CS518 - Final Project Report

Nora Quick

Executive Summary

This report covers the a cencus dataset about adult income information. There are two main questions to disect thoughout this report. Firstly, do either age or hours worked increase the likelyhood of earning more than $50K a year and if so which one and what is most significant age/hours worked? Secondly, what are the causes, if any, of divorce?

Focusing on age and hours worked in relation to income it was found that there are significant correlation between age and income as well as hours worked and income. It was found that of the people who make make less than 50K a year are on average younger than 37 years old while the people who make more than 50K a year are on average older than 37 years old. In addition to this, of the people who make less than 50K a year the younger group works less than 34.4 hours a week while of the people who make more than 50K a year the older group work less than 34.4 hours a week.

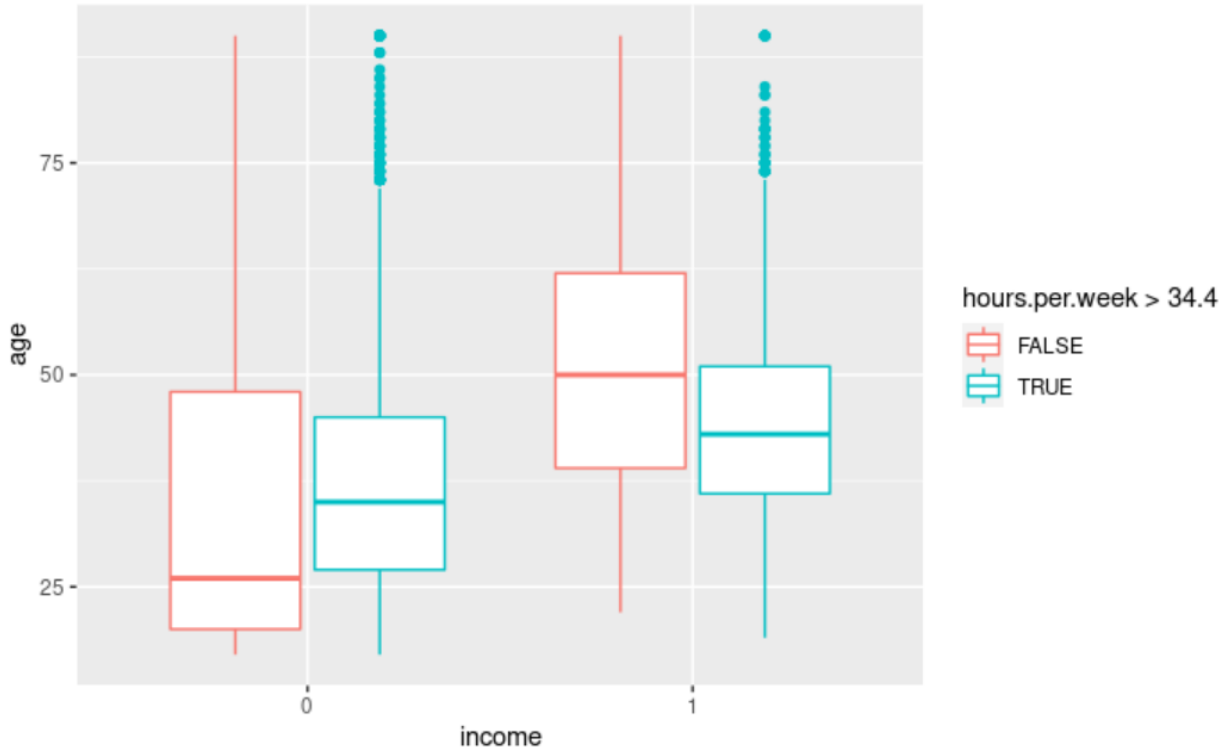
Focusing on only the data that seems most relatable to marrital status it was found that there were quite a few variables that were significant to divorce. Age, work class (job), some education level, race, hours worked and income are all significant factors of divorce.

Body

Dataset Desciption & Questions

For question one I deviated on the methodology I was originally going to use. To determine if either age or hours worked a week were significant to income I decided to use a binomial logit method. For this I made income greater than 50K "1" and income less than 50K "0". This methodology showed that both age and hours worked were significant with both showing p value of <2e-16.

