Interim Report

Group 1

4/20/2019

Introduction

Based on the questions that we proposed from the last assignment, and the provided feedback, our team decided to start from there as a start point. Total person income per year (WAGP) in Oregon was the question of interest for this report. There are many factors that can influence the rate of income; however, considering the variables which have reasonable explanation would be the strategy of fitting the model.

Objective

The purpose of this project is to identify variables that may be considered as a contributing factor in the analysis of a total person income annually. In other words, we would like to find a reasonable answer for: What really makes a person having higher or lower income per year? And how other factors affect this rate.

Data

The data was provided by Dr. Sharmodeep from the American Community Survey (ACS). In this report we considered Public Use Microdata Samples (PUMPS) data for the years between (2013 to 2017) in the State of Oregon.

Exploratory Analysis

One of the most recommended strategy to fit any model is to visualize the variables. Accordingly,

Preliminary Models

MU{ WAGP | ... } = B0 + B1*JWAP + B2*AGEP + B3*WKHP + B4*SCHL + B5*PWGTP + B6*COW1 + B7*COW2 + B8*COW3 + B9*COW4 + B10*JWTR1 + B11*JWTR2 + B12*JWTR3 + B13*JWTR4 + B12*MAR + B13*MIL + B14*SEX + B15*WKW1 + B16*WKW2 + B17*WKW3 + B18*INSUR

Model Checking and Model Selection

Model Checking

In order to run the linear regression model, the assumptions of this model should be met. Since the assumptions apply on the response variable, WAGP was investigated. 1) multicollinearity: the pairs of observations are independent of each other, by checking the variance inflation factor (VIF), this assumption seems met (see table xx). 2) normality: this assumption was checked by using (Q-Q plot), figure xx shows that this assumption is not met so we transfer the response variable by using log (PINCP) as shown in Figure

xx and this assumption seems better with log. Also, this assumption should not be a concern if we have a big sample size. 3) linearity: this assumption was check by using scatter plot (ggpairs function) between the response variable and the explanatory variables, this assumption seems met as shown in figure xx . 4) constant variance and outliers: these two assumptions were assessed by using residual vs fitted as shown in figure xx and these assumptions seems met too.

Model Selection

As it was mentioned earlier xxx explanatory variables were provided for this study. Lasso technique were applied to get the best model. First, all the variables were included in the model. Then, Lasso technique were used to dropped all the non-significant variables as shown in equation xx. Our target was to get a final model that would be the one with the best and most significant explanatory variables in it. Finally, xxx variables were selected to be the most significant one (p-value < 0.05). Table xx provides description of all the explanatory variables in the final model.

MU{ WAGP | ... } = B0 + B1*AGEP + B2*WKHP + B3*SCHL + B4*MAR + B5*SEX + B6*WKW3 + B7*INSUR + B8*JWMNP

While table xx presents the final results of fitting the best multiple linear regression model to the dataset including estimates of coefficient, standard error, z-value and corresponding p-value.

Conclusion

In this section, only the variables that were significant will be discussed Variable xx

