Untitled

GOVT\_650 FINAL

12/6/2020

library(tidyverse)

## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.2 v purrr 0.3.4  
## v tibble 3.0.1 v dplyr 1.0.0  
## v tidyr 1.1.2 v stringr 1.4.0  
## v readr 1.4.0 v forcats 0.5.0

## -- Conflicts ------------------------------------------ tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

library(ggplot2)  
library(readr)  
library(dplyr)  
library(reshape2)

##   
## Attaching package: 'reshape2'

## The following object is masked from 'package:tidyr':  
##   
## smiths

#Ques1: Load the data

finaldata<-read.csv("yu2017sample.csv", header=T)  
head(finaldata)

## X PUBID year weight selfemp wage age numChildren educ school  
## 1 65 15 2002 225405 0 750 20 0 2.High school TRUE  
## 2 66 15 2004 225405 0 900 22 0 2.High school TRUE  
## 3 67 15 2005 225405 0 700 22 0 2.High school TRUE  
## 4 68 15 2006 225405 0 400 23 0 2.High school TRUE  
## 5 69 15 2009 225405 0 1200 27 0 2.High school FALSE  
## 6 70 15 2010 225405 0 947 28 0 2.High school FALSE  
## experience tenure tenure2 fullTime multipleLocations unionized  
## 1 14 0.26923077 0.072485207 FALSE NA NA  
## 2 35 0.09615385 0.009245562 FALSE NA NA  
## 3 61 0.23076923 0.053254438 TRUE 1 0  
## 4 114 0.57692308 0.332840237 TRUE 1 0  
## 5 282 1.75000000 3.062500000 FALSE 1 0  
## 6 336 2.78846154 7.775517751 FALSE 1 0  
## firmSize marstat region urban  
## 1 <NA> No romantic union 1 1  
## 2 <NA> No romantic union 4 0  
## 3 1. Less than 30 No romantic union 4 0  
## 4 1. Less than 30 No romantic union 4 1  
## 5 <NA> No romantic union 1 1  
## 6 3. 300+ No romantic union 1 1  
## industry occupation  
## 1 Health Care 3640  
## 2 Health Care 5400  
## 3 Arts, Entertainment, Recreation, Accommodations, and Food Services 4110  
## 4 Arts, Entertainment, Recreation, Accommodations, and Food Services 4040  
## 5 Retail Trade 4760  
## 6 Retail Trade 4760  
## autonomy competitiveness hazardous regularity teamwork  
## 1 3.759670 2.406207 2.218415 1.072356 4.494050  
## 2 4.013030 3.231223 1.017998 1.173900 4.447300  
## 3 3.619571 2.606161 1.234400 1.327400 4.477848  
## 4 4.139598 2.701270 1.197880 1.593700 4.381400  
## 5 4.011500 3.585000 1.257400 1.196200 4.292729  
## 6 4.011500 3.585000 1.257400 1.196200 4.292729

#Total Years included in the dataset =16

levels(factor(finaldata$year))

## [1] "1997" "1998" "1999" "2000" "2001" "2002" "2003" "2004" "2005" "2006"  
## [11] "2007" "2008" "2009" "2010" "2011" "2013"

finaldata%>%   
distinct(year) %>%   
count()

## n  
## 1 16

#Number of Women in the dataset = 1569

length(unique(finaldata[,"PUBID"]))

## [1] 1569

finaldata%>%   
distinct(PUBID) %>%   
count()

## n  
## 1 1569

#person-years in the dataset #Highest number (1293) for year 2002 #Lowest number (374) for year 1997

finaldata%>%   
 group\_by(year)%>%  
 summarise(n=n\_distinct(PUBID))

## `summarise()` ungrouping output (override with `.groups` argument)

## Warning: `...` is not empty.  
##   
## We detected these problematic arguments:  
## \* `needs\_dots`  
##   
## These dots only exist to allow future extensions and should be empty.  
## Did you misspecify an argument?

