471 Midterm

Nik Krieger

2018-03-09

library(class)  
library(rms)  
library(MASS)  
#library(boot)  
#library(ISLR)  
#library(glmnet)  
#library(pls)  
#library(leaps)  
#library(simputation)  
library(tidyverse)

We will import the raw data from the .csv file and immediately make a copy that we will work with.

census\_train\_raw = read\_csv("census\_train.csv")  
census\_train = census\_train\_raw  
Hmisc::describe(census\_train)

## census\_train   
##   
## 15 Variables 25000 Observations  
## ---------------------------------------------------------------------------  
## age   
## n missing distinct Info Mean Gmd .05 .10   
## 25000 0 73 1 38.6 15.41 19 22   
## .25 .50 .75 .90 .95   
## 28 37 48 58 63   
##   
## lowest : 17 18 19 20 21, highest: 85 86 87 88 90  
## ---------------------------------------------------------------------------  
## workclass   
## n missing distinct   
## 25000 0 9   
##   
## ? (1404, 0.056), Federal-gov (734, 0.029), Local-gov (1606, 0.064),  
## Never-worked (7, 0.000), Private (17447, 0.698), Self-emp-inc (856,  
## 0.034), Self-emp-not-inc (1943, 0.078), State-gov (993, 0.040),  
## Without-pay (10, 0.000)  
## ---------------------------------------------------------------------------  
## fnlwgt   
## n missing distinct Info Mean Gmd .05 .10   
## 25000 0 17865 1 189625 112271 39054 65474   
## .25 .50 .75 .90 .95   
## 117833 178302 236986 328705 379069   
##   
## lowest : 12285 13769 14878 18827 19214  
## highest: 1161363 1184622 1226583 1366120 1484705  
## ---------------------------------------------------------------------------  
## education   
## n missing distinct   
## 25000 0 16   
##   
## 10th (726, 0.029), 11th (897, 0.036), 12th (355, 0.014), 1st-4th (129,  
## 0.005), 5th-6th (255, 0.010), 7th-8th (498, 0.020), 9th (402, 0.016),  
## Assoc-acdm (822, 0.033), Assoc-voc (1062, 0.042), Bachelors (4105, 0.164),  
## Doctorate (311, 0.012), HS-grad (8064, 0.323), Masters (1304, 0.052),  
## Preschool (35, 0.001), Prof-school (450, 0.018), Some-college (5585,  
## 0.223)  
## ---------------------------------------------------------------------------  
## education-num   
## n missing distinct Info Mean Gmd .05 .10   
## 25000 0 16 0.951 10.07 2.749 5 7   
## .25 .50 .75 .90 .95   
## 9 10 12 13 14   
##   
## Value 1 2 3 4 5 6 7 8 9 10  
## Frequency 35 129 255 498 402 726 897 355 8064 5585  
## Proportion 0.001 0.005 0.010 0.020 0.016 0.029 0.036 0.014 0.323 0.223  
##   
## Value 11 12 13 14 15 16  
## Frequency 1062 822 4105 1304 450 311  
## Proportion 0.042 0.033 0.164 0.052 0.018 0.012  
## ---------------------------------------------------------------------------  
## marital-status   
## n missing distinct   
## 25000 0 7   
##   
## Divorced (3425, 0.137), Married-AF-spouse (18, 0.001), Married-civ-spouse  
## (11493, 0.460), Married-spouse-absent (327, 0.013), Never-married (8196,  
## 0.328), Separated (790, 0.032), Widowed (751, 0.030)  
## ---------------------------------------------------------------------------  
## occupation   
## n missing distinct   
## 25000 0 15   
##   
## ? (1411, 0.056), Adm-clerical (2885, 0.115), Armed-Forces (6, 0.000),  
## Craft-repair (3160, 0.126), Exec-managerial (3146, 0.126), Farming-fishing  
## (770, 0.031), Handlers-cleaners (1051, 0.042), Machine-op-inspct (1534,  
## 0.061), Other-service (2536, 0.101), Priv-house-serv (114, 0.005),  
## Prof-specialty (3186, 0.127), Protective-serv (487, 0.019), Sales (2786,  
## 0.111), Tech-support (697, 0.028), Transport-moving (1231, 0.049)  
## ---------------------------------------------------------------------------  
## relationship   
## n missing distinct   
## 25000 0 6   
##   
## Value Husband Not-in-family Other-relative Own-child  
## Frequency 10117 6370 776 3887  
## Proportion 0.405 0.255 0.031 0.155  
##   
## Value Unmarried Wife  
## Frequency 2626 1224  
## Proportion 0.105 0.049  
## ---------------------------------------------------------------------------  
## race   
## n missing distinct   
## 25000 0 5   
##   
## Value Amer-Indian-Eskimo Asian-Pac-Islander Black  
## Frequency 239 806 2369  
## Proportion 0.010 0.032 0.095  
##   
## Value Other White  
## Frequency 216 21370  
## Proportion 0.009 0.855  
## ---------------------------------------------------------------------------  
## sex   
## n missing distinct   
## 25000 0 2   
##   
## Value Female Male  
## Frequency 8213 16787  
## Proportion 0.329 0.671  
## ---------------------------------------------------------------------------  
## capital-gain   
## n missing distinct Info Mean Gmd .05 .10   
## 25000 0 118 0.23 1074 2076 0 0   
## .25 .50 .75 .90 .95   
## 0 0 0 0 5013   
##   
## lowest : 0 114 401 594 914, highest: 25124 25236 27828 34095 99999  
## ---------------------------------------------------------------------------  
## capital-loss   
## n missing distinct Info Mean Gmd .05 .10   
## 25000 0 88 0.135 88.31 169.1 0 0   
## .25 .50 .75 .90 .95   
## 0 0 0 0 0   
##   
## lowest : 0 155 213 323 419, highest: 3004 3683 3770 3900 4356  
## ---------------------------------------------------------------------------  
## hours-per-week   
## n missing distinct Info Mean Gmd .05 .10   
## 25000 0 91 0.897 40.41 12.31 16 24   
## .25 .50 .75 .90 .95   
## 40 40 45 55 60   
##   
## lowest : 1 2 3 4 5, highest: 95 96 97 98 99  
## ---------------------------------------------------------------------------  
## native-country   
## n missing distinct   
## 25000 0 42   
##   
## lowest : ? Cambodia Canada China Columbia   
## highest: Thailand Trinadad&Tobago United-States Vietnam Yugoslavia   
## ---------------------------------------------------------------------------  
## income   
## n missing distinct   
## 25000 0 2   
##   
## Value <=50K >50K  
## Frequency 19002 5998  
## Proportion 0.76 0.24  
## ---------------------------------------------------------------------------

