Final Progress Report

Yessica Rubio

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.4.4     v purrr   1.0.2
## v tibble  3.2.1     v dplyr   1.1.3
## v tidyr   1.3.0     v stringr 1.5.0
## v readr   2.1.2     v forcats 0.5.1

## Warning: package 'tibble' was built under R version 4.1.2

## Warning: package 'tidyr' was built under R version 4.1.2

## Warning: package 'readr' was built under R version 4.1.2

## Warning: package 'stringr' was built under R version 4.1.2

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(openintro) # contains email data
```

```
## Warning: package 'openintro' was built under R version 4.1.2

## Loading required package: airports

## Loading required package: cherryblossom

## Loading required package: usdata
```

```
library(ggplot2)  #access ggplot
library(vcdExtra)
```

```
## Loading required package: vcd

## Loading required package: grid

## Loading required package: gnm
```

```
## Warning: package 'gnm' was built under R version 4.1.2
```

```
##
## Attaching package: 'vcdExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     summarise
```

```
library(magrittr)
```

```
## Warning: package 'magrittr' was built under R version 4.1.2
```

```
##
## Attaching package: 'magrittr'
```

```
## The following object is masked from 'package:purrr':
##
##     set_names
```

```
## The following object is masked from 'package:tidyr':
##
##     extract
```

```
library(MASS)
```

```
## Warning: package 'MASS' was built under R version 4.1.2
```

```
##
## Attaching package: 'MASS'
```

```
## The following objects are masked from 'package:openintro':
##
##     housing, mammals
```

```
## The following object is masked from 'package:dplyr':
##
##     select
```

```
library(lme4)     # access the mixed functions
```

```
## Warning: package 'lme4' was built under R version 4.1.2
```

```
## Loading required package: Matrix
```

```
## Warning: package 'Matrix' was built under R version 4.1.2
```

```
##
## Attaching package: 'Matrix'
```

```
## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack
```

```
library(VGAM)     # contains crash data
```

```
## Warning: package 'VGAM' was built under R version 4.1.2
```

```
## Loading required package: stats4
```

```
## Loading required package: splines
```

```
##
## Attaching package: 'VGAM'
```

```
## The following object is masked from 'package:tidyr':
##
##     fill
```

```
library(tree)     # for classification trees
```

```
## Warning: package 'tree' was built under R version 4.1.2
```

```
library(pROC)     # ROC curves
```

```
## Type 'citation("pROC")' for a citation.
```

```
##
## Attaching package: 'pROC'
```

```
## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

```
library(boot)     # contains the cv.glm function
```

```
##
## Attaching package: 'boot'
```

```
## The following objects are masked from 'package:VGAM':
##
##     logit, simplex
```

```
## The following object is masked from 'package:openintro':
##
##     salinity
```

```
library(car)      # check for multicollinearity
```

```
## Warning: package 'car' was built under R version 4.1.2
```

```
## Loading required package: carData
```

```
## Warning: package 'carData' was built under R version 4.1.2

##
## Attaching package: 'carData'

## The following object is masked from 'package:vcdExtra':
##
##     Burt

##
## Attaching package: 'car'

## The following object is masked from 'package:boot':
##
##     logit

## The following object is masked from 'package:VGAM':
##
##     logit

## The following object is masked from 'package:openintro':
##
##     densityPlot

## The following object is masked from 'package:dplyr':
##
##     recode

## The following object is masked from 'package:purrr':
##
##     some
```

DATA DESCRIPTION

I chose the census income data from the UCI machine learning repository. It contains 32,561 observations on 15 variables, which include "age", "workclass", "final-weight","education","education-num","marital-status","occupation", "relationship","race","sex","capital-gain","capital-loss","hours-per-week", "native-country",and "earnings".

The data was contained by following the conditions for people over the age of 16 and having worked at least one hour.

The data is clean from unknown values as they were removed to successfully predict or determine whether a person makes over 50K a year.

RESPONSE AND METHODS USED TO ANSWER QUESTIONS OF INTEREST

I'll be using the earnings variable as the response variable in the case whether or not a person makes more than 50K.

My method includes fitting a model to answer the question of interest. I plan on using a binomial logistic regression to attempt to predict whether a person falls into one of the two categories for earnings, whether a person makes more than 50K or not.

I will look at some exploratory plots to look at different relationships between the explanatory variables to then fit my model.

After looking at the summary of my model I will check for over dispersion and account for that over dispersion with a negative binomial model.