## # A tibble: 16 x 2  
## year n  
## <int> <int>  
## 1 1997 374  
## 2 1998 803  
## 3 1999 1042  
## 4 2000 1189  
## 5 2001 1239  
## 6 2002 1293  
## 7 2003 1267  
## 8 2004 1257  
## 9 2005 1259  
## 10 2006 1284  
## 11 2007 1267  
## 12 2008 1263  
## 13 2009 1231  
## 14 2010 1173  
## 15 2011 1157  
## 16 2013 1184

#Ques2: #Importance of variable transformation: #Transformation can normalize than before by removing skewness #It is mostly done to improve (removing Heteroscedasticity) #and have more valid results (overall better model accuracy) #for unbiased inferences.

finaldata%>%   
 mutate(logwage=(log(wage))) -> finaldata1

#Correlation Matrix for (Continous Variables): #logwage and numChildren: 0.049088172 (Extremely Low) #logwage and experience: 0.499130160 (Moderately +ve) #logwage and competitiveness: 0.201258731 (Low but +ve) #logwage and autonomy: 0.390023923 (Moderately +ve)

finaldata1%>%  
 select(-educ,-school,-fullTime,-firmSize,-marstat,-industry) ->cordata  
  
cor(na.omit(cordata))

## Warning in cor(na.omit(cordata)): the standard deviation is zero

## X PUBID year weight selfemp  
## X 1.000000000 0.999869016 0.007684057 -0.40778843 NA  
## PUBID 0.999869016 1.000000000 0.007148887 -0.40569350 NA  
## year 0.007684057 0.007148887 1.000000000 -0.02140492 NA  
## weight -0.407788429 -0.405693497 -0.021404919 1.00000000 NA  
## selfemp NA NA NA NA 1  
## wage -0.038995391 -0.039433961 0.122242732 0.02931890 NA  
## age 0.023520394 0.023252048 0.934830173 -0.03554663 NA  
## numChildren 0.134402947 0.134471927 0.342940309 -0.22143645 NA  
## experience -0.032533045 -0.031754377 0.808097046 0.09312644 NA  
## tenure -0.029315245 -0.028816002 0.461924613 0.01980893 NA  
## tenure2 -0.023971037 -0.023404541 0.393514556 0.02043617 NA  
## multipleLocations -0.018022159 -0.017843712 0.034680618 -0.03505527 NA  
## unionized 0.026597731 0.025731913 0.080390414 -0.07270275 NA  
## region 0.210597165 0.209428130 0.037553696 -0.16106160 NA  
## urban -0.042701988 -0.043676092 0.029625224 -0.17937722 NA  
## occupation 0.073675919 0.073714839 -0.209742608 -0.08204672 NA  
## autonomy -0.071320247 -0.071209421 0.280702609 0.10884523 NA  
## competitiveness -0.099807676 -0.099744273 0.079893866 0.07941536 NA  
## hazardous 0.022871691 0.023248037 0.051785385 0.01642760 NA  
## regularity -0.003640882 -0.003620622 -0.079165563 0.05287380 NA  
## teamwork -0.039798105 -0.039934233 0.059000039 0.01454220 NA  
## logwage -0.086968521 -0.087829490 0.483870632 0.04273824 NA  
## wage age numChildren experience tenure  
## X -0.03899539 0.02352039 0.134402947 -0.032533045 -0.02931525  
## PUBID -0.03943396 0.02325205 0.134471927 -0.031754377 -0.02881600  
## year 0.12224273 0.93483017 0.342940309 0.808097046 0.46192461  
## weight 0.02931890 -0.03554663 -0.221436447 0.093126443 0.01980893  
## selfemp NA NA NA NA NA  
## wage 1.00000000 0.12884105 0.017555049 0.129042308 0.07797608  
## age 0.12884105 1.00000000 0.379429581 0.850350465 0.47750475  
## numChildren 0.01755505 0.37942958 1.000000000 0.229699803 0.13737355  
## experience 0.12904231 0.85035046 0.229699803 1.000000000 0.56669016  
## tenure 0.07797608 0.47750475 0.137373550 0.566690160 1.00000000  
## tenure2 0.05917252 0.41310982 0.124964549 0.494118439 0.92610270  
## multipleLocations 0.01661552 0.04001192 0.007935002 0.050766786 0.03713218  
## unionized 0.03805894 0.07889220 0.044360103 0.066955613 0.06737355  
## region -0.02134581 0.03151713 0.065533049 -0.028666320 -0.01783364  
## urban 0.01171065 0.02447524 -0.011111091 -0.003491223 -0.03333502  
## occupation -0.05616367 -0.19986060 0.036277922 -0.205540183 -0.09952954  
## autonomy 0.11428469 0.28867165 -0.018341341 0.306859752 0.13495585  
## competitiveness 0.07041761 0.08144980 -0.072056762 0.087149706 0.04792152  
## hazardous 0.03591287 0.06385309 0.056624427 0.046840178 0.07187600  
## regularity -0.03016119 -0.08470554 -0.020353164 -0.100542464 -0.03346072  
## teamwork -0.02576260 0.05825150 -0.012332758 0.087525835 0.02612910  
## logwage 0.41587192 0.48969378 0.049088172 0.499130160 0.32069062  
## tenure2 