**Midterm Project: Person’s Income Analysis**

**SAS Code:**

**Sample Data selection and Format Values from the dataset:**

%let path=/folders/myfolders/census;

libname census "&path";

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

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

/\*dataset variables\*/

%let categorical = Education Occupation MAR SEX Race Country Has\_Children Working\_for;

proc format;

value bonusfmt 0="No"

1="Yes";

value $marfmt "1"="Married"

"2"="Widowed"

"3"="Divorced"

"4"="Separated"

"5"="Never married";

value $sexfmt "1"="Male"

"2"="Female"

run;

/\*Select sample data from dataset\*/

data census.psam\_p17\_subset1;

set census.psam\_p17;

format HINCP Z9.;

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

PINCP=PINCP\*(ADJINC/1000000);

/\* Setup bonus for mid income= 56,210 \*/

format bonus Z1.;

bonus=0;

IF PINCP >=56210 then bonus = 1;

/\* Clean Data \*/

if AGEP <= 16 then delete;

if PINCP <= 100 then delete;

if WKHP <= 0 then delete;

/\* Find the band of Education \*/

format Education $20.;

Education = 'HS-Not Grad';

if SCHL = 16 then Education = 'HS-Grad';

if SCHL = 17 then Education = 'HS-Grad';

if SCHL = 18 then Education = 'Some College';

if SCHL = 19 then Education = 'Some College';

if SCHL = 20 then Education = 'Associate Degree';

if SCHL = 21 then Education = 'Bachelor Degree';

if SCHL = 22 then Education = 'Master Degree';

if SCHL = 23 then Education = 'Professional Degree';

if SCHL = 24 then Education = 'Doctorate Degree';

/\* Find the band of Occupation \*/

format Occupation $10.;

if OCCP10 >= 0010 and OCCP10 <= 0430 then Occupation = 'Manager';

if OCCP10 >= 0500 and OCCP10 <= 3955 then Occupation = 'Specialist';

if OCCP10 >= 4000 and OCCP10 <= 4650 then Occupation = 'Service';

if OCCP10 >= 4700 and OCCP10 <= 4965 then Occupation = 'Sales';

if OCCP10 >= 5000 and OCCP10 <= 5940 then Occupation = 'Office';

if OCCP10 >= 6005 and OCCP10 <= 9750 then Occupation = 'Technician';

if OCCP10 >= 9800 and OCCP10 <= 9830 then Occupation = 'Military';

if OCCP10 = 9920 then Occupation = 'Unemployed';

/\* Find the band of Native Country \*/

format Country $12.;

Country = 'Others';

if ANC1P05 = 211 then Country = 'American';

if ANC1P05 = 227 then Country = 'American';

if ANC1P05 = 249 then Country = 'American';

if ANC1P05 = 250 then Country = 'American';

if ANC1P05 = 295 then Country = 'American';

if ANC1P05 = 900 then Country = 'American';

if ANC1P05 = 917 then Country = 'American';

if ANC1P05 = 924 then Country = 'American';

if ANC1P05 = 939 then Country = 'American';

if ANC1P05 = 940 then Country = 'American';

if ANC1P05 = 983 then Country = 'American';

if ANC1P05 = 994 then Country = 'American';

/\* Find the Family Relationship \*/

format Has\_Children $4.;

Has\_Children = 'Yes';

if SFR = 1 then Has\_Children = 'No';

/\* Find the band of Class of Worker \*/

format Working\_for $16.;

if COW = 1 then Working\_for = 'Private Company';

if COW = 2 then Working\_for = 'Private Company';

if COW = 3 then Working\_for = 'Government';

if COW = 4 then Working\_for = 'Government';

if COW = 5 then Working\_for = 'Government';

if COW = 6 then Working\_for = 'Self-employed';

if COW = 7 then Working\_for = 'Self-employed';

if COW = 8 then Working\_for = 'Self-employed';

if COW = 9 then Working\_for = 'Unemployed';

/\* Find the band of Race \*/

format Race $16.;

Race = 'American Native';

if RAC1P = 1 then Race = 'White';

if RAC1P = 2 then Race = 'Black';

if RAC1P = 6 then Race = 'Asian';

if RAC1P = 8 then Race = 'Others';

if RAC1P = 9 then Race = 'Others';

keep PINCP Education Occupation AGEP WKHP MAR SEX Race Country Has\_Children Working\_for bonus;

label PINCP = 'Total person income (signed)'

AGEP = 'Age'

WKHP = 'Hours worked per week'

MAR = 'Marital status'

Country = 'Native Country'

Has\_Children = 'Has Children'

Working\_for = 'Working for'

bonus = 'Income >= $56,210'

run;