EXPLORATORY ANALYSIS

```
# Load and look at the data. srip.white gets rid of any whitespace in the data, for example in V15 " >5
AdultData <- read.csv('adult.data', strip.white = TRUE, header=FALSE)
head(AdultData)
```

```
##   V1               V2     V3        V4 V5                  V6                 V7
## 1 39        State-gov  77516 Bachelors 13       Never-married       Adm-clerical
## 2 50 Self-emp-not-inc  83311 Bachelors 13  Married-civ-spouse    Exec-managerial
## 3 38          Private 215646   HS-grad  9            Divorced  Handlers-cleaners
## 4 53          Private 234721      11th  7  Married-civ-spouse  Handlers-cleaners
## 5 28          Private 338409 Bachelors 13  Married-civ-spouse     Prof-specialty
## 6 37          Private 284582   Masters 14  Married-civ-spouse    Exec-managerial
##                V8    V9    V10  V11 V12 V13           V14   V15
## 1 Not-in-family White   Male 2174   0  40 United-States <=50K
## 2       Husband White   Male    0   0  13 United-States <=50K
## 3 Not-in-family White   Male    0   0  40 United-States <=50K
## 4       Husband Black   Male    0   0  40 United-States <=50K
## 5          Wife Black Female    0   0  40          Cuba <=50K
## 6          Wife White Female    0   0  40 United-States <=50K
```

```
AdultData2 <- AdultData
```

```
# Create new column that will take value 1 if earnings greater than 50k and 0 otherwise. This will be t
AdultData2 %<>% mutate(.,V16 = ifelse((V15 == ">50K"),1,0))
head(AdultData2)
```

```
##   V1               V2     V3        V4 V5                  V6                 V7
## 1 39        State-gov  77516 Bachelors 13       Never-married       Adm-clerical
## 2 50 Self-emp-not-inc  83311 Bachelors 13  Married-civ-spouse    Exec-managerial
## 3 38          Private 215646   HS-grad  9            Divorced  Handlers-cleaners
## 4 53          Private 234721      11th  7  Married-civ-spouse  Handlers-cleaners
## 5 28          Private 338409 Bachelors 13  Married-civ-spouse     Prof-specialty
## 6 37          Private 284582   Masters 14  Married-civ-spouse    Exec-managerial
##                V8    V9    V10  V11 V12 V13           V14   V15 V16
## 1 Not-in-family White   Male 2174   0  40 United-States <=50K   0
## 2       Husband White   Male    0   0  13 United-States <=50K   0
## 3 Not-in-family White   Male    0   0  40 United-States <=50K   0
## 4       Husband Black   Male    0   0  40 United-States <=50K   0
## 5          Wife Black Female    0   0  40          Cuba <=50K   0
## 6          Wife White Female    0   0  40 United-States <=50K   0
```

```r
# Change the names of the columns
colnames(AdultData2) <- c("age", "workclass", "final-weight","education","education-num","marital-status
head(AdultData2)
```

```
##   age        workclass final-weight education education-num     marital-status
## 1  39        State-gov        77516 Bachelors            13      Never-married
## 2  50 Self-emp-not-inc        83311 Bachelors            13 Married-civ-spouse
## 3  38          Private       215646   HS-grad             9           Divorced
## 4  53          Private       234721      11th             7 Married-civ-spouse
## 5  28          Private       338409 Bachelors            13 Married-civ-spouse
## 6  37          Private       284582   Masters            14 Married-civ-spouse
##          occupation  relationship  race    sex capital-gain capital-loss
## 1      Adm-clerical Not-in-family White   Male         2174            0
## 2   Exec-managerial       Husband White   Male            0            0
## 3 Handlers-cleaners Not-in-family White   Male            0            0
## 4 Handlers-cleaners       Husband Black   Male            0            0
## 5    Prof-specialty          Wife Black Female            0            0
## 6   Exec-managerial          Wife White Female            0            0
##   hours-per-week native-country earnings earnings50K
## 1             40  United-States    <=50K           0
## 2             13  United-States    <=50K           0
## 3             40  United-States    <=50K           0
## 4             40  United-States    <=50K           0
## 5             40           Cuba    <=50K           0
## 6             40  United-States    <=50K           0
```

```r
# Here I will clean up the data some more and state the variables as factors that need to be factors.
AdultData2[AdultData2 == "?"] <- NA


AdultData2$workclass <- as.factor(AdultData2$workclass)
AdultData2$education <- as.factor(AdultData2$education)
AdultData2$`marital-status` <- as.factor(AdultData2$`marital-status`)
AdultData2$relationship <- as.factor(AdultData2$relationship)
AdultData2$race <- as.factor(AdultData2$race)
AdultData2$sex <- as.factor(AdultData2$sex)
AdultData2$`native-country` <- as.factor(AdultData2$`native-country`)
AdultData2$earnings <- as.factor(AdultData2$earnings)
AdultData2$occupation <- as.factor(AdultData2$occupation)
```

```r
# After all the data cleaning and data manipulation check the summary and structure of the data
str(AdultData2)
```

```
## 'data.frame':    32561 obs. of  16 variables:
##  $ age           : int  39 50 38 53 28 37 49 52 31 42 ...
##  $ workclass     : Factor w/ 8 levels "Federal-gov",..: 7 6 4 4 4 4 4 6 4 4 ...
##  $ final-weight  : int  77516 83311 215646 234721 338409 284582 160187 209642 45781 159449 ...
##  $ education     : Factor w/ 16 levels "10th","11th",..: 10 10 12 2 10 13 7 12 13 10 ...
##  $ education-num : int  13 13 9 7 13 14 5 9 14 13 ...
##  $ marital-status: Factor w/ 7 levels "Divorced","Married-AF-spouse",..: 5 3 1 3 3 3 4 3 5 3 ...
##  $ occupation    : Factor w/ 14 levels "Adm-clerical",..: 1 4 6 6 10 4 8 4 10 4 ...
##  $ relationship  : Factor w/ 6 levels "Husband","Not-in-family",..: 2 1 2 1 6 6 2 1 2 1 ...
```

```
##  $ race          : Factor w/ 5 levels "Amer-Indian-Eskimo",..: 5 5 5 3 3 5 3 5 5 5 ...
##  $ sex           : Factor w/ 2 levels "Female","Male": 2 2 2 2 1 1 1 2 1 2 ...
##  $ capital-gain  : int  2174 0 0 0 0 0 0 0 14084 5178 ...
##  $ capital-loss  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ hours-per-week: int  40 13 40 40 40 40 16 45 50 40 ...
##  $ native-country: Factor w/ 41 levels "Cambodia","Canada",..: 39 39 39 39 5 39 23 39 39 39 ...
##  $ earnings      : Factor w/ 2 levels "<=50K",">50K": 1 1 1 1 1 1 1 2 2 2 ...
##  $ earnings50K   : num  0 0 0 0 0 0 0 1 1 1 ...
```

```
summary(AdultData2)
```

```
##       age                     workclass      final-weight
##  Min.   :17.00   Private         :22696   Min.   :  12285
##  1st Qu.:28.00   Self-emp-not-inc: 2541   1st Qu.: 117827
##  Median :37.00   Local-gov       : 2093   Median : 178356
##  Mean   :38.58   State-gov       : 1298   Mean   : 189778
##  3rd Qu.:48.00   Self-emp-inc    : 1116   3rd Qu.: 237051
##  Max.   :90.00   (Other)         :  981   Max.   :1484705
##                  NA's            : 1836
##         education     education-num                marital-status
##  HS-grad     :10501   Min.   : 1.00   Divorced             : 4443
##  Some-college: 7291   1st Qu.: 9.00   Married-AF-spouse    :   23
##  Bachelors   : 5355   Median :10.00   Married-civ-spouse   :14976
##  Masters     : 1723   Mean   :10.08   Married-spouse-absent:  418
##  Assoc-voc   : 1382   3rd Qu.:12.00   Never-married        :10683
##  11th        : 1175   Max.   :16.00   Separated            : 1025
##  (Other)     : 5134                   Widowed              :  993
##            occupation            relationship                    race
##  Prof-specialty : 4140   Husband       :13193   Amer-Indian-Eskimo:  311
##  Craft-repair   : 4099   Not-in-family : 8305   Asian-Pac-Islander: 1039
##  Exec-managerial: 4066   Other-relative:  981   Black             : 3124
##  Adm-clerical   : 3770   Own-child     : 5068   Other             :  271
##  Sales          : 3650   Unmarried     : 3446   White             :27816
##  (Other)        :10993   Wife          : 1568
##  NA's           : 1843
##      sex         capital-gain    capital-loss    hours-per-week
##  Female:10771   Min.   :    0   Min.   :   0.0   Min.   : 1.00
##  Male  :21790   1st Qu.:    0   1st Qu.:   0.0   1st Qu.:40.00
##                 Median :    0   Median :   0.0   Median :40.00
##                 Mean   : 1078   Mean   :  87.3   Mean   :40.44
##                 3rd Qu.:    0   3rd Qu.:   0.0   3rd Qu.:45.00
##                 Max.   :99999   Max.   :4356.0   Max.   :99.00
##
##        native-country   earnings      earnings50K
##  United-States:29170   <=50K:24720   Min.   :0.0000
##  Mexico       :  643   >50K : 7841   1st Qu.:0.0000
##  Philippines  :  198                 Median :0.0000
##  Germany      :  137                 Mean   :0.2408
##  Canada       :  121                 3rd Qu.:0.0000
##  (Other)      : 1709                 Max.   :1.0000
##  NA's         :  583
```

```r
# Are there any NA values?
nrow(AdultData2[is.na(AdultData2$education) | is.na(AdultData2$age) | is.na(AdultData2$occupation),])
```

```
## [1] 1843
```

```r
# Number of observations before removing NAs
nrow(AdultData2)
```

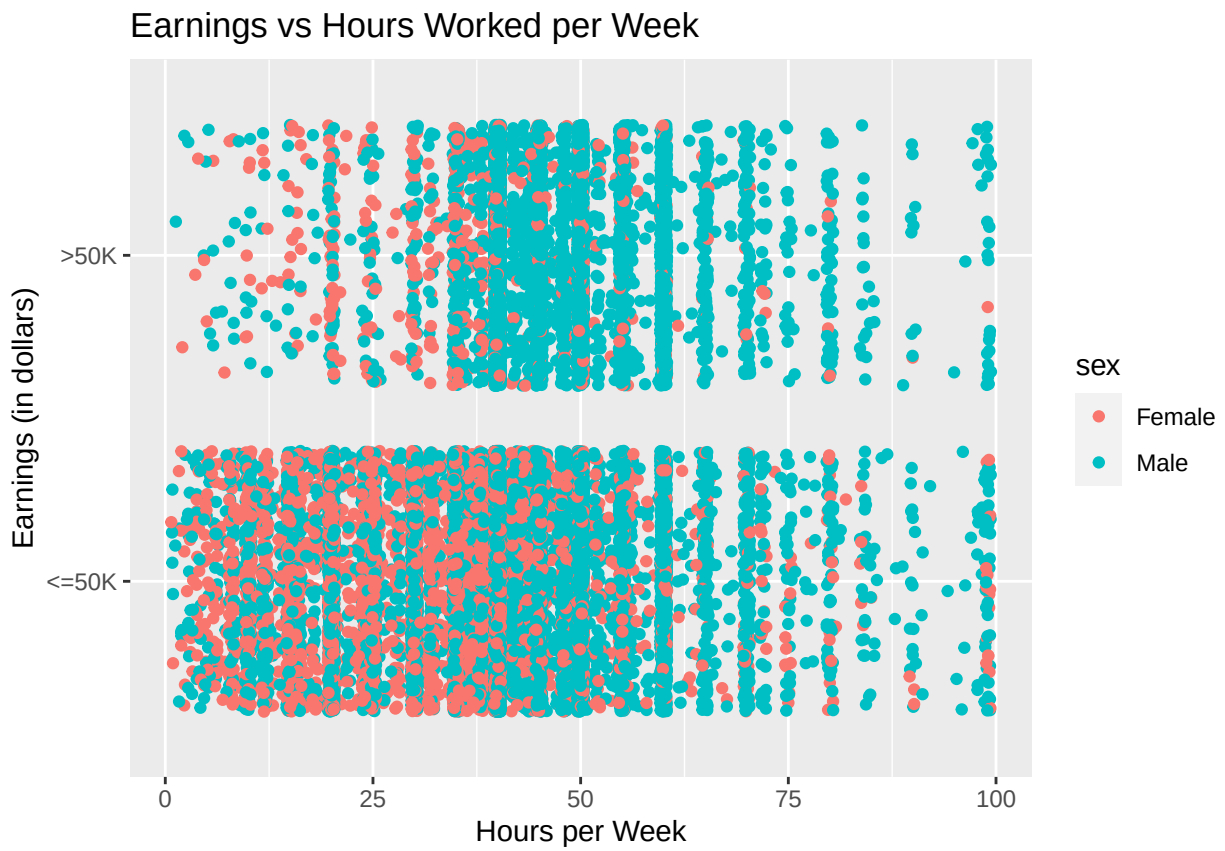
```
## [1] 32561
```

```r
# Remove the NAs
AdultData2 <- AdultData2[!(is.na(AdultData2$education) | is.na(AdultData2$age) | is.na(AdultData2$occupa

# Number of observations after removing NAs
nrow(AdultData2)
```