Additionally below there is a graph. On this graph I have the x-axis as income and the y-axis as age. Additionally there is color indicating if the person works more or less than 34.4 hours a week. First of all we can see that there is quite a large age discrepensy between the two incomes. Secondly, the color shows that whole hours worked is important to income it is not a clear indication like age. The number 34.4 comes from the average hours worked a week in the US.



One concern I have from this model is overfitting. There is quite a high residual deviance and degrees of freedom.

For question two I deviated fron my original model again. Instead of a mixed model I chose to do a poisson distribution instead. Additionally, instead of using all variables I chose to look at the ones that made the most sense for marital status. These variables are age, work-class, education, race, and income. Many of the work class have significant p-values as seen below.

workclassFederal-gov 4.92e-06

workclassLocal-gov 3.75e-11

workclassNever-worked 0.366148

workclassPrivate 1.99e-11

workclassSelf-emp-inc 0.000762

workclassSelf-emp-not-inc 1.88e-10

workclassState-gov 4.31e-10

workclassWithout-pay 0.911512

The model also shows some significance for education level but shockingly it was not all lower level education as I was expecting. There was only one race that showed significance with a significant p-value of 0.000705. This shocked me as I didn't expect only one race to be significant and so much more than others. Hours worked is significant with a p-value of $< 2e\text{-}16$ which was expected along with income which has a significant p-value of $< 2e\text{-}16$.

Again I have a concern with this model as the residual deviance and degree of freedome is so large as well as the AIC. However, the fisher score is reasonable and only some variables, variabes that make sense, are significant so I feel comfortable concluding from this mode.

Conclusions/Discussion

Focusing on age and hours worked in relation to income it was found that there are significant correlation between age and income as well as hours worked and income. It was found that of the people who make make less than 50K a year are on average younger than 37 years old while the people who make more than 50K a year are on average older than 37 years old. In addition to this, of the people who make less than 50K a year the younger group works less than 34.4 hours a week while of the people who make more than 50K a year the older group work less than 34.4 hours a week.

This makes a lot of sense. The older we get the more work experience we have and we usually move up in the job market. As we increase our positions and work experience companies are more willing to pay people more. In addition younger people work less hours due to school and child labor laws and older people work less hours due to closer to retirement or a position that allows for it.

Focusing on only the data that seems most relatable to marrital status it was found that there were quite a few variables that were significant to divorce. Age, work class (job), some education level, race, hours per week, and income are all significant factors of divorce.

Education level often determines the work that someone has. The lower level education typically means lower level paying jobs. Government jobs, self employement, and no employement are all jobs that indicated in divorce. Lots of these are independent in nature which could be correlated to marrital status as well. Finally, an income of greater than 50K is correlated to divorce. Money can complicate things and it can also make someone able to support themself more comfortably. Both of those things could be correlated to why someone is/can be divorced. Age makes some sense because some people marry young for one reason or another and realize later that they need to divorce for one reason or another. Finally, hours per week makes a lot of sense. If someone works too much they may alienate themself from their family and, therefore, end up divorcing.

Appendices

Techincal Appendice

First Question:

```
bin_income <- cencus
bin_income$income <- ifelse(bin_income$income=="<=50K", 0, 1)

#bin_income

q1_mod1 <- glm(income ~ age + hours.per.week, family = binomial(link = "logit"), data = bin_income)
summary(q1_mod1)
```

```
##
## Call:
## glm(formula = income ~ age + hours.per.week, family = binomial(link = "logit"),
##     data = bin_income)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7312  -0.7523  -0.5552  -0.2416   2.6339
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -4.913396   0.074406  -66.03   <2e-16 ***
```