CODE

Setup and Transformations

```
library(tidyverse)
library(GGally)
library(ggplot2)
library(glmnet)
library(faraway)

pop_data <- read_csv("psam_p41.csv")

#pop_data <- tbl_df(pop_data)
pop_new <- select(pop_data, AGEP, JWMNP)

#pop_new$PINCP <- as.numeric(pop_data$PINCP)
pop_new$WAGP <- as.numeric(pop_data$WAGP)
pop_new$JWMNP <- as.numeric(pop_data$JWMNP)

pop_new$JWAP <- as.numeric(pop_data$JWAP)
pop_new$AGEP <- as.numeric(pop_data$AGEP)
pop_new$SCHL <- as.numeric(pop_data$SCHL)
pop_new$WKHP <- as.numeric(pop_data$WKHP)
pop_new$PWGTP <- as.numeric(pop_data$PWGTP)
#pop_new$WKW <- as.numeric(pop_data$WKW)

cow1_f <- function(x){
  if(!is.na(x) && (x == 1 || x == 2)){
    return(1)
  } else {
    return(0)
  }
}
cow4_f <- function(x){
  if(!is.na(x) && (x == 8 || x == 9)){
    return(1)
  } else {
    return(0)
  }
}
cow2_f <- function(x){
  if(!is.na(x) && (x == 3 || x == 4 || x == 5)){
    return(1)
  } else {
    return(0)
  }
}
cow3_f <- function(x){
  if(!is.na(x) && (x == 6 || x == 7)){
    return(1)
  } else {
    return(0)
  }
```

```
}

jwtr1_f <- function(x){
  if(!is.na(x) && (x == 1 || x == 8)){
    return(1)
  } else {
    return(0)
  }
}
jwtr2_f <- function(x){
  if(!is.na(x) && (x >=2 && x <= 7)){
    return(1)
  } else {
    return(0)
  }
}
jwtr3_f <- function(x){
  if(!is.na(x) && (x == 9 || x == 10)){
    return(1)
  } else {
    return(0)
  }
}
jwtr4_f <- function(x){
  if(!is.na(x) && (x == 11 || x == 12)){
    return(1)
  } else {
    return(0)
  }
}

wkw1_f <- function(x){
  if(!is.na(x) && ( x >= 5)){
    return(1)
  } else {
    return(0)
  }
}
wkw2_f <- function(x){
  if(!is.na(x) && ( x == 4)){
    return(1)
  } else {
    return(0)
  }
}
wkw3_f <- function(x){
  if(!is.na(x) && (x <= 3 )){
    return(1)
  } else {
    return(0)
  }
}
```

```
mar_f <- function(x){
  if(!is.na(x) && (x == 1)){
    return(1)
  } else {
    return(0)
  }
}

mils_f <- function(x){
  if(!is.na(x) && (x == 1 || x == 2 || x == 3)){
    return(1)
  } else {
    return(0)
  }
}

sex_f <- function(x){
  if(!is.na(x) && (x == 1)){
    return(1)
  } else {
    return(0)
  }
}

insur_f <- function(x,y){
  if(!is.na(x) && (x == 2 && y == 2)){
    return(0)
  } else {
    return(1)
  }
}

pop_new$COW1 <- as.numeric(lapply(as.numeric(pop_data$COW), cow1_f))
pop_new$COW2 <- as.numeric(lapply(as.numeric(pop_data$COW), cow2_f))
pop_new$COW3 <- as.numeric(lapply(as.numeric(pop_data$COW), cow3_f))
pop_new$COW3 <- as.numeric(lapply(as.numeric(pop_data$COW), cow4_f))

pop_new$JWTR1 <- as.numeric(lapply(as.numeric(pop_data$JWTR), jwtr1_f))
pop_new$JWTR2 <- as.numeric(lapply(as.numeric(pop_data$JWTR), jwtr2_f))
pop_new$JWTR3 <- as.numeric(lapply(as.numeric(pop_data$JWTR), jwtr3_f))
pop_new$JWTR4 <- as.numeric(lapply(as.numeric(pop_data$JWTR), jwtr4_f))

pop_new$WKW1 <- as.numeric(lapply(as.numeric(pop_data$WKW), wkw1_f))
pop_new$WKW2 <- as.numeric(lapply(as.numeric(pop_data$WKW), wkw2_f))
pop_new$WKW3 <- as.numeric(lapply(as.numeric(pop_data$WKW), wkw3_f))

pop_new$MART <- as.numeric(lapply(as.numeric(pop_data$MAR), mar_f))
pop_new$MILT <- as.numeric(lapply(as.numeric(pop_data$MIL), mils_f))
pop_new$SEXT <- as.numeric(lapply(as.numeric(pop_data$SEX), sex_f))

pop_new$INSUR <- as.numeric(mapply(insur_f, pop_data$PRIVCOV, pop_data$PUBCOV))

pop_omit <- na.omit(pop_new)
```

```
y <- pop_omit$WAGP
j_rem <- pop_omit$JWMNP

pop_omit$WAGP <- NULL
pop_omit$JWMNP <- NULL

X <- as.matrix(pop_omit)

lasso <- glmnet(X, y)
lasso.cv <- cv.glmnet(X, y)
```

Figure 1: LASSO Results

```
coef(lasso.cv)
```

```
## 20 x 1 sparse Matrix of class "dgCMatrix"
##                            1
## (Intercept) -101686.2840
## AGEP            249.4693
## JWAP              .
## SCHL           3392.9480
## WKHP           1063.7469
## PWGTP             .
## COW1           6893.1108
## COW2              .
## COW3              .
## JWTR1             .
## JWTR2             .
## JWTR3             .
## JWTR4             .
## WKW1              .
## WKW2              .
## WKW3          12419.2682
## MART           8554.4386
## MILT              .
## SEXT           7672.5970
## INSUR          7226.2810
```

