Project 1 Report

Group 1: Finn Womack, Hisham Jashami, and Rama Krishna Baisetti

April 2019

Background

Question 1: Is there any relationship between work travel time and wage for employed Oregonians?

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. Is there a relationship between travel time for work (JWMNP) and Total occupational 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 strategy recomemded in the model selection lectures is to fit a model using explanetry variables which are not of interest, preform model selection, and then add the variable of interest into the final selected model. (Variables defined in table 1)

Preliminary Models

$$\begin{aligned}\mu\{WAGP|...\} =&\beta_0 + \beta_1 * JWAP + \beta_2 * AGEP + \beta_3 * WKHP + \beta_4 * SCHL + \beta_5 * PWGTP \\ &+ \beta_6 * COW1 + \beta_7 * COW2 + \beta_8 * COW3 + \beta_9 * COW4 + \beta_{10} * JWTR1 \\ &+ \beta_{11} * JWTR2 + \beta_{12} * JWTR3 + \beta_{13} * JWTR4 + \beta_{14} * MAR + \beta_{15} * MIL \\ &+ \beta_{16} * SEX + \beta_{17} * WKW1 + \beta_{18} * WKW2 + \beta_{19} * WKW3 + \beta_{20} * INSUR\end{aligned}$$

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 figure 4). 2) normality: this assumption was checked by using (Q-Q plot), figure 2 shows that this assumption is not met so we transform the response variable by using log(WAGP) as shown in Figure 3 and this assumption seems better with log and through filtering ouliers who made less that $2500 last year. 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 5 . 4) constant variance: this assumption was assessed by using residual vs fitted as shown in figure 3 and this assumption seems met too.

Model Selection

As it was mentioned earlier 20 explanatory variables were initially examined for this study. Lasso technique were applied to get the best model. First, all the variables were included in the model. Then, we used lasso to find significant variables (figure 1) and dropped all the non-significant variables as shown in the equation below. Our target was to get a final model that would be the one with the best and most significant explanatory variables in it. Finally, the variable of interest was added as an aditional variable to the model.

$$\begin{aligned}\mu\{WAGP|...\} =&\beta_0 + \beta_1 * AGEP + \beta_2 * WKHP + \beta_3 * SCHL + \beta_4 * MAR + \beta_5 * SEX \\ &+ \beta_6 * WKW3 + \beta_7 * INSUR + \beta_8 * JWMNP\end{aligned}$$

While figure 6 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.

Inference

Variable JWMNP is a statistically significant (P-value < 0.001). Keeping all other variables constant, the outcome of a single person increases by (exp(0.0018)-1)*100 = 18% for each minute increase in the travel time. This can have multiple interpretation, one possible meaning can be as some people prefer a job with high salary even though the job is higher. However, some they prefer less salary than traveling longer distances.

Question 2: Is there a relationship between worker class and hours of work per week?

Introduction

Objective

Data

Exploritory Analysis

Preliminary Models

Model Checking and Model Selection

Model Checking

Model Selection

Inference

Question 3: Is there a relationship between veteran term of service and veteran disability percentage?

Introduction

Objective

Data

Exploritory Analysis

Preliminary Models

Model Checking and Model Selection

Model Checking

Model Selection

Inference

Obstacles

Discussion

Appendix: R Code and Plots

Table 1: Variable Descriptions

VARS	DESCRIPTIONS
WAGP	Ocupational income for the year
JWMNP	Travel time to work in minutes
JWAP	Arrival time in 15 minutes after 12 AM
AGEP	Age in years
WKHP	Normal hors worked in a week
SCHL	Educational attainment
PWGTP	Weight
COW1	Private employee indicator
COW2	Public employee indicator
COW3	Self-employed indicator
COW4	Unemployed/unpaid indicator
JWTR1	Drove to work indicator
JWTR2	Rode pubic transit to work indicator
JWTR3	Biked/Walked to work indicator
JWTR4	Worked from home indicator
MAR	Married indicator
MIL	Military service indicator
SEX	Sex indicator (1 male/0 female)
WKW1	Worked $<$ 20 weeks worked in past year indicator
WKW2	Worked between 20 and 40 week in last year indicator
WKW3	Worked $>$ 40 weeks in past year indicator
INSUR	Have insurance indicator
VPS1	Served in either Gulf war
VPS2	Served between WWII and Vietnam
VPS3	Served before WWII
DRAT	Veteran Disability Percentage

