Project 1 Report

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

April 2019

Background

Dr. Sharmodeep provided us with the data through the US Census Bureau website which is related to the American Community Survey (ACS). The data is compiled into different years of period. The data used for this report is a 5 year Public Use Microdata Samples (PUMS) population records for the years 2013 to 2017 in the State of Oregon. The observations include a total of 200,159 (N) data points with 286 (p) variables. Based on the strategy mentioned in the model selection lectures we fit a model using explanatory variables which are not of interest, perform model selection, and then add the variable of interest into the final selected model. (Variables defined in table 1)

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

Exploratory Analysis

The initial exploration of WAGP (wages) with JWMNP (travel time for work) shows the data with high spread for which we performed log transformation on WAGP. Figure 5 shows the exploration of Person's salary income associated with travel time to work. In addition, we explored how the Person's school attainment, age and hours of work per week are associated to WAGP.

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

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. We started with checking the variance inflation factor (VIF, see figure 4). 1) 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. 2) 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 . 3) constant variance: this assumption was assessed by using residual vs fitted as shown in figure 3 and this assumption seems met too.

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.

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

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?

Exploratory Analysis

The question we are trying to understand, whether there is a relationship between the class of worker (COW) and the hours of work per week (WKHP). Figure 10 shows the initial exploration of Person's Work hours per week associated with the class of worker. In addition, we explored how the Person's wage per hour, travel time to work and time of arrival is associated to WKHP.

Preliminary Models

$$\begin{aligned}\mu\{WKHP|...\} =&\beta_0 + \beta_1 * JWAP + \beta_2 * AGEP + \beta_3 * WAGP + \beta_4 * SCHL + \beta_5 * PWGTP \\ &+ \beta_6 * JWTR1 + \beta_7 * MAR + \beta_8 * MIL + \beta_9 * SEX + \beta_{10} * WKW3+ \\ &+ \beta_{11} * INSUR + \beta_{12} * JWMNP\end{aligned}$$

Model Checking and Model Selection

In order to run the linear regression model, the assumptions of this model should be met. Since the assumptions apply on the response variable, WKHP was investigated. We started with checking the variance inflation factor (VIF, see figure 9). 1) normality: this assumption was checked by using (Q-Q plot), figure 8 shows that this assumption is met. Also, this assumption should not be a concern if we have a big sample size. 2) 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 10 . 3) constant variance: this assumption was assessed by using residual vs fitted as shown in figure 8 and this assumption seems met too.

Of the 12 explanatory variables that were initially examined for this study 4 non-significant variables were removed after Lasso technique was applied. 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 which in this case was a categorical variable with 4 states thus 3 indicators were added with the 4th being considered as a reference for the others.

$$\begin{aligned}\mu\{WKHP|...\} =&\beta_0 + \beta_1 * JWAP + \beta_3 * WAGP + \beta_4 * JWTR1 + \beta_5 * MAR \\ &+ \beta_6 * SEX + \beta_7 * WKW3 + \beta_8 * JWMNP + \beta_9 * COW2 + \beta_{10} * COW3 \\ &+ \beta_{11} * COW4\end{aligned}$$

Figure 11 presents the summary of results for fitting this linear model.

Inference

Based on the first question of interest, linear regression model was used to predict normal hours worked per week (WKHP) by mainly using class of work as an explanatory variable. As mentioned, we broke down this variable into 4 categorical variables. In the model, we used private employee as the reference. The other three were statistically significant (P-value < 0.001). For example, a person who works as a public employee, it is more likely that he/she will work more hours in a week than being a private employee by (0.275). One possible explanation could be that usually public employee has to work the regular hours per week while on the other hand, for private sector, it could depend on the workload of the company or how fast an employee finishes their work. While being a self-employed or unemployed, it is less likely that they work less hours than being a private employee by 1.06 and 2.68 respectively. This also makes sense and it is self-explanatory. Other explanatory variables were also statistically significant (P-value < 0.001) (β_p not equal zero) except for the marriage status was not statistically significant.

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

Exploratory Analysis

The question we are trying to understand, whether there is a relationship between the veteran period of service (VPS) and the veteran disability percentage (DRAT). With the initial exploration of veteran period of service which is categorized into 15 categories we transformed into two groups I & II Gulf war, Between Vietnam & WWII to easily understand the relationship which major period of service has the effect on DRAT.