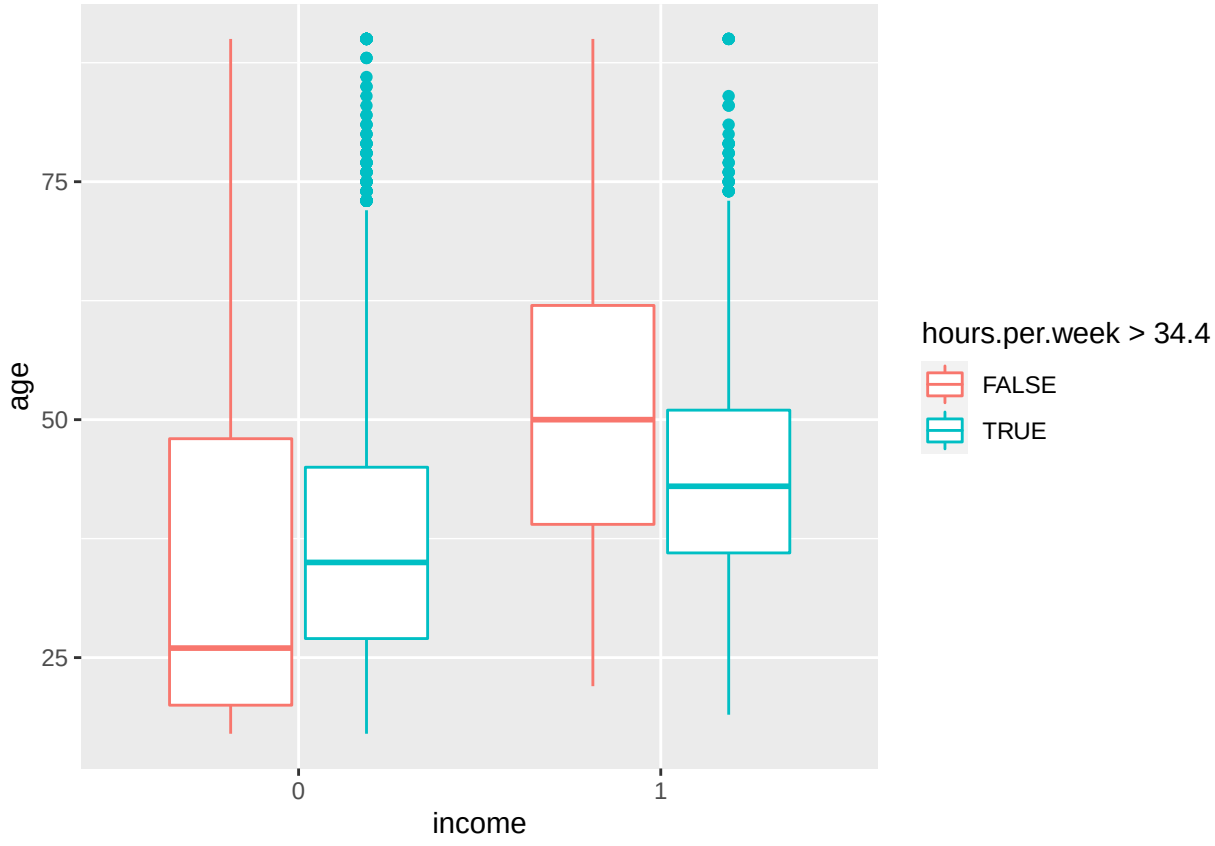
```
## age             0.043146   0.001050   41.09   <2e-16 ***
## hours.per.week  0.047836   0.001212   39.48   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 35948  on 32560  degrees of freedom
## Residual deviance: 32421  on 32558  degrees of freedom
## AIC: 32427
##
## Number of Fisher Scoring iterations: 4
```

```r
#q1_mod2 <- glm(income ~ age * hours.per.week, family = binomial(link = "logit"), data = bin_income)
#summary(q1_mod2)

#rand_bin_income <- bin_income

#q1_mod3 <- glm(income ~ age + hours.per.week, family = binomial(link = "logit"), data = rand_bin_incom
#summary(q1_mod3)

p <- ggplot(bin_income) +
  geom_boxplot(aes(x=factor(income), y=age, col=hours.per.week>34.4)) +
  xlab('income') + ylab('age')
p
```