multipleLocations unionized region  
## X -0.02397104 -0.018022159 0.02659773 0.2105971647  
## PUBID -0.02340454 -0.017843712 0.02573191 0.2094281303  
## year 0.39351456 0.034680618 0.08039041 0.0375536959  
## weight 0.02043617 -0.035055275 -0.07270275 -0.1610615981  
## selfemp NA NA NA NA  
## wage 0.05917252 0.016615521 0.03805894 -0.0213458133  
## age 0.41310982 0.040011916 0.07889220 0.0315171274  
## numChildren 0.12496455 0.007935002 0.04436010 0.0655330488  
## experience 0.49411844 0.050766786 0.06695561 -0.0286663203  
## tenure 0.92610270 0.037132177 0.06737355 -0.0178336360  
## tenure2 1.00000000 0.026273444 0.04856500 -0.0192984162  
## multipleLocations 0.02627344 1.000000000 0.05232560 -0.0032171596  
## unionized 0.04856500 0.052325596 1.00000000 -0.0234014988  
## region -0.01929842 -0.003217160 -0.02340150 1.0000000000  
## urban -0.04000113 0.064229861 0.01891156 0.0405207619  
## occupation -0.07509926 -0.003576230 -0.02082666 0.0003864585  
## autonomy 0.09753537 0.066255145 0.04139885 0.0057531299  
## competitiveness 0.03973444 0.117189663 -0.06750630 -0.0171432323  
## hazardous 0.07015111 -0.015845143 0.02694122 -0.0454990351  
## regularity -0.02097535 0.011955542 -0.08212747 -0.0419699159  
## teamwork 0.00727416 0.013006519 0.05644289 0.0146473612  
## logwage 0.25384266 0.086965740 0.15999622 -0.0037230058  
## urban occupation autonomy competitiveness  
## X -0.042701988 0.0736759195 -0.07132025 -0.099807676  
## PUBID -0.043676092 0.0737148390 -0.07120942 -0.099744273  
## year 0.029625224 -0.2097426076 0.28070261 0.079893866  
## weight -0.179377218 -0.0820467203 0.10884523 0.079415360  
## selfemp NA NA NA NA  
## wage 0.011710652 -0.0561636706 0.11428469 0.070417612  
## age 0.024475238 -0.1998606044 0.28867165 0.081449802  
## numChildren -0.011111091 0.0362779218 -0.01834134 -0.072056762  
## experience -0.003491223 -0.2055401831 0.30685975 0.087149706  
## tenure -0.033335024 -0.0995295388 0.13495585 0.047921517  
## tenure2 -0.040001132 -0.0750992561 0.09753537 0.039734436  
## multipleLocations 0.064229861 -0.0035762302 0.06625515 0.117189663  
## unionized 0.018911562 -0.0208266628 0.04139885 -0.067506300  
## region 0.040520762 0.0003864585 0.00575313 -0.017143232  
## urban 1.000000000 -0.0502828393 0.04302378 0.028652501  
## occupation -0.050282839 1.0000000000 -0.35127582 -0.267035840  
## autonomy 0.043023783 -0.3512758217 1.00000000 0.412749361  
## competitiveness 0.028652501 -0.2670358403 0.41274936 1.000000000  
## hazardous -0.052987933 0.1116800658 0.05942516 0.049747397  
## regularity -0.007676564 0.0023227967 -0.09224524 0.230246933  
## teamwork 0.012838746 -0.2009668810 0.25352352 0.005365306  
## logwage 0.058467976 -0.2385737759 0.39002392 0.201258731  
## hazardous regularity teamwork logwage  
## X 0.02287169 -0.003640882 -0.039798105 -0.086968521  
## PUBID 0.02324804 -0.003620622 -0.039934233 -0.087829490  
## year 0.05178539 -0.079165563 0.059000039 0.483870632  
## weight 0.01642760 0.052873803 0.014542198 0.042738235  
## selfemp NA NA NA NA  
## wage 0.03591287 -0.030161189 -0.025762599 0.415871923  
## age 0.06385309 -0.084705543 0.058251501 0.489693779  
## numChildren 0.05662443 -0.020353164 -0.012332758 0.049088172  
## experience 0.04684018 -0.100542464 0.087525835 0.499130160  
## tenure 0.07187600 -0.033460720 0.026129100 0.320690619  
## tenure2 0.07015111 -0.020975347 0.007274160 0.253842660  
## multipleLocations -0.01584514 0.011955542 0.013006519 0.086965740  
## unionized 0.02694122 -0.082127471 0.056442893 0.159996223  
## region -0.04549904 -0.041969916 0.014647361 -0.003723006  
## urban -0.05298793 -0.007676564 0.012838746 0.058467976  
## occupation 0.11168007 0.002322797 -0.200966881 -0.238573776  
## autonomy 0.05942516 -0.092245236 0.253523516 0.390023923  
## competitiveness 0.04974740 0.230246933 0.005365306 0.201258731  
## hazardous 1.00000000 0.239719251 -0.085979216 0.046546598  
## regularity 0.23971925 1.000000000 -0.151835074 -0.175812725  
## teamwork -0.08597922 -0.151835074 1.000000000 0.020327716  
## logwage 0.04654660 -0.175812725 0.020327716 1.000000000