Preliminary Models

$$\begin{aligned}\mu\{DRAT|...\} =&\beta_0 + \beta_1 * JWMNP + \beta_2 * AGEP + \beta_3 * WAGP + \beta_4 * SCHL + \beta_5 * PWGTP \\ &+ \beta_6 * WKHP\end{aligned}$$

Model Checking and Model Selection

The assumptions for the linear regression model are observed with DRAT as response. 1) Normality: Based on the Q-Q plot of the linear model fit the data follows the straight line with slight variation assuming it should not be a concern for large amounts of data 2) Equal Variance: Based on the residual plot of the linear model fit (residual vs fitted) we observe the residuals are centered around 0 and follows equal spread which shows the assumption seems valid. 3) Linearity: With the initial exploration using scatterplot there seems a linear relation exists between response and explanatory variables included in the model. Finally, we checked

whether there exists multicollinearity. Based on variance inflation factor (VIF) we didn't observe any high values.

Of the 6 explanatory variables that were initially examined for this study 5 remained after LASSO was applied. Then, the variable of interest was added as an aditional variable to the model which in this case was a categorical variable with 2 states thus 1 indicator was added with the 2nd being considered as a reference.

$$\mu\{DRAT|...\} = \beta_0 + \beta_1 * AGEP + \beta_2 * WAGP + \beta_3 * SCHL + \beta_4 * PWGTP + \beta_5 * WKHP + \beta_6 * VPS2$$

Figure 15 presents the summary of results for fitting this linear model.

Inference

The first question of interest was veteran disability percentage, as shown in the linear regression model, the variable of interest, which is veteran period of service, is statistically significant (p-value = 0.01). veterans who served in Vietnam war and before are less likely to have higher percentage of disability rate than who served in the 1st Gulf war and beyond by (0.18). all other variables were statistically significant which mean we reject the null hypothesis that their coefficients are equal to zeros (p-value < 0.05).

Obstacles

Sampling: For the initial exploratory analysis it is hard to analyze the large sample of data since we are missing the underlying patterns. We then sampled the data, performed initial exploration, and based on the results we formulated the final model on which we concluded the results with the full data.

More Categorical Variables: Explanatory variables like SCHL, COW, VPS have many categories which made hard to include and understand in the context of the problem. We transformed the initial list of categories, grouped in to sub categories and defined as Indicator variables. This transformation approach is very helpful while performing the model selection using LASSO, since LASSO takes the input x as numerical matrix only.

Technical issue: Since the data is very big, we had different technical issues especially with R. R seems very sensitive to a large sample size. Initially we tried Tidyr package to transform into tibbles as it easy to explore the data but other possible solution could be using different powerful statistical software that can handle big data. However, it is out of scope for this class.

Discussion

Based on the questions we have suggested in our proposal, it seems the three questions were answered statistically in a meaningful way. 1) Is there any relationship between work travel time and wage for employed Oregonians? From this question we found that people who spend more time traveling to work, are more likely to have higher salary (statistically significant, P-value < 0.001). 2) Is there a relationship between worker class and hours of work per week? For this question we found that public employee usually works higher hours than other classifications (statistically significant). Finally, 3) Is there a relationship between veteran term of service and veteran disability percentage? For the last question, we found that veterans who participated in Vietnam war and before are less likely to have higher disability percentage than who participated in Gulf war and beyond (statistically significant, p-value = 0.01). In terms of future work, we could use more sophisticated models than multiple linear regression such as (binary logit model with random parameters or multinomial logit model), which can help us answering different important questions.

Additionally, considering the interaction terms (e.g. gender) in the model is another way of improving our conclusion.

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 hours 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	Verteran ended service after or in 1st gulf war indicator
VPS2	Veteran ended service in vietnam or before
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)
  }
}


cow_f <- function(x){
  if (cow1_f(x)){
    return (1)
  } else if(cow2_f(x)){
    return (2)
  } else if(cow3_f(x)){
    return (3)
  } else if(cow4_f(x)){
    return (4)
```

```
  } else {
    return (NA)
  }
}

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) -109583.9844
## AGEP            276.0056
## JWAP              .
## SCHL           3487.2810
## WKHP           1073.5359
## PWGTP             .
## COW1          10097.4024
## COW2           3291.9064
## COW3              .
## JWTR1             .
## JWTR2             .
## JWTR3             .
## JWTR4             .
## WKW1           -196.2987
## WKW2              .
## WKW3          12872.6460
## MART           8969.8814
## MILT              .
## SEXT           8490.9707
## INSUR          7948.5070
```

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)
```