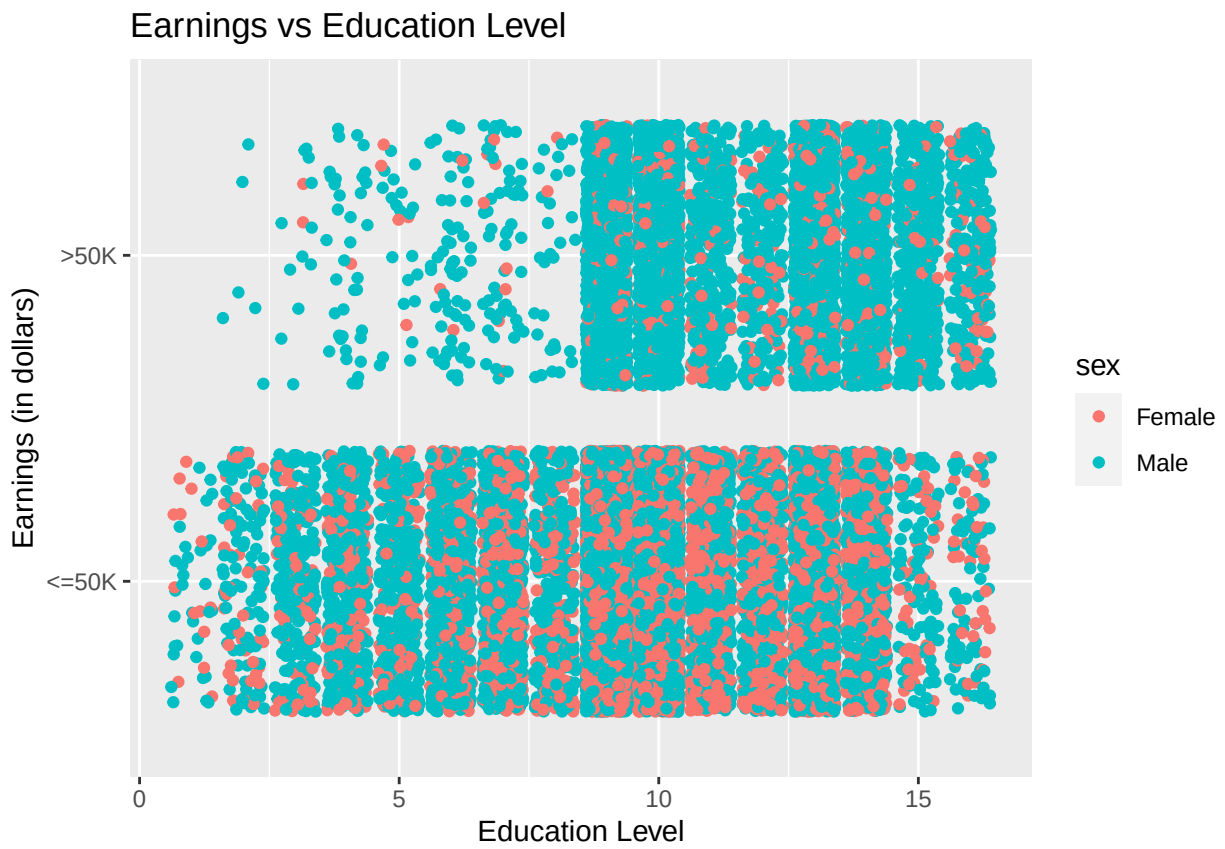
```
## [1] 30718
```

```r
# Make Exploratory Data Analysis (EDA) with ggplot2 to then fit a model
# These plots are AFTER removing the NAs


# USE jitter for the same plots above to see more of the observations without
# the overlapping
ggplot(data = AdultData2) +
  geom_jitter(aes(x = age, y = earnings, color = sex)) +
  ggtitle("Earnings vs Age") +
  labs(
    x = "Age",          # X-axis title
    y = "Earnings (in dollars)"       # Y-axis title
  )
```


```
ggplot(data = AdultData2) +
  geom_jitter(aes(x = `hours-per-week`, y = earnings, color = sex))+
  ggtitle("Earnings vs Hours Worked per Week") +
  labs(
    x = "Hours per Week",          # X-axis title
    y = "Earnings (in dollars)"       # Y-axis title
  )
```




```
ggplot(data = AdultData2) +
  geom_jitter(aes(x = `education-num`, y = earnings, color = sex))+
  ggtitle("Earnings vs Education Level") +
  labs(
    x = "Education Level",          # X-axis title
    y = "Earnings (in dollars)"       # Y-axis title
  )
```



Some observation include a visible concentration around 40 hours worked per week among both genders and a possible trend where individuals with more years of education could be more likely to make more than $50,000. We also see that the majority of individuals who earn more than $50,000 are male across all ages. Overall, we notice the majority of individuals fall in the category that make $50,000 or less.