#Another method to find correlation #We can also generate scaterrplot for this anlaysis

naomitfinaldata1<- na.omit(finaldata1)  
  
cor(naomitfinaldata1$logwage, naomitfinaldata1$numChildren)

## [1] 0.05322835

cor(naomitfinaldata1$logwage, naomitfinaldata1$experience)

## [1] 0.5008612

cor(naomitfinaldata1$logwage, naomitfinaldata1$competitiveness)

## [1] 0.2036232

cor(naomitfinaldata1$logwage, naomitfinaldata1$autonomy)

## [1] 0.3916341

#Ques3: #Regression model #(Even though the data consits of missing values, it might be #a good idea to retain the values for model generation) #Estimated coefficient(beta “numChildren”=-0.0748283,pvalue=< 2e-16 #Interpretation:For every unit increase in variable“numChildren”, #there will be variation (avergae decrease)in the response #variable “logwage” by 0.0748283 (holding others fixed).

regfinal1<- lm(logwage ~ numChildren+factor(year)+factor(PUBID), data=finaldata1)

#Using SUMMARY(regfinal1) #Adjusted R^2 = 0.4877 or 48.77% approx #Generally, the higher the better #RMSE(measures the model prediction error) #0.4206701 or 42.06% approx #Generally, the lower the better

res.1<- resid(regfinal1)  
sqrt(mean(res.1^2))

## [1] 0.4206701

#Ques4: #(variable “numChildren”) Standard Error = 0.0065134 #coefficient estimate divided by standard error = t statistic #(variable “numChildren”) t-value = -11.488 #A t-test is performed to check whether or not the #beta coefficients are significantly different from zero. #A non-zero beta coefficients means that there is a significant #relationship between the predictor “numChildren” and “logwage”. #T value is looked at with the p-value (Here,p-value=< 2e-16) #i.e. Approx(2e-16) < 0.05.