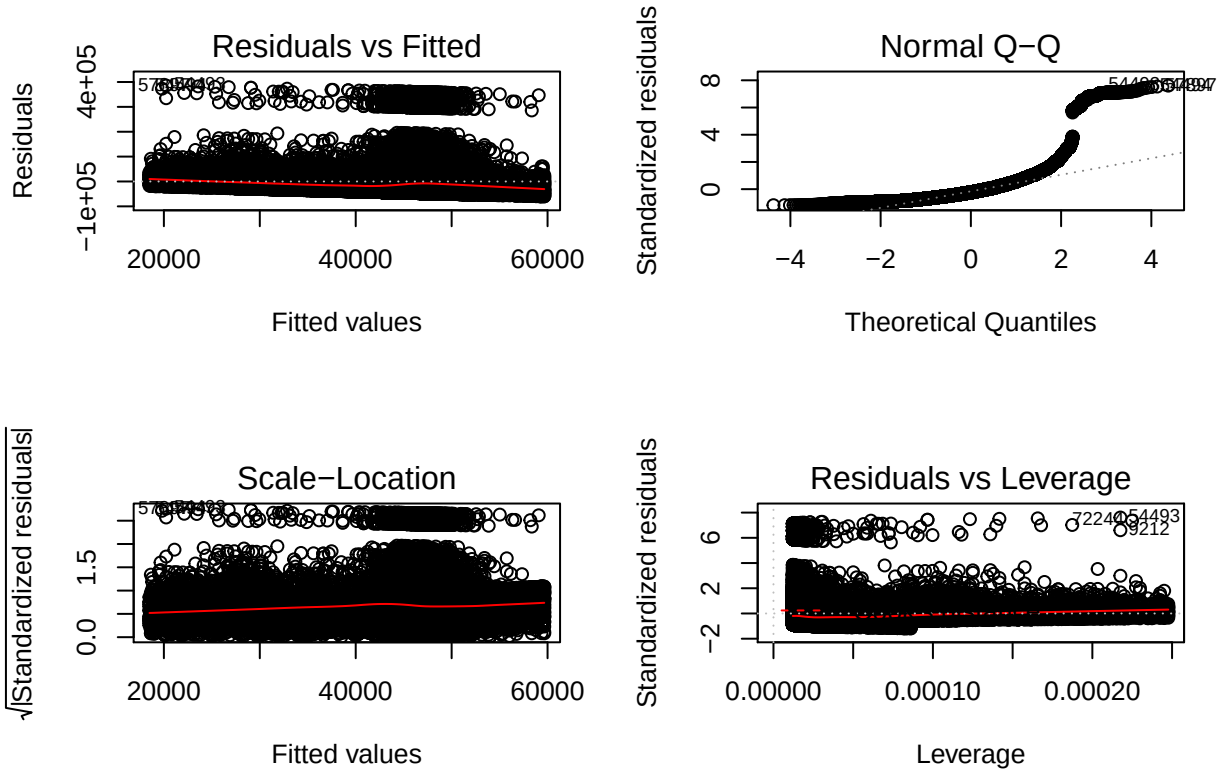

Figure 3: Take Log of WAGP

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

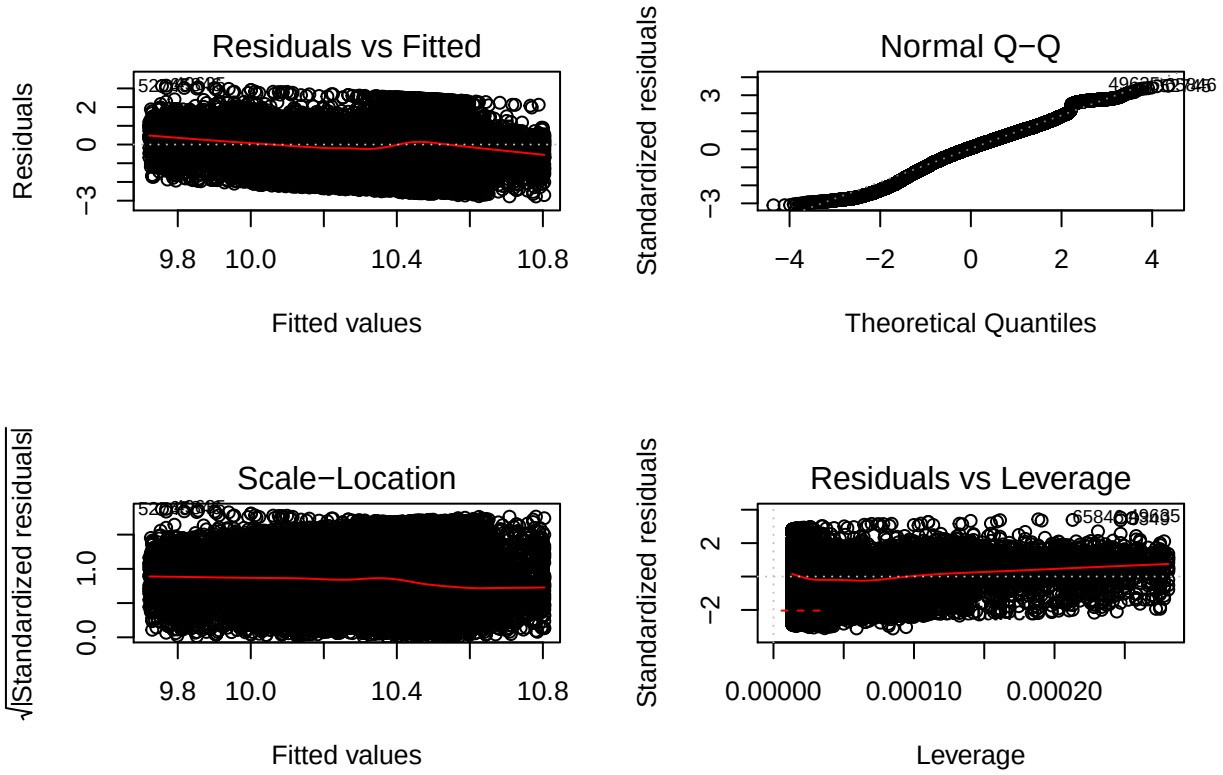

Figure 4: Check the Multicolinearity