QUESTION 1

Is there an association between earnings over 50K and hours worked per week, after accounting for other factors (sex, age, education)?

```
I could fit two models to see if there is an association between
earnings and hours worked per week. The first will be a rich model
without the hours worked per week variable then refit the model
adding the hours worked per week variable.
QUESTION 1
```

```
# ALL VARIABLES IN THE MODEL excluding hours per week
richMod1 <- glm(earnings50K ~ sex + age + education + workclass + `final-weight` +
                  `marital-status` + occupation + relationship + race + `capital-gain` +
```

```
                  `capital-loss` + `native-country` , data = AdultData2,
                family = binomial)
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
summary(richMod1)
```

```
##
## Call:
## glm(formula = earnings50K ~ sex + age + education + workclass +
##     `final-weight` + `marital-status` + occupation + relationship +
##     race + `capital-gain` + `capital-loss` + `native-country`,
##     family = binomial, data = AdultData2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -5.1363  -0.5253  -0.1956   0.0000   3.8174
##
## Coefficients:
##                                            Estimate Std. Error z value
## (Intercept)                               -5.067e+00  7.569e-01  -6.694
## sexMale                                    9.295e-01  8.009e-02  11.605
## age                                        2.063e-02  1.665e-03  12.392
## education11th                              5.767e-02  2.131e-01   0.271
## education12th                              4.429e-01  2.775e-01   1.596
## education1st-4th                          -5.475e-01  4.941e-01  -1.108
## education5th-6th                          -4.202e-01  3.554e-01  -1.182
## education7th-8th                          -5.274e-01  2.406e-01  -2.192
## education9th                              -2.369e-01  2.684e-01  -0.883
## educationAssoc-acdm                        1.284e+00  1.786e-01   7.189
## educationAssoc-voc                         1.262e+00  1.719e-01   7.337
## educationBachelors                         1.910e+00  1.599e-01  11.941
## educationDoctorate                         3.099e+00  2.222e-01  13.944
## educationHS-grad                           7.789e-01  1.556e-01   5.006
## educationMasters                           2.308e+00  1.709e-01  13.510
## educationPreschool                        -2.100e+01  2.059e+02  -0.102
## educationProf-school                       2.948e+00  2.063e-01  14.288
## educationSome-college                      1.123e+00  1.579e-01   7.116
## workclassLocal-gov                        -7.038e-01  1.123e-01  -6.265
## workclassPrivate                          -5.059e-01  9.343e-02  -5.415
## workclassSelf-emp-inc                     -1.814e-01  1.231e-01  -1.473
## workclassSelf-emp-not-inc                 -9.447e-01  1.093e-01  -8.643
## workclassState-gov                        -8.819e-01  1.246e-01  -7.077
## workclassWithout-pay                      -1.359e+01  1.949e+02  -0.070
## `final-weight`                             6.694e-07  1.743e-07   3.841
## `marital-status`Married-AF-spouse          2.886e+00  5.686e-01   5.076
## `marital-status`Married-civ-spouse         2.096e+00  2.714e-01   7.721
## `marital-status`Married-spouse-absent     -1.619e-03  2.383e-01  -0.007
## `marital-status`Never-married             -5.497e-01  8.854e-02  -6.208
## `marital-status`Separated                 -1.363e-01  1.650e-01  -0.826
## `marital-status`Widowed                    6.844e-02  1.565e-01   0.437
## occupationArmed-Forces                    -1.180e+00  1.576e+00  -0.749
```

```
## occupationCraft-repair                       1.059e-01  8.017e-02   1.321
## occupationExec-managerial                    9.125e-01  7.715e-02  11.827
## occupationFarming-fishing                   -6.868e-01  1.367e-01  -5.024
## occupationHandlers-cleaners                 -7.080e-01  1.439e-01  -4.919
## occupationMachine-op-inspct                 -2.338e-01  1.022e-01  -2.287
## occupationOther-service                     -8.330e-01  1.179e-01  -7.062
## occupationPriv-house-serv                   -3.922e+00  1.903e+00  -2.062
## occupationProf-specialty                     5.607e-01  8.177e-02   6.857
## occupationProtective-serv                    6.877e-01  1.251e-01   5.498
## occupationSales                              3.857e-01  8.224e-02   4.690
## occupationTech-support                       6.614e-01  1.112e-01   5.948
## occupationTransport-moving                   5.227e-02  9.883e-02   0.529
## relationshipNot-in-family                    4.856e-01  2.682e-01   1.811
## relationshipOther-relative                  -4.852e-01  2.450e-01  -1.981
## relationshipOwn-child                       -8.563e-01  2.670e-01  -3.207
## relationshipUnmarried                        3.507e-01  2.840e-01   1.235
## relationshipWife                             1.232e+00  1.045e-01  11.781
## raceAsian-Pac-Islander                       8.049e-01  2.841e-01   2.833
## raceBlack                                    4.217e-01  2.393e-01   1.762
## raceOther                                    6.128e-02  3.762e-01   0.163
## raceWhite                                    5.875e-01  2.277e-01   2.580
## `capital-gain`                               3.239e-04  1.072e-05  30.225
## `capital-loss`                               6.474e-04  3.823e-05  16.936
## `native-country`Canada                      -7.719e-01  6.866e-01  -1.124
## `native-country`China                       -1.963e+00  7.028e-01  -2.793
## `native-country`Columbia                    -3.160e+00  1.016e+00  -3.110
## `native-country`Cuba                        -8.036e-01  7.025e-01  -1.144
## `native-country`Dominican-Republic          -2.778e+00  1.223e+00  -2.272
## `native-country`Ecuador                     -1.409e+00  9.529e-01  -1.479
## `native-country`El-Salvador                 -1.772e+00  7.874e-01  -2.251
## `native-country`England                     -7.556e-01  7.001e-01  -1.079
## `native-country`France                      -4.457e-01  8.069e-01  -0.552
## `native-country`Germany                     -6.502e-01  6.763e-01  -0.961
## `native-country`Greece                      -2.008e+00  8.453e-01  -2.375
## `native-country`Guatemala                   -1.432e+00  9.794e-01  -1.463
## `native-country`Haiti                       -1.165e+00  9.243e-01  -1.261
## `native-country`Holand-Netherlands          -1.154e+01  8.827e+02  -0.013
## `native-country`Honduras                    -2.056e+00  2.242e+00  -0.917
## `native-country`Hong                        -1.354e+00  8.920e-01  -1.518
## `native-country`Hungary                     -1.327e+00  9.758e-01  -1.360
## `native-country`India                       -1.651e+00  6.665e-01  -2.477
## `native-country`Iran                        -1.034e+00  7.566e-01  -1.366
## `native-country`Ireland                     -5.394e-01  8.830e-01  -0.611
## `native-country`Italy                       -2.450e-01  7.063e-01  -0.347
## `native-country`Jamaica                     -1.114e+00  7.697e-01  -1.448
## `native-country`Japan                       -8.042e-01  7.249e-01  -1.109
## `native-country`Laos                        -1.875e+00  1.045e+00  -1.794
## `native-country`Mexico                      -1.602e+00  6.633e-01  -2.416
## `native-country`Nicaragua                   -1.957e+00  1.026e+00  -1.906
## `native-country`Outlying-US(Guam-USVI-etc) -1.340e+01  2.088e+02  -0.064
## `native-country`Peru                        -2.135e+00  1.052e+00  -2.029
## `native-country`Philippines                 -8.247e-01  6.431e-01  -1.282
## `native-country`Poland                      -1.195e+00  7.483e-01  -1.596
## `native-country`Portugal                    -1.085e+00  8.885e-01  -1.221
```

```
## `native-country`Puerto-Rico               -1.452e+00  7.362e-01  -1.973
## `native-country`Scotland                   -1.296e+00  1.071e+00  -1.211
## `native-country`South                      -2.273e+00  7.306e-01  -3.111
## `native-country`Taiwan                     -1.461e+00  7.472e-01  -1.956
## `native-country`Thailand                   -1.291e+00  1.009e+00  -1.279
## `native-country`Trinadad&Tobago            -1.498e+00  1.054e+00  -1.421
## `native-country`United-States              -9.064e-01  6.294e-01  -1.440
## `native-country`Vietnam                    -2.321e+00  8.361e-01  -2.776
## `native-country`Yugoslavia                 -3.036e-01  9.092e-01  -0.334
##                                           Pr(>|z|)
## (Intercept)                               2.16e-11 ***
## sexMale                                    < 2e-16 ***
## age                                        < 2e-16 ***
## education11th                             0.786651
## education12th                             0.110431
## education1st-4th                          0.267781
## education5th-6th                          0.237055
## education7th-8th                          0.028382 *
## education9th                              0.377454
## educationAssoc-acdm                       6.52e-13 ***
## educationAssoc-voc                        2.19e-13 ***
## educationBachelors                         < 2e-16 ***
## educationDoctorate                         < 2e-16 ***
## educationHS-grad                          5.56e-07 ***
## educationMasters                           < 2e-16 ***
## educationPreschool                        0.918752
## educationProf-school                       < 2e-16 ***
## educationSome-college                     1.11e-12 ***
## workclassLocal-gov                        3.73e-10 ***
## workclassPrivate                          6.12e-08 ***
## workclassSelf-emp-inc                     0.140649
## workclassSelf-emp-not-inc                  < 2e-16 ***
## workclassState-gov                        1.47e-12 ***
## workclassWithout-pay                      0.944413
## `final-weight`                            0.000122 ***
## `marital-status`Married-AF-spouse         3.86e-07 ***
## `marital-status`Married-civ-spouse        1.16e-14 ***
## `marital-status`Married-spouse-absent     0.994578
## `marital-status`Never-married             5.37e-10 ***
## `marital-status`Separated                 0.408746
## `marital-status`Widowed                   0.661802
## occupationArmed-Forces                    0.453911
## occupationCraft-repair                    0.186481
## occupationExec-managerial                  < 2e-16 ***
## occupationFarming-fishing                 5.06e-07 ***
## occupationHandlers-cleaners               8.71e-07 ***
## occupationMachine-op-inspct               0.022167 *
## occupationOther-service                   1.64e-12 ***
## occupationPriv-house-serv                 0.039241 *
## occupationProf-specialty                  7.03e-12 ***
## occupationProtective-serv                 3.85e-08 ***
## occupationSales                           2.73e-06 ***
## occupationTech-support                    2.71e-09 ***
## occupationTransport-moving                0.596856
```

```
## relationshipNot-in-family                  0.070183 .
## relationshipOther-relative                 0.047613 *
## relationshipOwn-child                      0.001339 **
## relationshipUnmarried                      0.216919
## relationshipWife                            < 2e-16 ***
## raceAsian-Pac-Islander                     0.004618 **
## raceBlack                                  0.078067 .
## raceOther                                  0.870609
## raceWhite                                  0.009890 **
## `capital-gain`                              < 2e-16 ***
## `capital-loss`                              < 2e-16 ***
## `native-country`Canada                     0.260929
## `native-country`China                      0.005230 **
## `native-country`Columbia                   0.001873 **
## `native-country`Cuba                       0.252675
## `native-country`Dominican-Republic         0.023103 *
## `native-country`Ecuador                    0.139177
## `native-country`El-Salvador                0.024392 *
## `native-country`England                    0.280460
## `native-country`France                     0.580716
## `native-country`Germany                    0.336347
## `native-country`Greece                     0.017548 *
## `native-country`Guatemala                  0.143583
## `native-country`Haiti                      0.207406
## `native-country`Holand-Netherlands         0.989569
## `native-country`Honduras                   0.359106
## `native-country`Hong                       0.128896
## `native-country`Hungary                    0.173902
## `native-country`India                      0.013241 *
## `native-country`Iran                       0.171822
## `native-country`Ireland                    0.541286
## `native-country`Italy                      0.728740
## `native-country`Jamaica                    0.147625
## `native-country`Japan                      0.267250
## `native-country`Laos                       0.072804 .
## `native-country`Mexico                     0.015697 *
## `native-country`Nicaragua                  0.056593 .
## `native-country`Outlying-US(Guam-USVI-etc) 0.948826
## `native-country`Peru                       0.042409 *
## `native-country`Philippines                0.199728
## `native-country`Poland                     0.110423
## `native-country`Portugal                   0.221969
## `native-country`Puerto-Rico                0.048534 *
## `native-country`Scotland                   0.226021
## `native-country`South                      0.001864 **
## `native-country`Taiwan                     0.050501 .
## `native-country`Thailand                   0.200833
## `native-country`Trinadad&Tobago            0.155198
## `native-country`United-States              0.149845
## `native-country`Vietnam                    0.005507 **
## `native-country`Yugoslavia                 0.738449
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 33851  on 30161  degrees of freedom