Second Question:

```
num_marriage <- transform(cencus,id=as.numeric(factor(marital.status)))
div_df <- subset(num_marriage, (id == 1))

q2_mod1 <- glm(id ~ age + workclass + education + race + hours.per.week + native.country + income, data
summary(q2_mod1)
```

```
##
## Call:
## glm(formula = id ~ age + workclass + education + race + hours.per.week +
##     native.country + income, family = "poisson", data = num_marriage)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.15253  -0.34087  -0.00819   0.42178   2.84901
##
## Coefficients:
##                                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)                             1.8443844  0.0434256  42.472  < 2e-16
## age                                    -0.0071509  0.0002347 -30.464  < 2e-16
## workclassFederal-gov                   -0.0970819  0.0212525  -4.568 4.92e-06
## workclassLocal-gov                     -0.1114973  0.0168590  -6.614 3.75e-11
## workclassNever-worked                  -0.1682409  0.1861663  -0.904 0.366148
## workclassPrivate                       -0.0824701  0.0122971  -6.706 1.99e-11
## workclassSelf-emp-inc                  -0.0705162  0.0209475  -3.366 0.000762
## workclassSelf-emp-not-inc              -0.1037563  0.0162852  -6.371 1.88e-10
## workclassState-gov                     -0.1190027  0.0190641  -6.242 4.31e-10
## workclassWithout-pay                    0.0150471  0.1353989   0.111 0.911512
## education11th                          -0.0033179  0.0221526  -0.150 0.880944
## education12th                           0.0161828  0.0290643   0.557 0.577670
## education1st-4th                        0.0936097  0.0435990   2.147 0.031788
## education5th-6th                        0.0300376  0.0345346   0.870 0.384419
## education7th-8th                        0.0413351  0.0267625   1.545 0.122464
## education9th                           -0.0021821  0.0283812  -0.077 0.938713
## educationAssoc-acdm                    -0.0642215  0.0235878  -2.723 0.006476
## educationAssoc-voc                     -0.0549090  0.0222604  -2.467 0.013638
## educationBachelors                      0.0258232  0.0184706   1.398 0.162092
## educationDoctorate                      0.0859457  0.0325266   2.642 0.008234
## educationHS-grad                       -0.0278324  0.0175267  -1.588 0.112287
## educationMasters                        0.0235112  0.0218454   1.076 0.281813
## educationPreschool                      0.1120146  0.0711333   1.575 0.115323
## educationProf-school                    0.0234201  0.0292838   0.800 0.423849
## educationSome-college                  -0.0193794  0.0178205  -1.087 0.276826
## raceAsian-Pac-Islander                  0.0686075  0.0387298   1.771 0.076488
## raceBlack                               0.1064710  0.0314306   3.387 0.000705
## raceOther                               0.0221030  0.0437658   0.505 0.613539
## raceWhite                               0.0189352  0.0302625   0.626 0.531513
## hours.per.week                         -0.0045934  0.0002520 -18.226  < 2e-16
## native.countryCambodia                 -0.0119627  0.1214908  -0.098 0.921562
## native.countryCanada                   -0.0229040  0.0534268  -0.429 0.668142
## native.countryChina                    -0.1308136  0.0696546  -1.878 0.060377
## native.countryColumbia                  0.0967008  0.0682575   1.417 0.156569
## native.countryCuba                     -0.0145053  0.0594082  -0.244 0.807104
```

```
## native.countryDominican-Republic          0.0070061  0.0650972   0.108 0.914293
## native.countryEcuador                     -0.0209340  0.1014014  -0.206 0.836442
## native.countryEl-Salvador                  0.0619268  0.0525897   1.178 0.238978
## native.countryEngland                     -0.0222610  0.0604959  -0.368 0.712892
## native.countryFrance                      -0.0460710  0.1033877  -0.446 0.655876
## native.countryGermany                     -0.0522233  0.0509524  -1.025 0.305389
## native.countryGreece                       0.0368341  0.1018993   0.361 0.717744
## native.countryGuatemala                    0.0779764  0.0645881   1.207 0.227321
## native.countryHaiti                       -0.0300719  0.0790528  -0.380 0.703647
## native.countryHoland-Netherlands           0.2605310  0.4477910   0.582 0.560692
## native.countryHonduras                    -0.0530821  0.1446233  -0.367 0.713591
## native.countryHong                        -0.0515897  0.1188513  -0.434 0.664238