```
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))
```

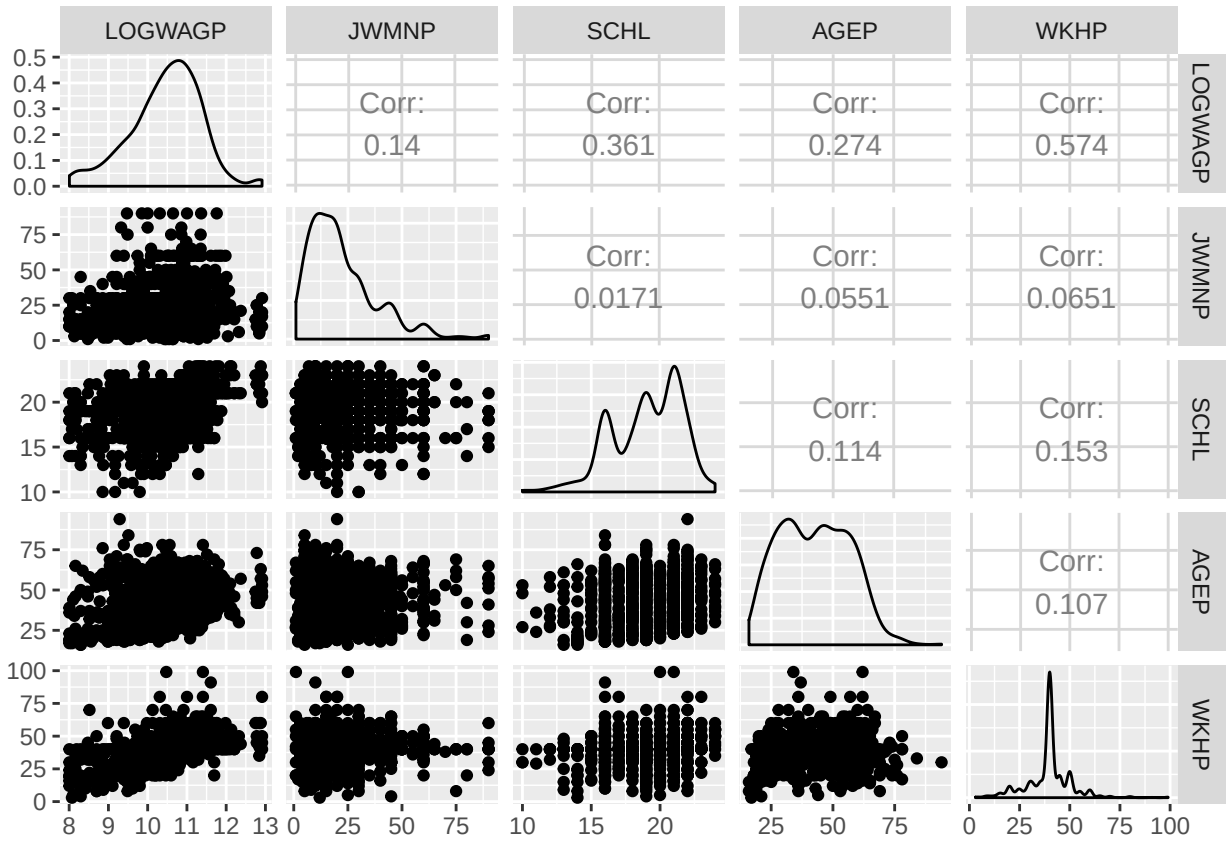

Figure 6: Summary of Linear Model

```
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
```