Load Packages

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

Load from csv

```
pop_data <- read_csv("psam_p41.csv")
```

Create Subset Data Frame

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

Convert variable to numerical type

```
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)
```

Define functions for indicator variabes for use in LASSO

```
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)
  }
}
```

Create Indicator Variables

```
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))
```

Remove na Values, responce, and Important Variable and Convert to Matrix

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

y <- pop_omit$WAGP

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

```
X <- as.matrix(pop_omit)
```

Preform LASSO

```
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) -96423.6663
## AGEP           231.4529
## JWAP             .
## SCHL          3300.3515
## WKHP          1054.1403
## PWGTP            .
## COW1          5906.7181
## COW2             .
## COW3             .
## JWTR1            .
## JWTR2            .
## JWTR3            .
## JWTR4            .
## WKW1             .
## WKW2             .
## WKW3         11933.5242
## MART          8216.5852
## MILT             .
## SEXT          7097.1165
## INSUR         6426.2068
```

Figure 2: Check assumptions

```
pop_lm_data <- na.omit(pop_new)
fit3 <- lm(WAGP~ JWAP, data=pop_lm_data)
 #summary(fit3)
 par(mfrow = c(2, 2))
 plot(fit3)
```

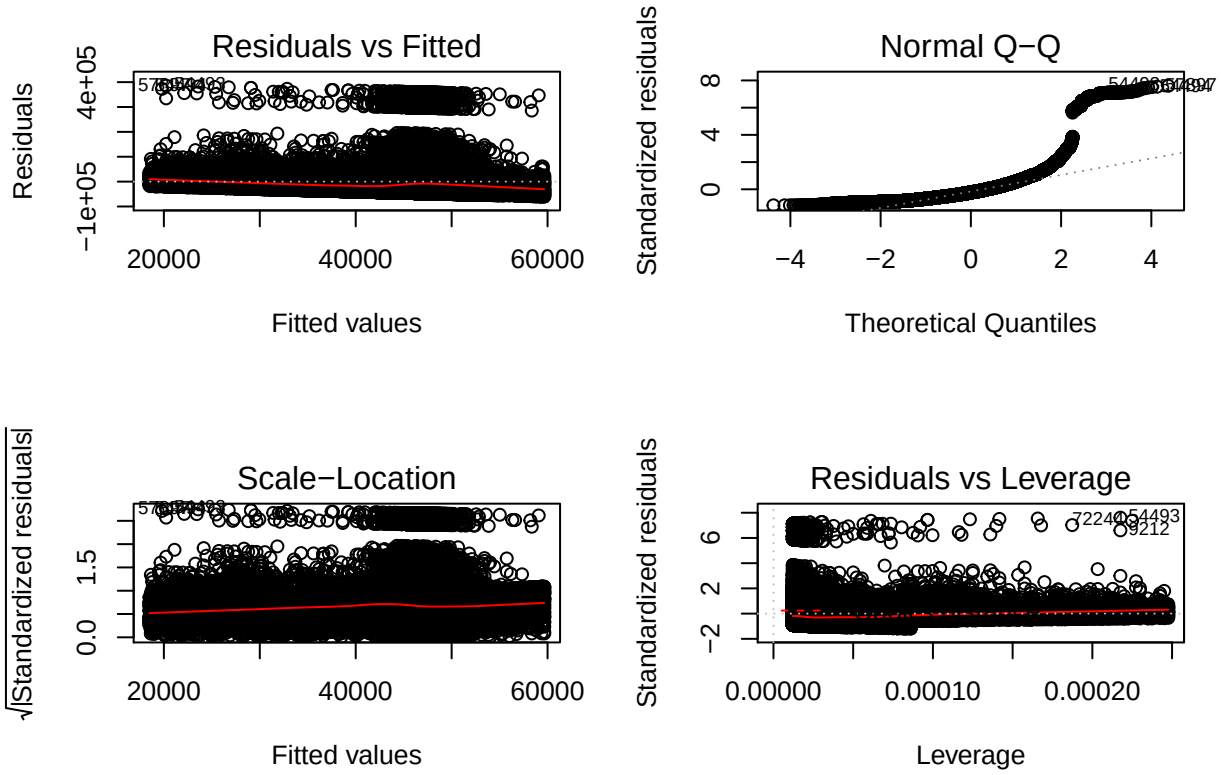

Figure 3: Take Log of WAGP

```
fit3 <- lm(log(WAGP)~ JWAP, data=filter(pop_lm_data, WAGP > 2500))
 #summary(fit3)
 par(mfrow = c(2, 2))
 plot(fit3)
```

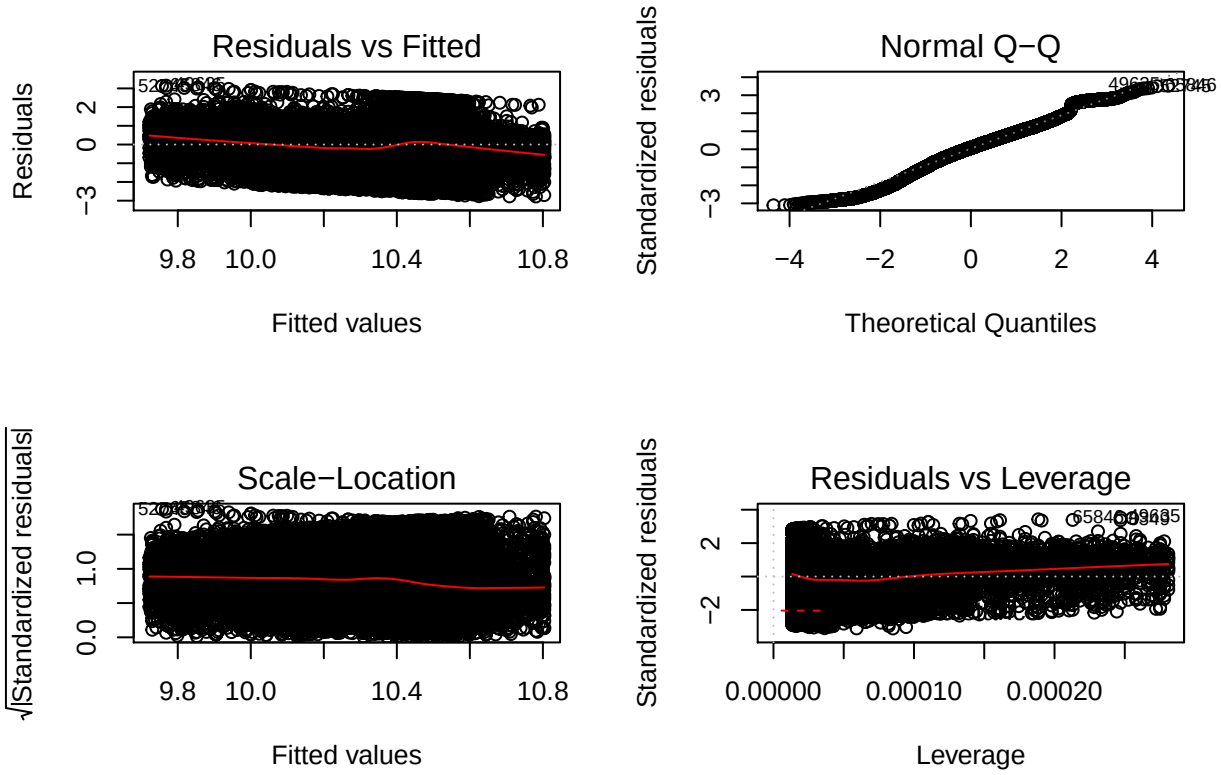

Figure 4: Check the Multicolinearity

```
pop_lm <- lm(log(WAGP) ~ JWMNP+AGEP+SCHL+WKHP+WKW3+MART+SEXT+INSUR,
             data = filter(pop_lm_data, WAGP > 2500))

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 5: 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), JWMNP < (120),
                        SCHL > 9), 1000))
```