The Hmisc::describe() function reports no missing data, but the categorical variables workclass, occupation, and native.country each contain a level named ?. We will flag these values as missing data and then check to see how much missingness there is. Below is the number of rows with mising values and the percentage of rows with missing values.

census\_train[census\_train == "?"] <- NA  
  
nrow(census\_train) - nrow(na.omit(census\_train))

## [1] 1829

(nrow(census\_train) - nrow(na.omit(census\_train)))/nrow(census\_train)

## [1] 0.07316

Hmisc::describe(census\_train)

## census\_train   
##   
## 15 Variables 25000 Observations  
## ---------------------------------------------------------------------------  
## age   
## n missing distinct Info Mean Gmd .05 .10   
## 25000 0 73 1 38.6 15.41 19 22   
## .25 .50 .75 .90 .95   
## 28 37 48 58 63   
##   
## lowest : 17 18 19 20 21, highest: 85 86 87 88 90  
## ---------------------------------------------------------------------------  
## workclass   
## n missing distinct   
## 23596 1404 8   
##   
## Federal-gov (734, 0.031), Local-gov (1606, 0.068), Never-worked (7,  
## 0.000), Private (17447, 0.739), Self-emp-inc (856, 0.036),  
## Self-emp-not-inc (1943, 0.082), State-gov (993, 0.042), Without-pay (10,  
## 0.000)  
## ---------------------------------------------------------------------------  
## fnlwgt   
## n missing distinct Info Mean Gmd .05 .10   
## 25000 0 17865 1 189625 112271 39054 65474   
## .25 .50 .75 .90 .95   
## 117833 178302 236986 328705 379069   
##   
## lowest : 12285 13769 14878 18827 19214  
## highest: 1161363 1184622 1226583 1366120 1484705  
## ---------------------------------------------------------------------------  
## education   
## n missing distinct   
## 25000 0 16   
##   
## 10th (726, 0.029), 11th (897, 0.036), 12th (355, 0.014), 1st-4th (129,  
## 0.005), 5th-6th (255, 0.010), 7th-8th (498, 0.020), 9th (402, 0.016),  
## Assoc-acdm (822, 0.033), Assoc-voc (1062, 0.042), Bachelors (4105, 0.164),  
## Doctorate (311, 0.012), HS-grad (8064, 0.323), Masters (1304, 0.052),  
## Preschool (35, 0.001), Prof-school (450, 0.018), Some-college (5585,  
## 0.223)  
## ---------------------------------------------------------------------------  
## education-num   
## n missing distinct Info Mean Gmd .05 .10   
## 25000 0 16 0.951 10.07 2.749 5 7   
## .25 .50 .75 .90 .95   
## 9 10 12 13 14   
##   
## Value 1 2 3 4 5 6 7 8 9 10  
## Frequency 35 129 255 498 402 726 897 355 8064 5585  
## Proportion 0.001 0.005 0.010 0.020 0.016 0.029 0.036 0.014 0.323 0.223  
##   
## Value 11 12 13 14 15 16  
## Frequency 1062 822 4105 1304 450 311  
## Proportion 0.042 0.033 0.164 0.052 0.018 0.012  
## ---------------------------------------------------------------------------  
## marital-status   
## n missing distinct   
## 25000 0 7   
##   
## Divorced (3425, 0.137), Married-AF-spouse (18, 0.001), Married-civ-spouse  
## (11493, 0.460), Married-spouse-absent (327, 0.013), Never-married (8196,  