prep data for Question 2

```
pop_new_2 <- tbl_df(pop_new)
tracemem(pop_new) == tracemem(pop_new_2)

pop_new_2$COW1 <- NULL
pop_new_2$COW2 <- NULL
pop_new_2$COW3 <- NULL
pop_new_2$COW4 <- NULL

pop_new_2$JWTR2 <- NULL
pop_new_2$JWTR3 <- NULL
pop_new_2$JWTR4 <- NULL

pop_new_2$WKW1 <- NULL
pop_new_2$WKW2 <- NULL

pop_new_2$COW <- as.numeric(lapply(as.numeric(pop_data$COW),cow_f))

pop_omit_2 <- na.omit(pop_new_2)

y2 <- pop_omit_2$WKHP

pop_omit_2$WKHP <- NULL
pop_omit_2$COW1 <- NULL
pop_omit_2$COW2 <- NULL
pop_omit_2$COW3 <- NULL
pop_omit_2$COW4 <- NULL
pop_omit_2$COW <- NULL

X2 <- as.matrix(pop_omit_2)

lasso <- glmnet(X2, y2)
lasso.cv <- cv.glmnet(X2, y2)
```

Question 2 Figures

Figure 7: LASSO Results

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

```
## 13 x 1 sparse Matrix of class "dgCMatrix"
##                            1
## (Intercept)  2.894301e+01
## AGEP         .
```

```
## JWMNP         8.712340e-03
## WAGP          6.179366e-05
## JWAP         -3.297663e-02
## SCHL          .
## PWGTP         .
## JWTR1         5.878605e-01
## WKW3          8.305241e+00
## MART          3.577828e-01
## MILT          .
## SEXT          3.197083e+00
## INSUR         .
```

Figure 8: Check assumptions

```
pop_lm_data_2 <- na.omit(pop_new_2)
pop_lm_test <- lm(WKHP ~ as.factor(COW), data = filter(pop_lm_data_2, WKHP < 50, WKHP > 10, JWMNP < 120
par(mfrow = c(2, 2))
plot(pop_lm_test)
```

