**Project : Income in Illinois, USA**

**Students: Aishwarya Ravi, Hong Zhang**

1. **Business Scenario**:

Everybody care about the same question how I can earn more money. "Success is a learnable skill," emphasizes T. Harv Eker in his book "Secrets of the Millionaire Mind." "If you want to be a great golfer, you can learn how to do it. If you want to be a great piano player, you can learn how to do it ... If you want to be rich, you can learn how to do it." If you want to learn how to get rich — how to grow and master your money — consider these variables which can affect your income.

1. **Objectives**:

American Community Survey Office is the premier source for detailed information about the American people and workforce. We will choose the dataset of its population survey from 2010 to 2014 and try to do some analysis. The study location will be our living state, Illinois. There are many factors describing a person, and they do not weigh equally in determining this person’s income. We will try to find them, cross compare and build a regression model to predict the personal income.

1. **Data Sources**:

This dataset contains the data of American Community Survey 2010-2014 ACS 5-year PUMS.

Prepared by: American Community Survey Office, U.S. Census Bureau

Create Date: ‎October ‎28, ‎2015

|  |  |  |  |
| --- | --- | --- | --- |
| Data Source | Files | total number of rows | time period |
| <http://www.census.gov/programs-surveys/acs/data/pums.html> | psam\_p17.sas7bdat | 635532 | 2010~2014 |

1. **Variable Selection:**

**Total: 293 variables**

|  |  |  |
| --- | --- | --- |
| **Predictor(X)** | **Variable** | **Type** |
| **SCHG (Grade Level Attending)** | **Categorical** |
| **COW (Class of Worker)** | **Categorical** |
| **SEX(Male or Female)** | **Categorical** |
| **MAR (Marital Status)** | **Categorical** |
| **AGEP(Age)** | **Continuous** |
| **WKHP (No. of hours worked in a week)** | **Continuous** |
| **Response(Y)** | **PINCP (Person’s Total Income)** | **Continuous** |

1. **SAS Code:**

**The SAS procedures we are planning to use are**

**Proc means**

**Proc freq**

**Proc univariate**

**Proc sgplot, etc.**

**Sample Data selection from the dataset:**

/\*Select sample data from dataset\*/

/\*Select particular variables from data set\*/

data census.psam\_p17\_subset1;

set census.psam\_p17;

if AGEP < 16 then delete;

if PINCP <=100 then delete;

if WKHP <=0 then delete;

if FOD1P <0 then delete;

if WRK<=0 then delete;

/\* Because it is data for 5 year, PINCP should use ADJINC to get current PINCP \*/

PINCP=PINCP\*(ADJINC/1000000);

keep SCH SEMP AGEP PINCP ADJINC WKHP WRK COW SCHL FOD1P MAR OCCP10 RELP RAC1P SEX

SCHG ANC1P05 ANC2P05;

label SCH='School Enrolment'

SEMP='Self employment Income'

AGEP='Age'

PINCP='Total person income (signed)'

ADJINC='Adjustment factor for income and earnings dollar amounts (6 implied decimal places)'

WKHP='Usual hours worked per week past 12 months'

WRK='worked last week'

COW='Class of worker'

SCHL='Educational attainment'

FOD1P='Recoded field of degree - first entry'

MAR='Marital status'

RELP='Relationship'

RAC1P='Recoded detailed race code - 0'

SEX='Sex'

OCCP10='Occupation recode'

SCHG='Grade Level Attending'

ANC1P05='Recoded Detailed Ancestry - first entry'

ANC2P05='Recoded Detailed Ancestry - second entry';

run;

**Exploratory Data Analysis**

/\*Exploration of all variables that are available for analysis.\*/

/\*%let statements define macro variables containing lists of continuous variables\*/

%let tfilename=census.psam\_p17\_subset1;

%let interval= AGEP WKHP;

/\*UNIVARIATE proc is used to plot histogram, probability graph and to display basic statistics\*/

proc univariate data=&tfilename;

var &interval;

histogram &interval / normal kernel;

inset n mean median std skewness kurtosis / position=ne;

probplot &interval / normal (mu=est sigma=est);

inset skewness kurtosis;

title 'Descriptive Statistics Using PROC UNIVARIATE';

run;

**Exploration of all categorical variables that are available for analysis**

/\*Exploration of all categorical variables that are available for analysis.\*/

%let categorical= SEX WRK ANC1P05 SCHG COW MAR;

%let tfilename=census.psam\_p17\_subset1;

proc means data=&tfilename maxdec=2 fw=10 printalltypes n mean median std var

q1 q3;

class &categorical;

var PINCP;

output out=means mean=PINCP;

title 'Selected Descriptive Statistics for Persons Income';

run; title;

**Association between categorical response and continuous predictors using SGPLOT (VBOX)**

/\*PROC SGPLOT is used here with the VBOX statement to produce vertical bar charts for various categorical variables\*/

ods graphics on/width=900;

proc sgplot data=census.data\_corrl;

vbox PINCP/ category=SCHG connect=mean;

run;

ods graphics on/width=900;

proc sgplot data=census.data\_corrl;

vbox PINCP/ category=SEX connect=mean;

run;

ods graphics on/width=900;

proc sgplot data=census.data\_corrl;

vbox PINCP/ category=COW connect=mean;

run;

ods graphics on/width=900;

proc sgplot data=census.data\_corrl;

vbox PINCP/ category=MAR connect=mean;

run;

ods graphics on/width=900;

proc sgplot data=census.data\_corrl;

vbox PINCP/ category=WRK connect=mean;

run;

**Produce Scatter Plot using PROC CORR**

/\*Produce Scatter Plot using PROC CORR \*/

proc corr data=census.data\_corrl rank;

var AGEP WKHP;

with PINCP;

title "Correlations with Total Person's Income";

run;

ods graphics / width=700;

proc corr data=census.data\_corrl nosimple

plots=matrix(nvar=all histogram);

var AGEP WKHP;

title "Correlations with Total Person's Income";

run;

/\*Scatter Plot\*/

%let interval=AGEP WKHP;

options nolabel;

proc sgscatter

data= census.data\_corrl;

plot PINCP\*(&interval)/reg;

title "Associations of Interval variables with Person's Income(PINCP)";

run;