## Residual deviance: 19794  on 30067  degrees of freedom
##   (556 observations deleted due to missingness)
## AIC: 19984
##
## Number of Fisher Scoring iterations: 13
```

```
# ALL VARIABLES IN THE MODEL + hours-per-week
richMod11 <- glm(earnings50K ~ sex + age + education + workclass + `final-weight` +
                   `marital-status` + occupation + relationship + race + `capital-gain` +
                   `capital-loss` + `native-country` + `hours-per-week`, data = AdultData2,
                 family = binomial)
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
summary(richMod11)
```

```
##
## Call:
## glm(formula = earnings50K ~ sex + age + education + workclass +
##     `final-weight` + `marital-status` + occupation + relationship +
##     race + `capital-gain` + `capital-loss` + `native-country` +
##     `hours-per-week`, family = binomial, data = AdultData2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -5.1182  -0.5148  -0.1885   0.0000   3.7839
##
## Coefficients:
##                                          Estimate Std. Error z value
## (Intercept)                            -6.408e+00  7.636e-01  -8.392
## sexMale                                 8.648e-01  8.091e-02  10.689
## age                                     2.550e-02  1.712e-03  14.890
## education11th                           9.462e-02  2.139e-01   0.442
## education12th                           4.443e-01  2.784e-01   1.596
## education1st-4th                       -4.398e-01  4.960e-01  -0.887
## education5th-6th                       -3.956e-01  3.590e-01  -1.102
## education7th-8th                       -5.640e-01  2.433e-01  -2.318
## education9th                           -2.372e-01  2.702e-01  -0.878
## educationAssoc-acdm                     1.269e+00  1.797e-01   7.063
## educationAssoc-voc                      1.268e+00  1.729e-01   7.332
## educationBachelors                      1.899e+00  1.608e-01  11.807
## educationDoctorate                      2.935e+00  2.231e-01  13.159
## educationHS-grad                        7.735e-01  1.564e-01   4.945
## educationMasters                        2.259e+00  1.719e-01  13.138
## educationPreschool                     -2.008e+01  1.987e+02  -0.101
## educationProf-school                    2.844e+00  2.071e-01  13.734
## educationSome-college                   1.109e+00  1.587e-01   6.989
## workclassLocal-gov                     -6.985e-01  1.130e-01  -6.184
## workclassPrivate                       -5.055e-01  9.379e-02  -5.390
## workclassSelf-emp-inc                  -3.293e-01  1.239e-01  -2.658
```

```
## workclassSelf-emp-not-inc                -9.972e-01  1.100e-01  -9.063
## workclassState-gov                       -8.207e-01  1.254e-01  -6.544
## workclassWithout-pay                     -1.329e+01  1.972e+02  -0.067
## `final-weight`                            7.515e-07  1.762e-07   4.264
## `marital-status`Married-AF-spouse         2.768e+00  5.766e-01   4.800
## `marital-status`Married-civ-spouse        2.105e+00  2.747e-01   7.663
## `marital-status`Married-spouse-absent     1.220e-02  2.404e-01   0.051
## `marital-status`Never-married            -4.861e-01  8.926e-02  -5.446
## `marital-status`Separated                -8.940e-02  1.656e-01  -0.540
## `marital-status`Widowed                   1.852e-01  1.582e-01   1.171
## occupationArmed-Forces                   -1.165e+00  1.547e+00  -0.753
## occupationCraft-repair                    6.369e-02  8.076e-02   0.789
## occupationExec-managerial                 8.054e-01  7.794e-02  10.334
## occupationFarming-fishing                -9.809e-01  1.408e-01  -6.968
## occupationHandlers-cleaners              -6.950e-01  1.447e-01  -4.803
## occupationMachine-op-inspct              -2.633e-01  1.027e-01  -2.564
## occupationOther-service                  -8.245e-01  1.191e-01  -6.920
## occupationPriv-house-serv                -4.153e+00  1.723e+00  -2.411
## occupationProf-specialty                  5.165e-01  8.253e-02   6.259
## occupationProtective-serv                 5.978e-01  1.263e-01   4.734
## occupationSales                           2.943e-01  8.318e-02   3.538
## occupationTech-support                    6.648e-01  1.117e-01   5.951
## occupationTransport-moving               -8.982e-02  1.001e-01  -0.898
## relationshipNot-in-family                 4.522e-01  2.716e-01   1.665
## relationshipOther-relative               -3.960e-01  2.477e-01  -1.599
## relationshipOwn-child                    -7.322e-01  2.706e-01  -2.706
## relationshipUnmarried                     3.358e-01  2.873e-01   1.169
## relationshipWife                          1.351e+00  1.057e-01  12.784
## raceAsian-Pac-Islander                    8.280e-01  2.860e-01   2.896
## raceBlack                                 4.359e-01  2.409e-01   1.810
## raceOther                                 1.255e-01  3.786e-01   0.332
## raceWhite                                 5.875e-01  2.291e-01   2.564
## `capital-gain`                            3.225e-04  1.074e-05  30.022
## `capital-loss`                            6.420e-04  3.845e-05  16.696
## `native-country`Canada                   -8.113e-01  6.890e-01  -1.178
## `native-country`China                    -1.916e+00  7.031e-01  -2.725
## `native-country`Columbia                 -3.275e+00  1.031e+00  -3.177
## `native-country`Cuba                     -7.738e-01  7.028e-01  -1.101
## `native-country`Dominican-Republic       -2.915e+00  1.220e+00  -2.390
## `native-country`Ecuador                  -1.400e+00  9.587e-01  -1.461
## `native-country`El-Salvador              -1.745e+00  7.922e-01  -2.203
## `native-country`England                  -8.348e-01  7.004e-01  -1.192
## `native-country`France                   -5.604e-01  8.137e-01  -0.689
## `native-country`Germany                  -6.860e-01  6.781e-01  -1.012
## `native-country`Greece                   -2.126e+00  8.369e-01  -2.540
## `native-country`Guatemala                -1.396e+00  9.798e-01  -1.424
## `native-country`Haiti                    -1.169e+00  9.273e-01  -1.261
## `native-country`Holand-Netherlands       -1.164e+01  8.827e+02  -0.013
## `native-country`Honduras                 -2.306e+00  2.607e+00  -0.885
## `native-country`Hong                     -1.355e+00  9.005e-01  -1.505
## `native-country`Hungary                  -1.254e+00  9.905e-01  -1.266
## `native-country`India                    -1.664e+00  6.682e-01  -2.491
## `native-country`Iran                     -1.123e+00  7.578e-01  -1.482
## `native-country`Ireland                  -6.158e-01  8.884e-01  -0.693
```

```
## `native-country`Italy                        -3.295e-01  7.089e-01  -0.465
## `native-country`Jamaica                      -1.125e+00  7.708e-01  -1.460
## `native-country`Japan                        -9.413e-01  7.294e-01  -1.290
## `native-country`Laos                         -1.883e+00  1.046e+00  -1.801
## `native-country`Mexico                       -1.649e+00  6.648e-01  -2.481
## `native-country`Nicaragua                    -1.880e+00  1.020e+00  -1.843
## `native-country`Outlying-US(Guam-USVI-etc) -1.342e+01  2.095e+02  -0.064
## `native-country`Peru                         -1.985e+00  1.053e+00  -1.884
## `native-country`Philippines                  -8.782e-01  6.441e-01  -1.363
## `native-country`Poland                       -1.146e+00  7.455e-01  -1.537
## `native-country`Portugal                     -1.122e+00  8.849e-01  -1.268
## `native-country`Puerto-Rico                  -1.440e+00  7.381e-01  -1.950
## `native-country`Scotland                     -1.407e+00  1.085e+00  -1.297
## `native-country`South                        -2.446e+00  7.356e-01  -3.325
## `native-country`Taiwan                       -1.384e+00  7.540e-01  -1.835
## `native-country`Thailand                     -1.831e+00  1.017e+00  -1.800
## `native-country`Trinadad&Tobago              -1.580e+00  1.060e+00  -1.490
## `native-country`United-States                -9.549e-01  6.302e-01  -1.515
## `native-country`Vietnam                      -2.395e+00  8.452e-01  -2.834
## `native-country`Yugoslavia                   -4.609e-01  9.193e-01  -0.501
## `hours-per-week`                              2.949e-02  1.702e-03  17.325
##                                             Pr(>|z|)
## (Intercept)                                  < 2e-16 ***
## sexMale                                      < 2e-16 ***
## age                                          < 2e-16 ***
## education11th                               0.658185
## education12th                               0.110525
## education1st-4th                            0.375228
## education5th-6th                            0.270456
## education7th-8th                            0.020461 *
## education9th                                0.379942
## educationAssoc-acdm                         1.63e-12 ***
## educationAssoc-voc                          2.27e-13 ***
## educationBachelors                           < 2e-16 ***
## educationDoctorate                           < 2e-16 ***
## educationHS-grad                            7.61e-07 ***
## educationMasters                             < 2e-16 ***
## educationPreschool                          0.919495
## educationProf-school                         < 2e-16 ***
## educationSome-college                       2.76e-12 ***
## workclassLocal-gov                          6.26e-10 ***
## workclassPrivate                            7.06e-08 ***
## workclassSelf-emp-inc                       0.007857 **
## workclassSelf-emp-not-inc                    < 2e-16 ***
## workclassState-gov                          6.00e-11 ***
## workclassWithout-pay                        0.946265
## `final-weight`                              2.01e-05 ***
## `marital-status`Married-AF-spouse           1.59e-06 ***
## `marital-status`Married-civ-spouse          1.82e-14 ***
## `marital-status`Married-spouse-absent       0.959518
## `marital-status`Never-married               5.16e-08 ***
## `marital-status`Separated                   0.589277
## `marital-status`Widowed                     0.241607
## occupationArmed-Forces                      0.451591
```

```
## occupationCraft-repair                      0.430362
## occupationExec-managerial                    < 2e-16 ***
## occupationFarming-fishing                   3.21e-12 ***
## occupationHandlers-cleaners                 1.57e-06 ***
## occupationMachine-op-inspct                 0.010360 *
## occupationOther-service                     4.50e-12 ***
## occupationPriv-house-serv                   0.015916 *
## occupationProf-specialty                    3.87e-10 ***
## occupationProtective-serv                   2.20e-06 ***
## occupationSales                             0.000403 ***
## occupationTech-support                      2.66e-09 ***
## occupationTransport-moving                  0.369448
## relationshipNot-in-family                   0.095982 .
## relationshipOther-relative                  0.109885
## relationshipOwn-child                       0.006812 **
## relationshipUnmarried                       0.242463
## relationshipWife                             < 2e-16 ***
## raceAsian-Pac-Islander                      0.003785 **
## raceBlack                                   0.070321 .
## raceOther                                   0.740237
## raceWhite                                   0.010343 *
## `capital-gain`                               < 2e-16 ***
## `capital-loss`                               < 2e-16 ***
## `native-country`Canada                      0.238944
## `native-country`China                       0.006439 **
## `native-country`Columbia                    0.001490 **
## `native-country`Cuba                        0.270924
## `native-country`Dominican-Republic          0.016847 *
## `native-country`Ecuador                     0.144101
## `native-country`El-Salvador                 0.027612 *
## `native-country`England                     0.233300
## `native-country`France                      0.491044
## `native-country`Germany                     0.311750
## `native-country`Greece                      0.011087 *
## `native-country`Guatemala                   0.154334
## `native-country`Haiti                       0.207295
## `native-country`Holand-Netherlands          0.989483
## `native-country`Honduras                    0.376267
## `native-country`Hong                        0.132285
## `native-country`Hungary                     0.205684
## `native-country`India                       0.012746 *
## `native-country`Iran                        0.138389
## `native-country`Ireland                     0.488248
## `native-country`Italy                       0.642116
## `native-country`Jamaica                     0.144361
## `native-country`Japan                       0.196881
## `native-country`Laos                        0.071769 .
## `native-country`Mexico                      0.013106 *
## `native-country`Nicaragua                   0.065264 .
## `native-country`Outlying-US(Guam-USVI-etc) 0.948933
## `native-country`Peru                        0.059554 .
## `native-country`Philippines                 0.172770
## `native-country`Poland                      0.124250
## `native-country`Portugal                    0.204960
```

```
## `native-country`Puerto-Rico                0.051129 .
## `native-country`Scotland                   0.194512
## `native-country`South                      0.000883 ***
## `native-country`Taiwan                     0.066463 .
## `native-country`Thailand                   0.071863 .
## `native-country`Trinadad&Tobago            0.136166
## `native-country`United-States              0.129711
## `native-country`Vietnam                    0.004601 **
## `native-country`Yugoslavia                 0.616151
## `hours-per-week`                            < 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: 33851  on 30161  degrees of freedom
## Residual deviance: 19486  on 30066  degrees of freedom
##   (556 observations deleted due to missingness)
## AIC: 19678
##
## Number of Fisher Scoring iterations: 13
```