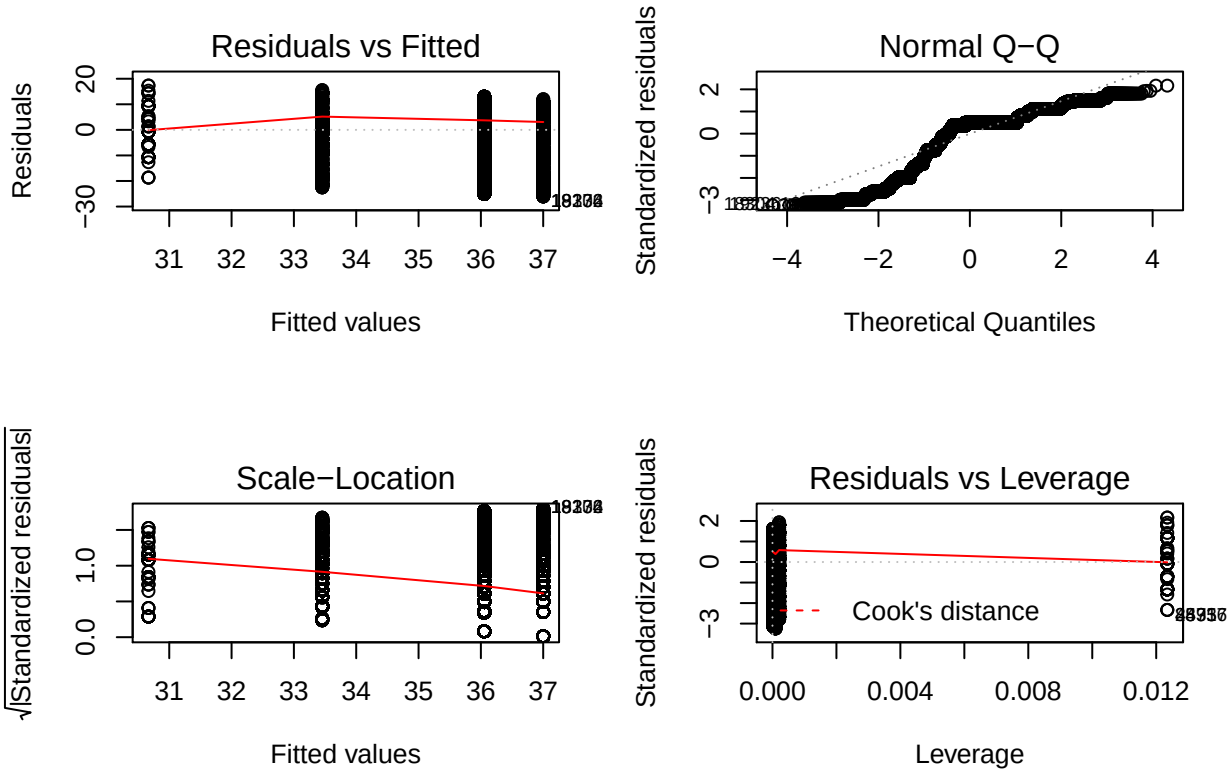

Figure 9: Check the Multicolinearity

```
pop_lm_2 <- lm(WKHP ~ as.factor(COW)+JWMNP+WAGP+JWAP+as.factor(JWTR1)+as.factor(WKW3)+as.factor(MART)+as
               data = filter(pop_lm_data_2, WKHP < 50, WKHP > 10, JWMNP < 120, WAGP < 200000))
```

```
vif(pop_lm_2)
```

```
##   as.factor(COW)2   as.factor(COW)3   as.factor(COW)4             JWMNP
##          1.039288          1.066069          1.001170          1.021208
##              WAGP              JWAP as.factor(JWTR1)1  as.factor(WKW3)1
##          1.260056          1.033742          1.031382          1.122473
##  as.factor(MART)1  as.factor(SEXT)1
##          1.105221          1.034097
```

Figure 10: Check linearity Assumptions

```
ggpairs(sample_n(select(filter(pop_lm_data_2, WKHP < 50, WKHP > 10, JWMNP < 120, WAGP < 200000), WKHP, W
```