## 0.328), Separated (790, 0.032), Widowed (751, 0.030)  
## ---------------------------------------------------------------------------  
## occupation   
## n missing distinct   
## 23589 1411 14   
##   
## Adm-clerical (2885, 0.122), Armed-Forces (6, 0.000), Craft-repair (3160,  
## 0.134), Exec-managerial (3146, 0.133), Farming-fishing (770, 0.033),  
## Handlers-cleaners (1051, 0.045), Machine-op-inspct (1534, 0.065),  
## Other-service (2536, 0.108), Priv-house-serv (114, 0.005), Prof-specialty  
## (3186, 0.135), Protective-serv (487, 0.021), Sales (2786, 0.118),  
## Tech-support (697, 0.030), Transport-moving (1231, 0.052)  
## ---------------------------------------------------------------------------  
## relationship   
## n missing distinct   
## 25000 0 6   
##   
## Value Husband Not-in-family Other-relative Own-child  
## Frequency 10117 6370 776 3887  
## Proportion 0.405 0.255 0.031 0.155  
##   
## Value Unmarried Wife  
## Frequency 2626 1224  
## Proportion 0.105 0.049  
## ---------------------------------------------------------------------------  
## race   
## n missing distinct   
## 25000 0 5   
##   
## Value Amer-Indian-Eskimo Asian-Pac-Islander Black  
## Frequency 239 806 2369  
## Proportion 0.010 0.032 0.095  
##   
## Value Other White  
## Frequency 216 21370  
## Proportion 0.009 0.855  
## ---------------------------------------------------------------------------  
## sex   
## n missing distinct   
## 25000 0 2   
##   
## Value Female Male  
## Frequency 8213 16787  
## Proportion 0.329 0.671  
## ---------------------------------------------------------------------------  
## capital-gain   
## n missing distinct Info Mean Gmd .05 .10   
## 25000 0 118 0.23 1074 2076 0 0   
## .25 .50 .75 .90 .95   
## 0 0 0 0 5013   
##   
## lowest : 0 114 401 594 914, highest: 25124 25236 27828 34095 99999  
## ---------------------------------------------------------------------------  
## capital-loss   
## n missing distinct Info Mean Gmd .05 .10   
## 25000 0 88 0.135 88.31 169.1 0 0   
## .25 .50 .75 .90 .95   
## 0 0 0 0 0   
##   
## lowest : 0 155 213 323 419, highest: 3004 3683 3770 3900 4356  
## ---------------------------------------------------------------------------  
## hours-per-week   
## n missing distinct Info Mean Gmd .05 .10   
## 25000 0 91 0.897 40.41 12.31 16 24   
## .25 .50 .75 .90 .95   
## 40 40 45 55 60   
##   
## lowest : 1 2 3 4 5, highest: 95 96 97 98 99  
## ---------------------------------------------------------------------------  
## native-country   
## n missing distinct   
## 24563 437 41   
##   
## lowest : Cambodia Canada China Columbia Cuba   
## highest: Thailand Trinadad&Tobago United-States Vietnam Yugoslavia   
## ---------------------------------------------------------------------------  
## income   
## n missing distinct   
## 25000 0 2   
##   
## Value <=50K >50K  
## Frequency 19002 5998  
## Proportion 0.76 0.24  
## ---------------------------------------------------------------------------