## native.countryHungary                      0.1746013  0.1391567   1.255 0.209584
## native.countryIndia                       -0.0479179  0.0596692  -0.803 0.421940
## native.countryIran                        -0.0632577  0.0864117  -0.732 0.464138
## native.countryIreland                      0.0507838  0.1065323   0.477 0.633577
## native.countryItaly                        0.0338097  0.0660487   0.512 0.608728
## native.countryJamaica                      0.0234554  0.0591281   0.397 0.691597
## native.countryJapan                       -0.0776631  0.0729817  -1.064 0.287263
## native.countryLaos                        -0.0734393  0.1259756  -0.583 0.559917
## native.countryMexico                      -0.0105772  0.0308663  -0.343 0.731840
## native.countryNicaragua                    0.0400814  0.0876891   0.457 0.647610
## native.countryOutlying-US(Guam-USVI-etc)   0.0088123  0.1378756   0.064 0.949038
## native.countryPeru                         0.0508127  0.0917807   0.554 0.579831
## native.countryPhilippines                 -0.0202698  0.0470474  -0.431 0.666585
## native.countryPoland                       0.0498694  0.0700289   0.712 0.476386
## native.countryPortugal                    -0.0633038  0.0910965  -0.695 0.487112
## native.countryPuerto-Rico                  0.0277310  0.0529634   0.524 0.600566
## native.countryScotland                    -0.0449353  0.1577359  -0.285 0.775738
## native.countrySouth                       -0.0056749  0.0643603  -0.088 0.929739
## native.countryTaiwan                      -0.0723482  0.0776829  -0.931 0.351684
## native.countryThailand                     0.0848317  0.1181518   0.718 0.472764
## native.countryTrinadad&Tobago             -0.0722477  0.1216760  -0.594 0.552665
## native.countryUnited-States               -0.0187238  0.0220253  -0.850 0.395266
## native.countryVietnam                     -0.0348251  0.0682473  -0.510 0.609856
## native.countryYugoslavia                  -0.0702666  0.1404229  -0.500 0.616799
## income>50K                                -0.1199414  0.0081334 -14.747  < 2e-16
##
## (Intercept)                              ***
## age                                      ***
## workclassFederal-gov                     ***
## workclassLocal-gov                       ***
## workclassNever-worked
## workclassPrivate                         ***
## workclassSelf-emp-inc                    ***
## workclassSelf-emp-not-inc                ***
## workclassState-gov                       ***
## workclassWithout-pay
## education11th
## education12th
## education1st-4th                         *
## education5th-6th
## education7th-8th
## education9th
```

```
## educationAssoc-acdm                      **
## educationAssoc-voc                       *
## educationBachelors
## educationDoctorate                       **
## educationHS-grad
## educationMasters
## educationPreschool
## educationProf-school
## educationSome-college
## raceAsian-Pac-Islander                   .
## raceBlack                                ***
## raceOther
## raceWhite
## hours.per.week                           ***
## native.countryCambodia
## native.countryCanada
## native.countryChina                      .
## native.countryColumbia
## native.countryCuba
## native.countryDominican-Republic
## native.countryEcuador
## native.countryEl-Salvador
## native.countryEngland
## native.countryFrance
## native.countryGermany
## native.countryGreece
## native.countryGuatemala
## native.countryHaiti
## native.countryHoland-Netherlands
## native.countryHonduras
## native.countryHong
## native.countryHungary
## native.countryIndia
## native.countryIran
## native.countryIreland
## native.countryItaly
## native.countryJamaica
## native.countryJapan
## native.countryLaos
## native.countryMexico
## native.countryNicaragua
## native.countryOutlying-US(Guam-USVI-etc)
## native.countryPeru
## native.countryPhilippines
## native.countryPoland
## native.countryPortugal
## native.countryPuerto-Rico
## native.countryScotland
## native.countrySouth
## native.countryTaiwan
## native.countryThailand
## native.countryTrinadad&Tobago
## native.countryUnited-States
## native.countryVietnam
```

```
## native.countryYugoslavia
## income>50K                                ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 22383  on 32560  degrees of freedom
## Residual deviance: 19822  on 32489  degrees of freedom
## AIC: 119796
##
## Number of Fisher Scoring iterations: 4
```