That's a pretty strong effect! r = p-value < 0.001. Recall: The coefficient in the output is log-odds

```
# Convert for odds ratio
exp(0.02949)
```

```
## [1] 1.029929
```

```
# Confidence interval
exp(0.02949+c(-1,1)*1.96*0.001702)
```

```
## [1] 1.026499 1.033371
```

```
# compare models with drop in deviance
anova(richMod1,richMod11,test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: earnings50K ~ sex + age + education + workclass + `final-weight` +
##     `marital-status` + occupation + relationship + race + `capital-gain` +
##     `capital-loss` + `native-country`
## Model 2: earnings50K ~ sex + age + education + workclass + `final-weight` +
##     `marital-status` + occupation + relationship + race + `capital-gain` +
##     `capital-loss` + `native-country` + `hours-per-week`
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1     30067      19794
## 2     30066      19486  1   308.11 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# compare with the other models with likelihood
LRstats(richMod1,richMod11)
```

```
## Likelihood summary table:
##             AIC   BIC LR Chisq    Df Pr(>Chisq)
## richMod1  19984 20774    19794 30067          1
## richMod11 19678 20476    19486 30066          1
```

```
# fit a rich model without the hours per week variable AFTER removing NAs
# check for over dispersion:sex, age, workclass, education, occupation,and race)?
# no over dispersion! yay
#richMod <- glm(earnings50K ~ sex + age + education,
 #              data = AdultData2, family = binomial)
#summary(richMod)
```

```
# fit a rich model with the hours per week variable AFTER removing NAs
# check for over dispersion:sex, age, workclass, education, occupation,and race)?
#hrsMod <- glm(earnings50K ~ sex + age +  education +
 #                 `hours-per-week`,data = AdultData2, family = binomial)
#summary(hrsMod)
```

```
#observe the coefficients
#coef(summary(richMod))

#coef(summary(hrsMod))
```

That's a pretty strong effect! r = p-value < 0.001.

```
#exp(0.0337)
#exp(0.0337+c(-1,1)*1.96*0.0014)
```

MODEL COMPARISON

Adding hours-per-week to the model significantly improves model fit.