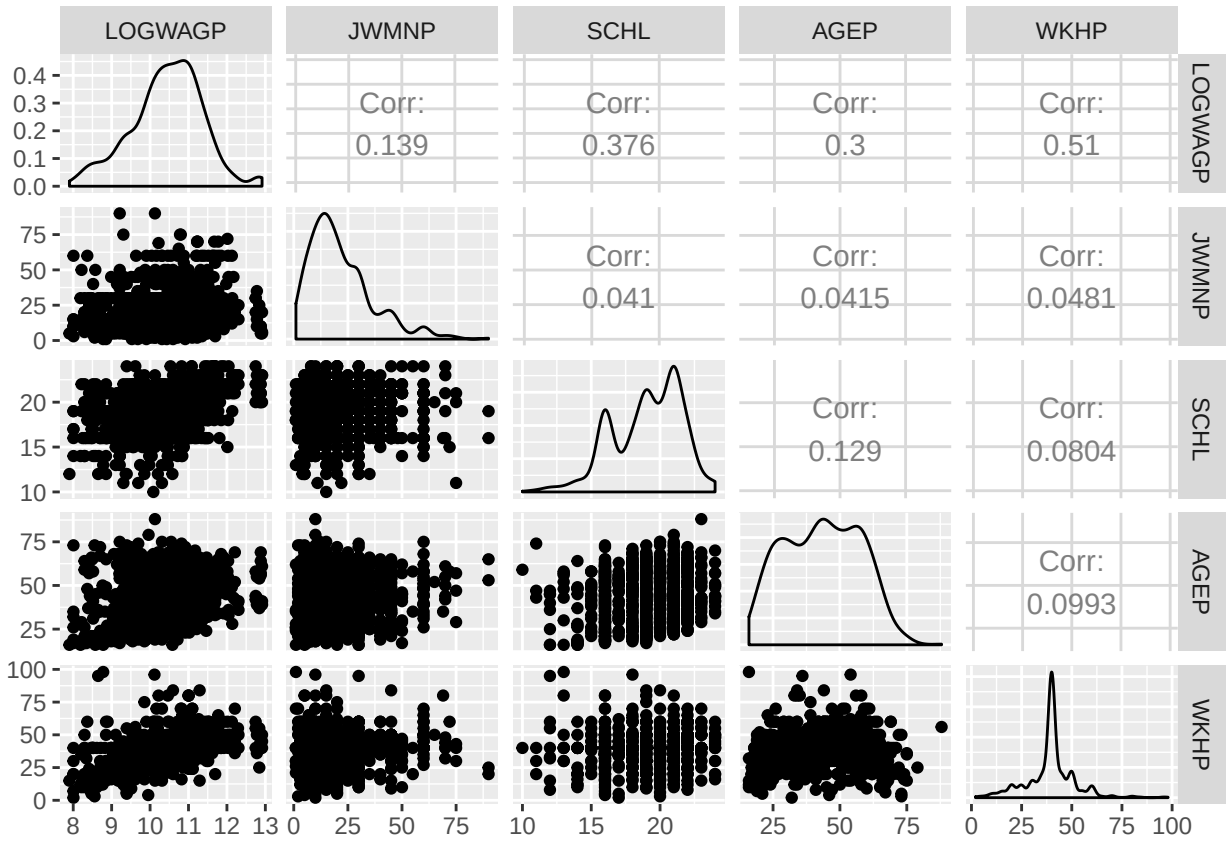

Figure 6: Summary of Linear Model

```
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
```