Only just over 7% of the observations contain a missing value, so we will omit these observations

census\_train <- census\_train %>% na.omit

Here is a cross-table of education and education-num:

table(census\_train$education, census\_train$`education-num`)

##   
## 1 2 3 4 5 6 7 8 9 10 11 12  
## 10th 0 0 0 0 0 636 0 0 0 0 0 0  
## 11th 0 0 0 0 0 0 802 0 0 0 0 0  
## 12th 0 0 0 0 0 0 0 314 0 0 0 0  
## 1st-4th 0 114 0 0 0 0 0 0 0 0 0 0  
## 5th-6th 0 0 225 0 0 0 0 0 0 0 0 0  
## 7th-8th 0 0 0 434 0 0 0 0 0 0 0 0  
## 9th 0 0 0 0 355 0 0 0 0 0 0 0  
## Assoc-acdm 0 0 0 0 0 0 0 0 0 0 0 772  
## Assoc-voc 0 0 0 0 0 0 0 0 0 0 998 0  
## Bachelors 0 0 0 0 0 0 0 0 0 0 0 0  
## Doctorate 0 0 0 0 0 0 0 0 0 0 0 0  
## HS-grad 0 0 0 0 0 0 0 0 7555 0 0 0  
## Masters 0 0 0 0 0 0 0 0 0 0 0 0  
## Preschool 34 0 0 0 0 0 0 0 0 0 0 0  
## Prof-school 0 0 0 0 0 0 0 0 0 0 0 0  
## Some-college 0 0 0 0 0 0 0 0 0 5125 0 0  
##   
## 13 14 15 16  
## 10th 0 0 0 0  
## 11th 0 0 0 0  
## 12th 0 0 0 0  
## 1st-4th 0 0 0 0  
## 5th-6th 0 0 0 0  
## 7th-8th 0 0 0 0  
## 9th 0 0 0 0  
## Assoc-acdm 0 0 0 0  
## Assoc-voc 0 0 0 0  
## Bachelors 3869 0 0 0  
## Doctorate 0 0 0 282  
## HS-grad 0 0 0 0  
## Masters 0 1231 0 0  
## Preschool 0 0 0 0  
## Prof-school 0 0 425 0  
## Some-college 0 0 0 0

Clearly, the variables education and education-num contain the same data in different formats, so we will ignore education-num.