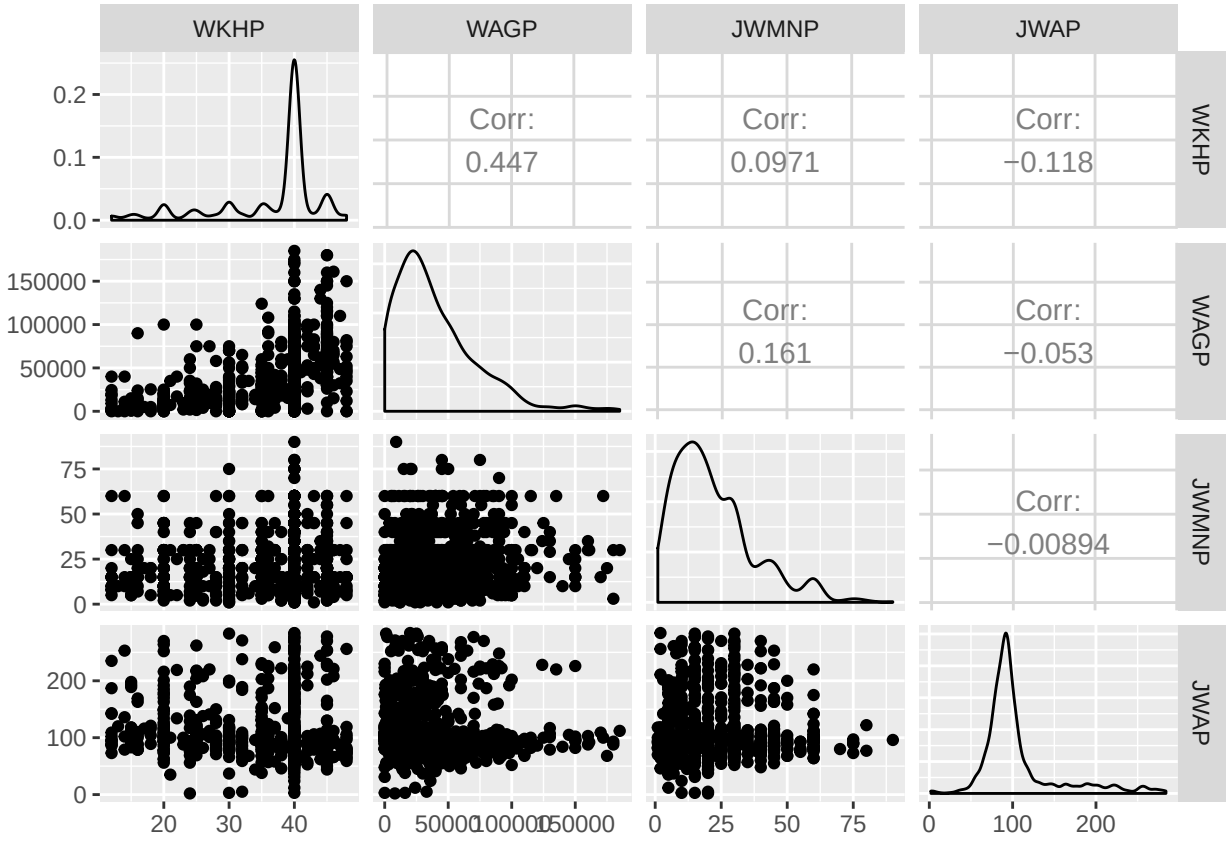

Figure 11: Summary of Linear Model

```
summary(pop_lm_2)
```

```
##
## Call:
## lm(formula = WKHP ~ as.factor(COW) + JWMNP + WAGP + JWAP + as.factor(JWTR1) +
##     as.factor(WKW3) + as.factor(MART) + as.factor(SEXT), data = filter(pop_lm_data_2,
##     WKHP < 50, WKHP > 10, JWMNP < 120, WAGP < 2e+05))
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -35.923  -3.297   1.607   4.453  24.881
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         2.834e+01  1.448e-01 195.619  < 2e-16 ***
## as.factor(COW)2     2.758e-01  7.519e-02   3.668 0.000244 ***
## as.factor(COW)3    -1.065e+00  1.072e-01  -9.933  < 2e-16 ***
## as.factor(COW)4    -2.681e+00  7.705e-01  -3.479 0.000503 ***
## JWMNP               2.055e-02  1.798e-03  11.426  < 2e-16 ***
## WAGP                8.963e-05  9.913e-07  90.415  < 2e-16 ***
## JWAP               -2.172e-02  6.667e-04 -32.584  < 2e-16 ***
## as.factor(JWTR1)1   9.882e-01  8.585e-02  11.512  < 2e-16 ***
## as.factor(WKW3)1    4.917e+00  8.944e-02  54.971  < 2e-16 ***
## as.factor(MART)1   -5.666e-03  5.769e-02  -0.098 0.921754
## as.factor(SEXT)1    1.911e+00  5.559e-02  34.371  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.926 on 64222 degrees of freedom
## Multiple R-squared:  0.2677, Adjusted R-squared:  0.2676
## F-statistic:  2348 on 10 and 64222 DF,  p-value: < 2.2e-16
```

Prep for question 3

```
pop_vet <- select(pop_data, AGEP, JWMNP)

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

pop_vet$AGEP <- as.numeric(pop_data$AGEP)

pop_vet$PWGTP <- as.numeric(pop_data$PWGTP)
pop_vet$SCHL <- as.numeric(pop_data$SCHL)

#pop_vet$DIS <- as.numeric(pop_data$DIS)

pop_vet$DRAT <- as.numeric(pop_data$DRAT)
pop_vet$VPS <- as.factor(pop_data$VPS)

na0 <- function(x){
  if (is.na(x)){
    return (0)
  } else {
    return (x)
  }
}

pop_vet$JWMNP <- as.numeric(lapply(as.numeric(pop_data$JWMNP), na0))
pop_vet$WKHP <- as.numeric(lapply(as.numeric(pop_data$WKHP), na0))

vet_omit <- na.omit(pop_vet)
```

```r
y3 <- vet_omit$DRAT

vet_omit$DRAT <- NULL
vet_omit$VPS <- NULL

# Gulf war erra
vps1_f <- function(x){
  if(!is.na(x) && (x <= 5)){
    return(1)
  } else {
    return(0)
  }
}
# WWII - Vietnam
vps2_f <- function(x){
  if(!is.na(x) && (x > 5 && x < 15 )){
    return(1)
  } else {
    return(0)
  }
}

vps_f <- function(x){
  if (vps1_f(x)){
    return (1)
  } else if(vps2_f(x)){
    return (2)
  } else {
    return (NA)
  }
}

pop_vet$VPS1 <- as.numeric(lapply(as.numeric(pop_data$VPS), vps1_f))
pop_vet$VPS2 <- as.numeric(lapply(as.numeric(pop_data$VPS), vps2_f))
pop_vet$VPS <- as.numeric(lapply(as.numeric(pop_data$VPS), vps_f))

vet_lm_data <- na.omit(pop_vet)

X3 <- as.matrix(vet_omit)

lasso <- glmnet(X3, y3)
lasso.cv <- cv.glmnet(X3, y3)
```

Figures for question 3

Figure 12: LASSO Results

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

```
## 7 x 1 sparse Matrix of class "dgCMatrix"
##                            1
## (Intercept)  4.282953e+00
```

```
## AGEP         -7.215658e-03
## JWMNP         .
## WAGP         -2.144904e-06
## PWGTP         1.240486e-03
## SCHL         -8.682399e-03
## WKHP         -1.111386e-02
```