The p-value (< 2.2e-16) means that the reduction in deviance (643.36) is not due to random chance.

Therefore, hours-per-week is a strong predictor of whether someone earns >50K or not after controlling for sex, age, and education.

```
#anova(richMod,hrsMod,test="Chisq")
```

QUESTION 2

After accounting for age, is there evidence of a difference in the earnings probability of men and women?

Since we will be concerned only with how ‘age‘, and ‘sex‘
explain whether someone makes more or less than 50K a year, we will drop
the other variables.

QUESTION #2

```
#Since we will be concerned only with how `age`,`education`, and `sex`
#      explain whether someone makes more or less than 50K a year, we will drop
 #     the other variables.
AdultData3 <- AdultData2
AdultData3 %<>% dplyr::select(.,sex, age, earnings, earnings50K)
head(AdultData3)
```

```
##      sex age earnings earnings50K
## 1   Male  39    <=50K           0
## 2   Male  50    <=50K           0
## 3   Male  38    <=50K           0
## 4   Male  53    <=50K           0
## 5 Female  28    <=50K           0
## 6 Female  37    <=50K           0
```

```
# Where >50K = 1 and <=50K = 0
(tbl1 <- xtabs(data = AdultData3, ~ sex + earnings50K))
```

```
##         earnings50K
## sex           0     1
##   Female   8803  1127
##   Male    14265  6523
```

```
#And see the proportions surviving in each `sex`

prop.table(tbl1, margin = 1) %>% round(2)
```

```
##         earnings50K
## sex          0    1
##   Female  0.89 0.11
##   Male    0.69 0.31
```

```
#Out of the 9,930 females 89% make 50K or less and 11% make more than 50K a year.
#Out of the 20,788 males 69% make 50K or less and 31% make more than 50K a year.
```

```
(male_odds <- 31/69)
```

```
## [1] 0.4492754
```

```
(female_odds <- 11/89)
```

```
## [1] 0.1235955
```

```r
#And the odds ratiosex:

(mf_odds_ratio <- male_odds/female_odds)
```

```
## [1] 3.635046
```

```r
#Ignoring age, the odds of earning >50K for males is 363%% of the odds of earning
#over 50K for females. This sounds like a large probability. There may be
#evidence of a difference in earning probability of men and women.
```

```r
glm_int <- glm(earnings50K ~ sex * age, family = binomial, data = AdultData3)
summary(glm_int)
```

```
##
## Call:
## glm(formula = earnings50K ~ sex * age, family = binomial, data = AdultData3)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8487  -0.7496  -0.5601  -0.3598   2.3062
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.206061   0.096732 -33.144  < 2e-16 ***
## sexMale      0.554760   0.110618   5.015 5.30e-07 ***
## age          0.029487   0.002219  13.291  < 2e-16 ***
## sexMale:age  0.016739   0.002541   6.587 4.48e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 34483  on 30717  degrees of freedom
## Residual deviance: 31207  on 30714  degrees of freedom
## AIC: 31215
##
## Number of Fisher Scoring iterations: 5
```

```r
# Looks like the interaction term between sexMale and age is significant. Will
# also refit without interaction terms

glm2 <- glm(earnings50K ~ sex + age, family = binomial, data = AdultData3)
summary(glm2)
```

```
##
## Call:
## glm(formula = earnings50K ~ sex + age, family = binomial, data = AdultData3)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7638  -0.7744  -0.5832  -0.3078   2.4135
```

```
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.745092   0.056241  -66.59   <2e-16 ***
## sexMale      1.255950   0.035842   35.04   <2e-16 ***
## age          0.042307   0.001073   39.43   <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: 34483  on 30717  degrees of freedom
## Residual deviance: 31250  on 30715  degrees of freedom
## AIC: 31256
##
## Number of Fisher Scoring iterations: 4
```

```
# compare models with drop in deviance
anova(glm2,glm_int,test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: earnings50K ~ sex + age
## Model 2: earnings50K ~ sex * age
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1     30715      31250
## 2     30714      31207  1   43.387 4.491e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

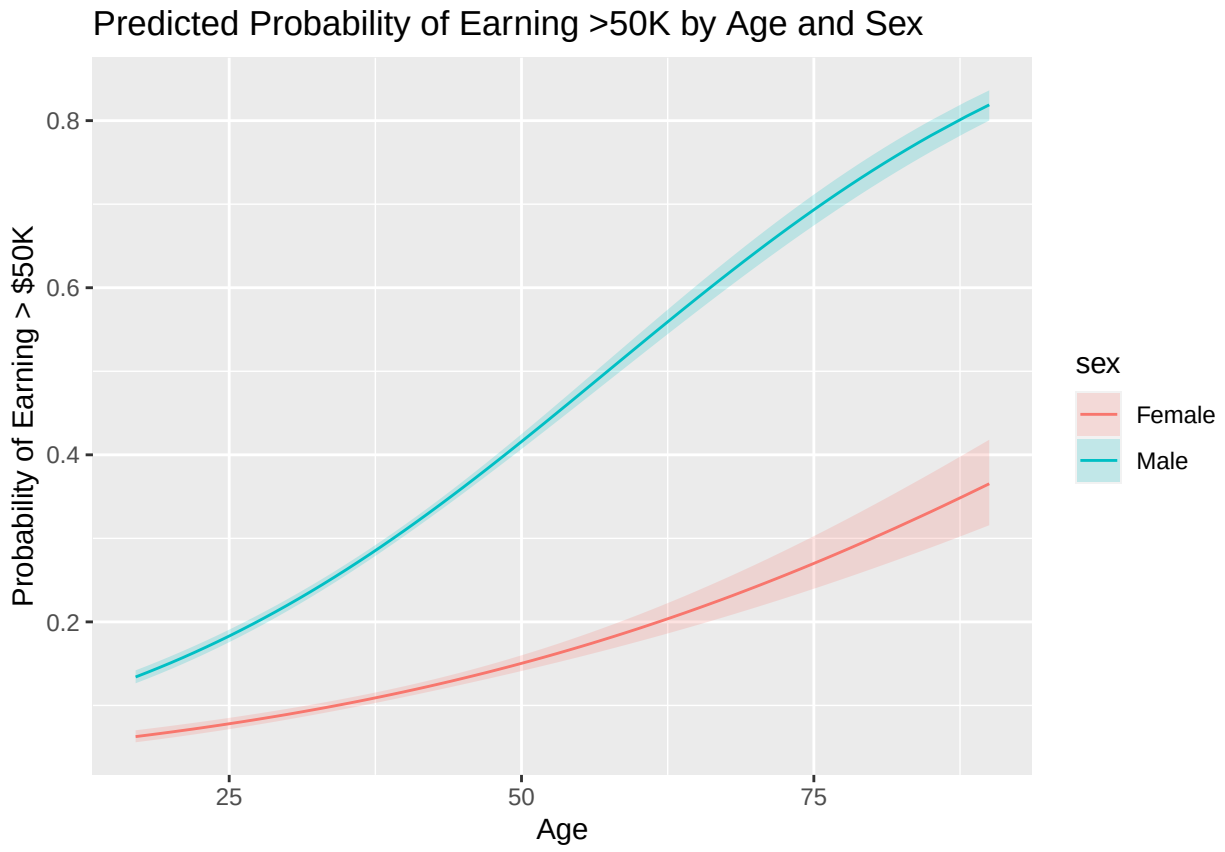
```
# compare with the other models with likelihood
LRstats(glm2,glm_int)
```

```
## Likelihood summary table:
##           AIC   BIC LR Chisq    Df Pr(>Chisq)
## glm2    31256 31281    31250 30715    0.01577 *
## glm_int 31215 31248    31207 30714    0.02379 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
logistic <- function(x){exp(x)/(1 + exp(x))}

# Obtain 95% pointwise confidence bands from predict.glm()
glm_pred <- predict.glm(glm_int, se.fit=TRUE)
low <- glm_pred$fit - 1.96 * glm_pred$se.fit
upp <- glm_pred$fit + 1.96 * glm_pred$se.fit