**Select first 500 Observations**

/\*Survey select only 500 obs\*/

proc surveyselect data=census.psam\_p17\_subset1 method=srs n=500

out=census.data\_corr;

run;

**Exploratory Data Analysis: Association between continuous predicotrs and response using scatter plot and correlation matrix**

/\*\*\*\*\*\*\*\*\*Continuous X: AGEP WKHP\*\*\*\*\*\*\*\*\*\*/

/\*Check PINCP and AGEP WKHP\*/

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

proc corr data=census.data\_sample rank;

var AGEP WKHP;

with PINCP;

title "Correlations with Total Person's Income";

run;

ods graphics / width=700;

proc corr data=census.data\_sample 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\_sample;

plot PINCP\*(&interval)/reg;

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

run;

**Exploratory Data Analysis**

/\*Exploration of all categorical variables that are available for analysis SEX WRK SCHG COW MAR\*/

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

%let categorical= Education Occupation;

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

;

class &categorical;

var PINCP;

output out=means mean=PINCP;

title 'Selected Descriptive Statistics for Persons Income';

run;

title;

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

/\*PROC SGPLOT is used here with the VBOX statement to produce vertical boxcharts\*/

ods graphics on/width=700;

proc sgplot data=census.data\_corr;

vbox PINCP/ category=Education connect=mean;

run;

proc sgplot data=census.data\_corr;

vbox PINCP/ category=Occupation connect=mean;

run;

proc sgplot data=census.data\_sample;

vbox PINCP/ category=MAR connect=mean;

format MAR $marfmt.;

run;

proc sgplot data=census.data\_sample;

vbox PINCP/ category=SEX connect=mean;

format SEX $sexfmt.;

run;

proc sgplot data=census.data\_sample;

vbox PINCP/ category=Race connect=mean;

run;

proc sgplot data=census.data\_sample;

vbox PINCP/ category=Country connect=mean;

run;

proc sgplot data=census.data\_sample;

vbox PINCP/ category=Has\_Children connect=mean;

run;

proc sgplot data=census.data\_sample;

vbox PINCP/ category=Working\_for connect=mean;

run;

**Proc Frequency for categorical Variables:**

/\*PROC FREQ is used with categorical variables\*/

ods graphics;

proc freq data=census.psam\_p17\_subset1;

tables &categorical / plots=freqplot ;

format SEX $sexfmt.

MAR $marfmt.;

title "Categorical Variable Frequency Analysis";

run;

**Linear Regression Modeling for Continuous predictors and Response**

/\*PROC REG using PINCP and AGEP WKHP\*/

ods graphics;

proc reg data=census.psam\_p17\_subset1;

model PINCP=AGEP;

title "Simple Regression with PINCP as Regressor";

run;

quit;

ods graphics;

proc reg data=census.psam\_p17\_subset1;

model PINCP=WKHP;

title "Simple Regression with PINCP as Regressor";

run;

quit;

**Regression modeling of categorical predictors and continuous response**

/\*PROC GLM using PINCP as response and Education Occupation MAR SEX Race Country Has\_Children Working\_for as predictor variable\*/

ods graphics;

proc glm data=census.psam\_p17\_subset1 plots=diagnostics;

class &categorical;

model PINCP=&categorical;

means &categorical / hovtest=levene;

format SEX $sexfmt.

MAR $marfmt.;

title "One-Way ANOVA with categorical variables as Predictor";

run;

quit;

**Categorical predictors' analysis using n-way ANCOVA with interaction**

/\*Categorical predictors' analysis using n-way ANCOVA with interaction\*/

ods graphics on / width=800;

proc glm data=census.psam\_p17\_subset1 plots(only)=intplot;

class Education Occupation;

model PINCP=Education Occupation WKHP/solution;

title "Analyze the overall effect of all Variables";

title2 "Including Interaction";

run;

quit;

**Chi-Square Test**

/\*Chi-Square Test\*/

ods graphics off;

proc freq data=census.psam\_p17\_subset1;

tables (&categorical)\*bonus /

chisq expected cellchi2 nocol nopercent relrisk;

title 'Associations with Bonus';

format SEX $sexfmt.

MAR $marfmt.

bonus bonusfmt.;

run;

ods graphics on;

**Proc Logistic for Bonus**

/\*PROC LOGOSTIC FOR BONUS\*/

ods graphics on;

proc logistic data=census.psam\_p17\_subset1 alpha=.05

plots(only)=(effect oddsratio);

class Occupation(ref='Manager') Education(ref='Master Degree') /param=ref;

model bonus(event='1')= WKHP Occupation Education / clodds=pl;

units WKHP=10;

title 'LOGISTIC MODEL: bonus = WKHP Occupation Education';

run;