Figure 2: Fit the New Linear Model

```
pop_lm_data <- na.omit(pop_new)
pop_lm <- lm(log(WAGP) ~ JWMNP+AGEP+SCHL+WKHP+WKW3+MART+SEXT+INSUR, data = filter(pop_lm_data, WAGP > 25
summary(pop_lm)
```

```
##
## Call:
## lm(formula = log(WAGP) ~ JWMNP + AGEP + SCHL + WKHP + WKW3 +
##     MART + SEXT + INSUR, data = filter(pop_lm_data, WAGP > 2500))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -4.1941 -0.3651  0.0193  0.3887  4.0095
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 6.2065572  0.0182897  339.35   <2e-16 ***
## JWMNP       0.0018104  0.0001119   16.18   <2e-16 ***
## AGEP        0.0093150  0.0001755   53.07   <2e-16 ***
## SCHL        0.0672650  0.0007620   88.28   <2e-16 ***
## WKHP        0.0337366  0.0002171  155.38   <2e-16 ***
## WKW3        0.8369684  0.0084225   99.37   <2e-16 ***
## MART        0.2096536  0.0049705   42.18   <2e-16 ***
## SEXT        0.1654019  0.0047828   34.58   <2e-16 ***
## INSUR       0.2276807  0.0086423   26.34   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.634 on 75117 degrees of freedom
## Multiple R-squared:  0.5107, Adjusted R-squared:  0.5106
## F-statistic:  9799 on 8 and 75117 DF,  p-value: < 2.2e-16
```

Figure 3: Check the Multicolinearity

```
vif(pop_lm)
```

```
##    JWMNP     AGEP     SCHL     WKHP     WKW3     MART     SEXT    INSUR
## 1.006606 1.116138 1.056481 1.130336 1.069587 1.129980 1.066073 1.065811
```

Figure 4: Check linearity Assumptions

```
pop_lm_data$LOGAGEP <- log(pop_lm_data$AGEP)
pop_lm_data$LOGWAGP <- log(pop_lm_data$WAGP)

ggpairs(sample_n(filter(select(pop_lm_data, LOGWAGP, JWMNP, SCHL, AGEP, WKHP),LOGWAGP > log(2500), JWMN
```