# back-transform everything to the data scale
glm_fit <- logistic(glm_pred$fit)
glm_lower <- logistic(low)
glm_upper <- logistic(upp)
```

```r
# augment  data frame
augment_Adult <- as.data.frame(cbind(AdultData3, glm_fit, glm_lower, glm_upper))


#ggplot(data = augment_Adult) +
  # plot jittered data
 # geom_jitter(aes(x = age,
  #               y = earnings50K,
   #              color = sex),
    #             height = 0.05, width = 0.2) +

  # plot loess smoother
  #geom_smooth(aes(x = age,
   #             y = earnings50K,
    #             color = sex)) +

#facet_grid(.~sex) +
 # ggtitle("Predicted Probability of Earning >50K by Age and Sex")+
  #labs(
   # x = "Age",          # X-axis title
    #y = "Probability of Earning > $50K"       # Y-axis title
#  )
```

```r
ggplot(augment_Adult, aes(x = age, y = glm_fit, color = sex)) +
  geom_line() +
  geom_ribbon(aes(ymin = glm_lower, ymax = glm_upper, fill = sex), alpha = 0.2, color = NA) +
  ggtitle("Predicted Probability of Earning >50K by Age and Sex")+
  labs(
    x = "Age",          # X-axis title
    y = "Probability of Earning > $50K"       # Y-axis title
  )
```




```
summary(glm_int)
```

```
##
## Call:
## glm(formula = earnings50K ~ sex * age, family = binomial, data = AdultData3)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8487  -0.7496  -0.5601  -0.3598   2.3062
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.206061   0.096732 -33.144  < 2e-16 ***
## sexMale      0.554760   0.110618   5.015 5.30e-07 ***
## age          0.029487   0.002219  13.291  < 2e-16 ***
## sexMale:age  0.016739   0.002541   6.587 4.48e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 34483  on 30717  degrees of freedom
## Residual deviance: 31207  on 30714  degrees of freedom
## AIC: 31215
##
## Number of Fisher Scoring iterations: 5
```

```r
exp(0.016739)#exponentiate sexMale:age
```

```
## [1] 1.01688
```

```r
exp(confint(glm_int))#obtain estimate of multiplicative difference in odds
```

```
## Waiting for profiling to be done...
```

```
##                 2.5 %     97.5 %
## (Intercept) 0.0334743 0.04891182
## sexMale     1.4032974 2.16514612
## age         1.0254559 1.03441462
## sexMale:age 1.0118292 1.02195964
```

```r
#You can also use the logistic back-transformation to make a comparison in
#terms of the probability of earnings >50K:
logistic(0.016739)
```

```
## [1] 0.5041847
```

```r
ci <- confint(glm_int)
```

```
## Waiting for profiling to be done...
```

```r
logistic(ci)
```

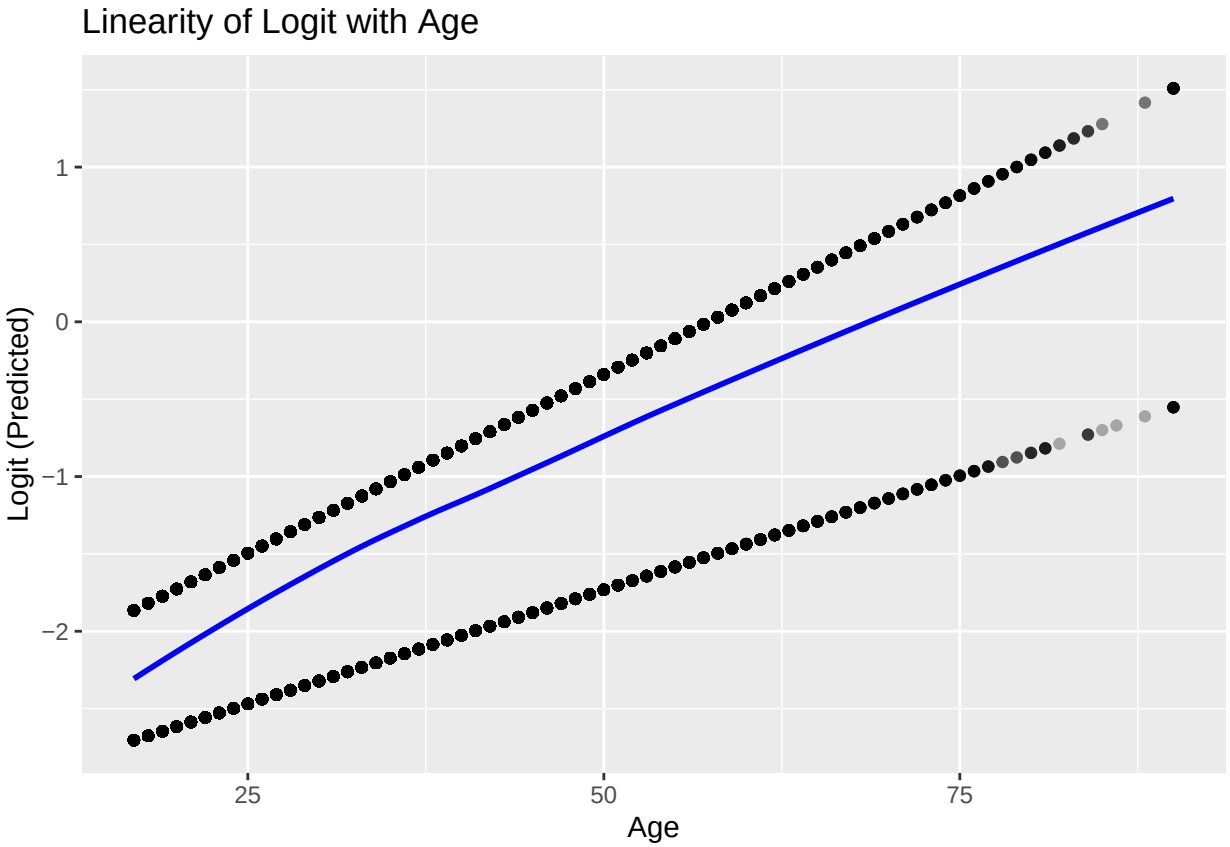
```
##                   2.5 %     97.5 %
## (Intercept) 0.03239006 0.04663101
## sexMale     0.58390501 0.68405882
## age         0.50628400 0.50845811
## sexMale:age 0.50293992 0.50543029
```

CHECK ASSUMPTIONS

1. Linearity of the logit (for age)

```r
# Add predicted logit values to dataset
AdultData3$logit <- predict(glm_int, type = "link")

# Plot logit vs age to visually assess linearity
library(ggplot2)
ggplot(AdultData3, aes(x = age, y = logit)) +
  geom_point(alpha = 0.3) +
  geom_smooth(method = "loess", se = FALSE, color = "blue") +
  labs(title = "Linearity of Logit with Age", x = "Age", y = "Logit (Predicted)")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```




```
#For a formal test (Box-Tidwell), your age variable must be positive:
AdultData3$log_age <- log(AdultData3$age)
library(car)
boxTidwell(earnings50K ~ age, ~ log_age, data = AdultData3)
```

```
##  MLE of lambda Score Statistic (z)  Pr(>|z|)
##         3.0071             -13.607 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## iterations =  23
```

2. Multicollinearity (variance inflation factor)

VIF $>$ 5 or 10 suggests multicollinearity (not likely with just sex, age, and their interaction, but worth checking).

```
library(car)
vif(glm_int)
```

```
## there are higher-order terms (interactions) in this model
## consider setting terms = 'marginal' or 'high-order'; see ?vif

##       sex       age   sex:age
##  9.701054  4.206361 13.003068
```

3. Goodness of fit (hosmer-lemeshow test)

A p-value > 0.05 suggests the model fits the data well.

```
library(ResourceSelection)
```

```
## ResourceSelection 0.3-6   2023-06-27
```

```
# Ensure earnings50K is binary (0/1)
hoslem.test(AdultData3$earnings50K, fitted(glm_int))
```

```
##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  AdultData3$earnings50K, fitted(glm_int)
## X-squared = 1102, df = 8, p-value < 2.2e-16
```

The probability of earning more than 50K a year for a male is estimated to be 0.64 higher than that of a female when we compare a male and a female of the same age. A 95% confidence interval for this (additive) difference runs from 0.58 to 0.68.