#Rejecting the Null in this exercise: #The alternate hypothesis that the coefficients are not #equal to zero (i.e. there exists a relationship between #numChildren and logwage).

#The level of alpha is mostly set at (convential threshold=5%) #Null Hypothesis will be rejected when p value < 0.05 and #the result is statistically significant. #But we also care about substantive significance. #However, when pvalue is 0.0499, it might be inconclusive #(Another case when p-value is 0.0512).

#(Generally)Any p-value >0.05 can be excluded from the model. #Based on the significance level,other variables can be added #for better prediction accuracy.

#However,the model in question has factor variables(year,PUBID). #A power analysis can also be done to find the relationship #rather than just relying on the p values.

#Ques5: #Variable “numChildren” #Estimated Coefficient: -0.0235627 #Standard Error: 0.0079786 #t statistic: -2.953, P-value: 0.003153 #(<0.05, statitically significant) #For every unit increase in variable“numChildren”, #there will be variation (avergae decrease)in the response #variable “logwage” by 2.356% (holding other predictors fixed).

regfinal2<- lm(logwage ~ numChildren+factor(year)+factor(PUBID)+factor(region)+  
 factor(urban)+factor(marstat)+factor(educ)+factor(school)+  
 experience+tenure+tenure2+factor(fullTime)+factor(firmSize)+  
 multipleLocations+factor(unionized)+  
 factor(industry), data=finaldata1)

#Using SUMMARY(regfinal2) #Adjusted R^2 = 0.6345 #RMSE(measures the model prediction error)= 0.3228371 #We see a higher (better) value of Adjusted R^2 than our 1st model #RMSE is lower than the 1st model #Overall, this model has better accuracy than the first

res.2<- resid(regfinal2)  
sqrt(mean(res.2^2))

## [1] 0.3228371

#Coefficients of two other variables (statistically significant) #factor(year)1999 = 0.082110001, p-value = 0.006630 #factor(year)2000 = 0.110267617, p-value = 0.000357

#For every unit increase in variable “(year)1999”, #there will be variation (avergae increase)in the response #variable “logwage” by 8.211% (holding other predictors fixed).

#For every unit increase in variable “(year)2000”, #there will be variation (avergae increase)in the response #variable “logwage” by 11.02% (holding other predictors fixed).

#Ques6: #Adding an interaction term, can change the #model coefficients significantly #It adds a combined efffect of variables to the model.

regfinal3<- lm(logwage ~ numChildren+factor(year)+factor(PUBID)+factor(region)+  
 factor(urban)+factor(marstat)+factor(educ)+factor(school)+  
 experience+tenure+tenure2+factor(fullTime)+factor(firmSize)+  
 multipleLocations+factor(unionized)+  
 factor(industry)+I(numChildren\*regularity)+regularity+  
 I(numChildren\*hazardous)+hazardous, data=finaldata1)

#Using SUMMARY(regfinal3) #Adjusted R^2 = 0.6469 #RMSE(measures the model prediction error)= 0.3141082 #We see a higher (better) value of Adjusted R^2 than our 1st and 2nd models #RMSE is lower than the other models #Overall, this model has better accuracy than the previous models #However, the overall accuracy is only slightly better than the 2nd model

res.3<- resid(regfinal3)  
sqrt(mean(res.3^2))

## [1] 0.3141082

#Coefficients #numChildren=0.0241048 S.E=0.0340750 p-value=0.479333 #regularity = S.E.= p-value = #I(numChindren*regularity)= S.E.= p-value = #hazardous = S.E.= p-value = #I(numChindren*hazardous)= S.E.= p-value =

#The pvalue of the numChilren variable > 0.05, #therefore, we fail to reject Null Hypothesis #(i.e. there exists no relationship between #numChildren and logwage).