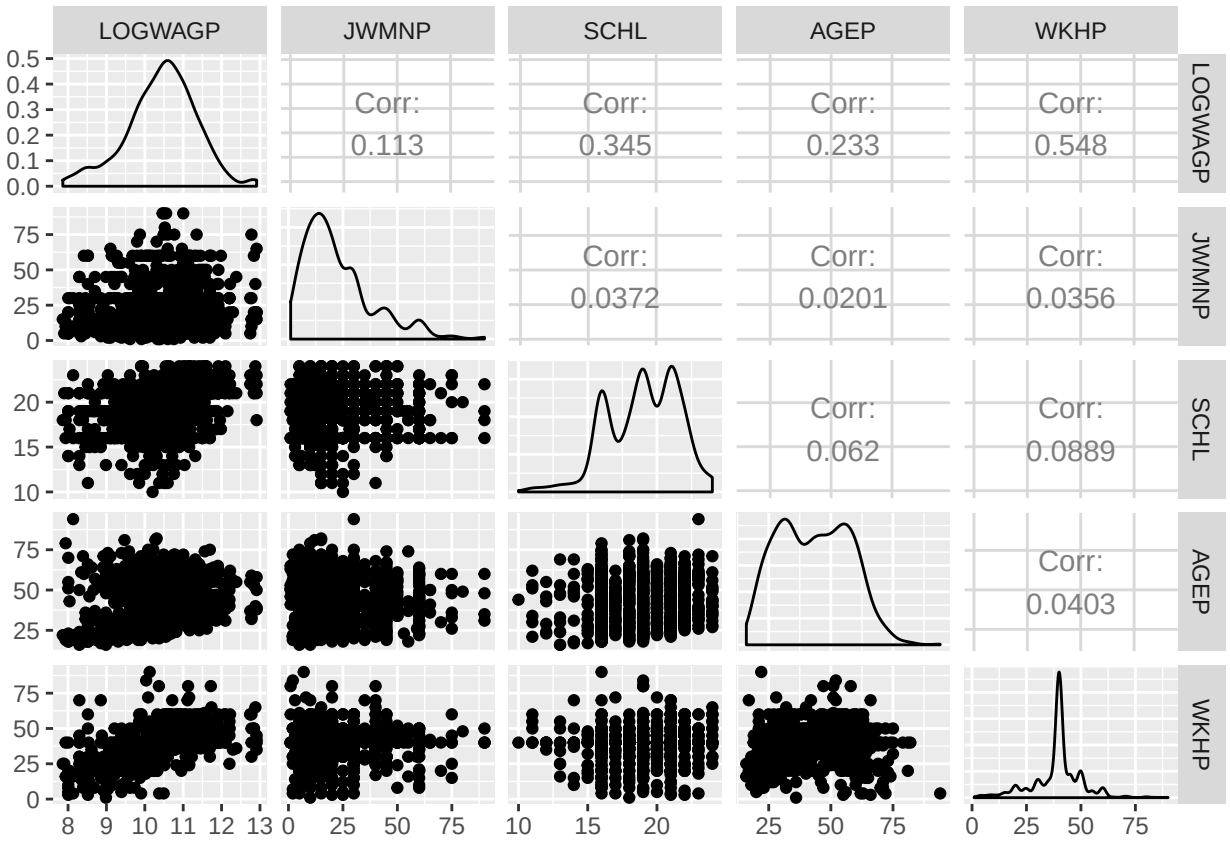
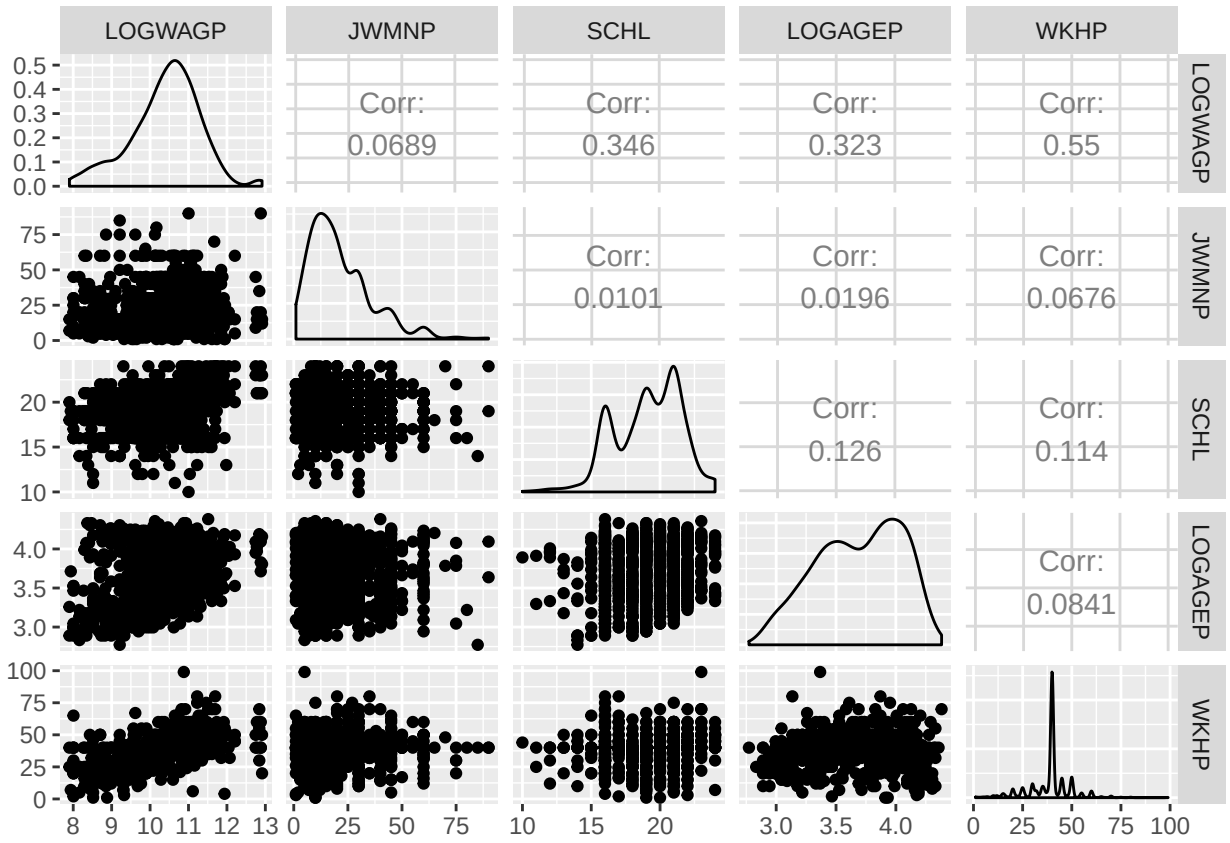


Figure 5: Log of Age

```
ggpairs(sample_n(filter(select(pop_lm_data, LOGWAGP, JWMNP, SCHL, LOGAGEP, WKHP),LOGWAGP > log(2500), JW
```


LOGWAGP
JWMNP
SCHL
LOGAGEP
WKHP
Corr:
0.0689
Corr:
0.346
Corr:
0.323
Corr:
0.55
Corr:
0.0101
Corr:
0.0196
Corr:
0.0676
Corr:
0.126
Corr:
0.114
Corr:
0.0841