are using “age_over_65” as the explanatory variable to measure poverty in a county for it can be assumed that the number of people above 65 do not have any sources of income. Categorized as “senior citizens” in any country, the people who fall under this age group are usually not capable of working further not having any sources of income. This is not a good model for measuring poverty, as we can see that the age_over_65 variable accounts for just about 1.3 percent. It does not provide us with enough variance to be calculated.

Model 2:

Code:-

```
#reading the file
group2 = read.csv("countyComplete.csv")

# Using cut() function to transform the numerical data to categorical
group2$gradhs = cut(group2$hs_grad, breaks=c(0,25,50,75,100),
labels=c("First","Second","Third","Fourth"))

#calculating mean
mn_hg = mean(group2$poverty ~ group2$gradhs)
mn_hg
#First   Second        Third         Fourth
#NaN   38.00000   22.77212   14.18886

mn_hgf <- function(x){return(mn_hg[[x]])}
mn_hgf("Second")
```

Homework 3

```
#[1] 38

mn_hgff <- function(b){sapply(b,mn_hgf)}
mn_hgff(group2$gradhs)

group2 = transform(group2,fitted=mn_hgff(group2$gradhs))
group2 = transform(group2,resid=(group2$poverty - fitted))



var(group2$fitted)
#[1] 9.64773

var(group2$resid)
#[1] 31.10595

var(group2$fitted) + var(group2$resid)
#[1] 40.75368

var(group2$poverty)
#[1] 40.75368

(var(group2$fitted))/(var(group2$poverty))
#[1] 0.2367327
```

In this model we have used “hs_grad” as the explanatory variable to measure the poverty. We can see that grouping by hs_grad accounts for just over 23 percent of the variance of poverty. As the number of people in a state having completed high school studies contributes directly somehow to the poverty in that particular state. Since, the maximum number of people who have not completed their high school tend to remain jobless without earning any respectable amount of income further contributing lesser to the overall state income.
Which in turn leads into more poverty. For example, higher the rate of illiterate people in a state, more the poverty in that particular state. Hence, this looks as a good model.