Figure 13: Check Assumptions

```
vet_lm <- lm(DRAT ~ as.factor(VPS) + AGEP + WKHP + SCHL + PWGTP + WAGP, data = vet_lm_data)
par(mfrow = c(2, 2))
plot(vet_lm)
```

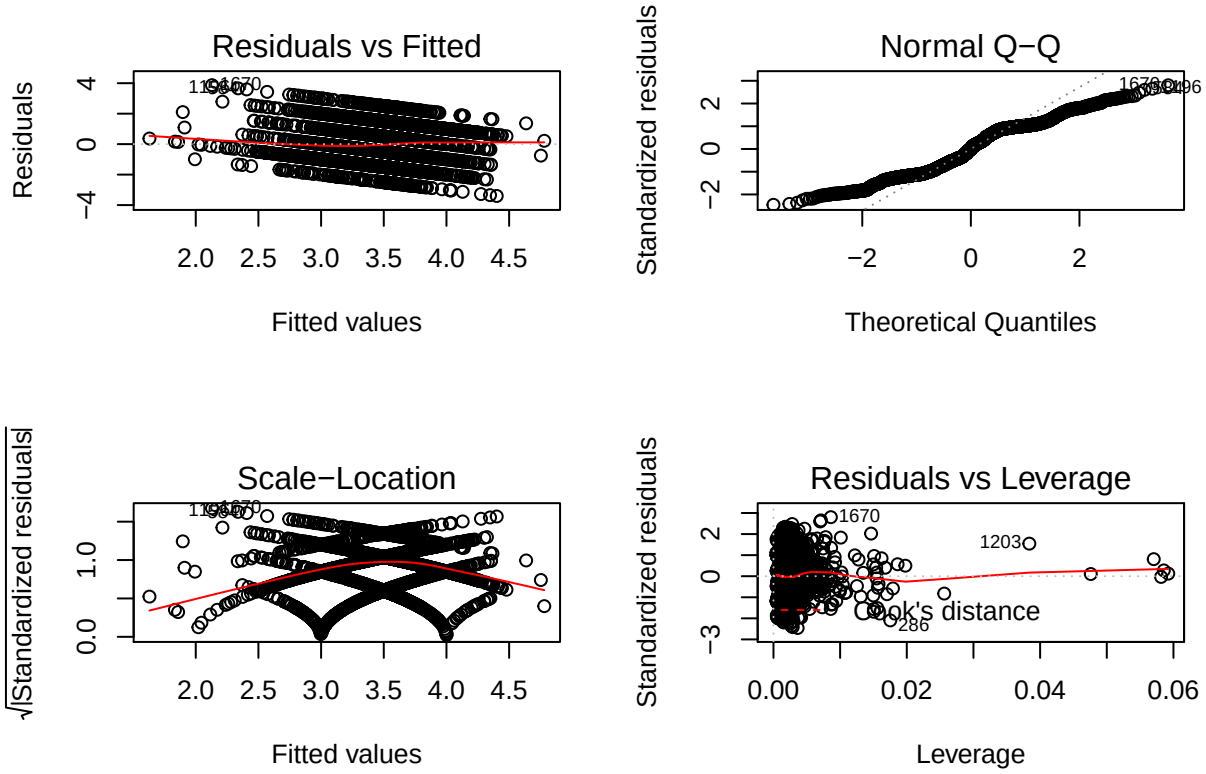

Figure 14: Check the Multicolinearity

```
vif(vet_lm)
```

```
## as.factor(VPS)2           AGEP           WKHP           SCHL
##        2.325580       2.450302       2.268074       1.057271
##          PWGTP           WAGP
##        1.022790       1.939351
```

Figure 15: Summary of Linear Model

```
summary(vet_lm)
```

```
##
## Call:
## lm(formula = DRAT ~ as.factor(VPS) + AGEP + WKHP + SCHL + PWGTP +
##     WAGP, data = vet_lm_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4009 -1.2664  0.0095  1.2716  3.8704
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      5.101e+00  2.218e-01  23.000  < 2e-16 ***
## as.factor(VPS)2 -1.877e-01  7.394e-02  -2.538 0.011181 *
## AGEP            -1.041e-02  2.274e-03  -4.578 4.86e-06 ***
## WKHP            -1.586e-02  1.674e-03  -9.475  < 2e-16 ***
## SCHL            -3.446e-02  9.607e-03  -3.587 0.000339 ***
## PWGTP            5.019e-03  1.766e-03   2.842 0.004503 **
## WAGP            -3.030e-06  9.629e-07  -3.147 0.001664 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.387 on 3603 degrees of freedom
## Multiple R-squared:  0.06786,    Adjusted R-squared:  0.06631
## F-statistic: 43.72 on 6 and 3603 DF,  p-value: < 2.2e-16
```