ASSIGNMENT 3 – Group-based models

In this assignment we will first reconstruct the model of per capita income by state using the group-based models in the Kaplan text. The dataset we will be working with will be **countyComplete.csv** and will also be making use of the mosaic package.

We read the file by using the read.csv command.

```
require( mosaic )
county = read.csv( "countyComplete.csv" )
```

We will then use the mean command to make sure that the mean of the per_capita_income is index by the grouping variable (in this case the state variable).

```
cm = mean( county$per_capita_income ~ county$state )
cm
```

```
           Alabama             Alaska            Arizona
          19822.18           26830.72           20578.20
          Arkansas         California           Colorado
          18726.41           27025.16           27202.88
       Connecticut           Delaware District of Columbia
          34873.25           27397.67           42078.00
           Florida            Georgia             Hawaii
          23129.94           19715.99           30942.20
             Idaho            Illinois            Indiana
          20518.84           23552.98           22592.82
              Iowa             Kansas            Kentucky
          23673.60           22584.90           19313.23
         Louisiana              Maine            Maryland
          20320.98           23765.75           31459.08
      Massachusetts           Michigan          Minnesota
          33547.07           22198.66           24861.59
       Mississippi           Missouri            Montana
          17624.95           20133.73           22130.21
          Nebraska             Nevada      New Hampshire
          22294.08           25955.06           28904.10
        New Jersey         New Mexico           New York
          34391.19           20629.39           26155.52
    North Carolina       North Dakota               Ohio
          21620.08           24572.98           22624.16
          Oklahoma             Oregon       Pennsylvania
          20497.09           23322.56           23645.55
      Rhode Island     South Carolina       South Dakota
          32741.80           20226.33           21934.17
         Tennessee              Texas               Utah
          20094.04           21521.11           21635.48
           Vermont           Virginia         Washington
          26235.43           25557.64           24445.97
     West Virginia          Wisconsin            Wyoming
          19443.75           24488.12           27239.61
```

We now define a function to map the **state** values with their corresponding **mean** values.

```
cmf = function(x){ return(km[[x]]) }
```

We now generate a function that applies the cmf function over the entire column values.

cmff = function(v) { sapply(v,kmf) }

We now use this function to generate fitted and residual values for the dataset using the transform command and adding those values as a new column in the dataset.

county = transform(county, fitted = cmff(state))
county = transform(county, resid = per_capita_income - fitted)

To demonstrate the partitioning property for the variance we now calculate the variance of the model values and the variance of residual values and show them to be same as the variance of the per_capita variable.

var(county$fitted)

```
[1] 8438371
```

var(county$resid)

```
[1] 20815321
```

var(county$per_capita_income)

```
[1] 29253692
```

var(county$fitted) + var(county$resid)

```
[1] 29253692
```

var(county$fitted)/var(county$per_capita_income)

```
[1] 0.2884549
```

As can be seen from above the variance of fitted and residual values equals to the per_capital_income variable. We can also observe that grouping by state account for about twenty-nine percent of the per capita income which is quite less. The per capita income for each state does not describe the states as having been rich in terms of population. Grouping the data in terms of states does not take into account of the low-income counties in any state.

To address these questions, we can look at the number of persons in the household, housing units and the median household income to look for the relative distribution of wealth in a county. This will reduce the effect of having few rich people living in the states. The modeling of per capita income with state does provide some meaningful insights as we can look for the relative number of firms in the state and also use it to identify high opportunity states where there is high growth.

Now we contrast two different group-based models. Both the models will have different explanatory variable but will try to explain the poverty variable.

First we take the **median household income** variable as an explanatory variable. We calculate the fitted and residual values for the variance as below.

medianf = sapply(county_median$median_household_income, function(x){ if(x < 40000) return("low income") else if (x >= 40000 & x < 60000) return("middle income") else return("high income")})

This is the function used to group the household groups into low income, middle income and high-income groups.

Now we copy our county dataset into a new variable called **county_median** for analysis.

county_median = county

mm = mean(county_median$poverty ~ medianf)

This is our modeling function. We take the mean of the household incomes for the poverty level.

mmf = function(x) { if(x < 40000) return(mm["low income"]) else if (x >= 40000 & x < 60000) return(mm["middle income"]) else return(mm["high income"]) }

mmff = function(v) { sapply(v, mmf) }

The above two statements are used to get the fitted values for the model.

county_median = transform(county_median, fitted = mmmf(county_median $ median_household_income))

This stores the fitted values of our model in a new column fitted in the dataset.

county_median = transform(county_median, resid = county_median$poverty – fitted)

We store the residual values in another column to get the variance of the values.

var(county_median$resid)

```
[1] 20.97673
```

var(county_median$poverty)

```
[1] 40.75368
```

var(county_median$fitted)

```
[1] 19.77695
```

var(county_median$fitted)/var(county_median$poverty)

```
[1] 0.4852802
```

We can see that the median household income accounts for over forty-nine percent of the variance of the poverty variable. This is not high by it does gives us a sense of the relation between the median household income and the poverty variable. The poverty rate is seen to be more in the counties where the median household income in low and vice versa.

We now look at another variable **per capita income** as our explanatory variable with respect to poverty variable.

percapf = sapply(county_percap$per_capita_income, function(x){ if(x<20000) return("low income") else if (x>=20000 & x < 40000) return("middle income") else return("high income")})

The above function is used to group the per capita income into low income, middle income, and high-income people.

We now create a new variable to store our dataset **county_percap** for the analysis with **per capita income** as explanatory variable.

county_percap = county

pm = mean(county_percap$poverty ~ percapf)

This is our modeling function which we will use to create per capita income model with respect to poverty.

pmf = function(x){ if(x<20000) return(pm["low income"]) else if (x>=20000 & x < 40000) return(pm["middle income"]) else return(pm["high income"])}

pmff = function(v){ sapply(v, pmf)}

The above code is used to get the model values in a vector format.

county_percap = transform(county_percap, fitted = pmff(county_percap$per_capita_income))

This code adds the fitted values as a new column in the **county_percap** data frame.

county_percap = transform(county_percap, resid = county_percap$poverty - fitted)

The above code does the same for the residual values.

var(county_percap$resid)

```
[1] 23.44392
```

var(county_percap$poverty)

```
[1] 40.75368
```

var(county_percap$fitted)

```
[1] 17.30976
```

var(county_percap$fitted)/var(county_percap$poverty)

```
[1] 0.4247409
```

The output above show the variance of fitted, residual and per capita income variables. Although the percentage of the percap income variance accounted for the poverty is less than the median household variable it does have some of the relationships with the poverty variable. The poverty rate also tends to be less where the per capita income is high and vice versa. The difference between the percap and household income is due to the fact that a median household income captures the general condition of the people living in the county in a better way than per capita income. The per capita income does get affected by the outliers such as super rich people living in a county with fewer members in the house whereas the median household income does not get affected by extreme values which is a better way to gauge poverty levels.