Here is a cross-table of marital-status and relationship:

table(census\_train$`marital-status`, census\_train$relationship)

##   
## Husband Not-in-family Other-relative Own-child  
## Divorced 0 1740 85 242  
## Married-AF-spouse 7 0 1 1  
## Married-civ-spouse 9561 13 84 58  
## Married-spouse-absent 0 140 21 30  
## Never-married 0 3410 435 3015  
## Separated 0 292 42 70  
## Widowed 0 339 33 9  
##   
## Unmarried Wife  
## Divorced 1187 0  
## Married-AF-spouse 0 8  
## Married-civ-spouse 0 1089  
## Married-spouse-absent 97 0  
## Never-married 598 0  
## Separated 317 0  
## Widowed 247 0

Clearly, there is some kind of redundancy here. Since marital-status is intuitive and relationship is decidedly not, we will drop relationship.

Here we turn character variables into factors and change hyphens in variable names to periods. We’ll also create a numerical version of income.

census\_train <- census\_train %>%   
 mutate(workclass = as.factor(workclass), education = as.factor(education),  
 marital.status = as.factor(`marital-status`), occupation = as.factor(occupation),  
 race = as.factor(race), sex = as.factor(sex), capital.gain = `capital-gain`,  
 capital.loss = `capital-loss`, hours.per.week = `hours-per-week`,  
 native.country = as.factor(`native-country`), income = as.factor(income),  
 income.num = ifelse(income==">50K", 1, 0)) %>%   
 select(age, workclass, fnlwgt, education, marital.status, occupation, race, sex,   
 capital.gain, capital.loss, hours.per.week, native.country, income, income.num)

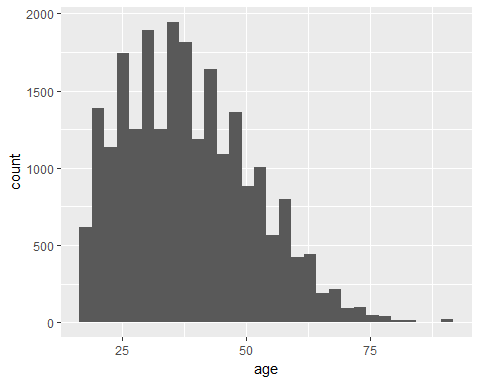
We’ll now look at numerical summaries and histograms of the quantitative predictors:

census\_train %>% select(age, fnlwgt, capital.gain, capital.loss, hours.per.week) %>%  
 Hmisc::describe()

## .   
##   
## 5 Variables 23171 Observations  
## ---------------------------------------------------------------------------  
## age   
## n missing distinct Info Mean Gmd .05 .10   
## 23171 0 72 0.999 38.47 14.86 20 22   
## .25 .50 .75 .90 .95   
## 28 37 47 57 62   
##   
## lowest : 17 18 19 20 21, highest: 84 85 86 88 90  
## ---------------------------------------------------------------------------  
## fnlwgt   
## n missing distinct Info Mean Gmd .05 .10   
## 23171 0 16680 1 189752 112568 39130 65475   
## .25 .50 .75 .90 .95   
## 117674 178344 237529 329144 379403   
##   
## lowest : 13769 14878 18827 19214 19302  
## highest: 1161363 1184622 1226583 1366120 1484705  
## ---------------------------------------------------------------------------  
## capital.gain   
## n missing distinct Info Mean Gmd .05 .10   
## 23171 0 114 0.232 1078 2081 0 0   
## .25 .50 .75 .90 .95   
## 0 0 0 0 5013   
##   
## lowest : 0 114 594 914 991, highest: 25124 25236 27828 34095 99999  
## ---------------------------------------------------------------------------  
## capital.loss   
## n missing distinct Info Mean Gmd .05 .10   
## 23171 0 85 0.137 89.11 170.5 0 0   
## .25 .50 .75 .90 .95   
## 0 0 0 0 0   
##   
## lowest : 0 155 213 323 419, highest: 2824 3004 3683 3770 3900  
## ---------------------------------------------------------------------------  
## hours.per.week   
## n missing distinct Info Mean Gmd .05 .10   
## 23171 0 91 0.894 40.93 11.89 20 25   
## .25 .50 .75 .90 .95   
## 40 40 45 55 60   
##   
## lowest : 1 2 3 4 5, highest: 95 96 97 98 99  
## ---------------------------------------------------------------------------

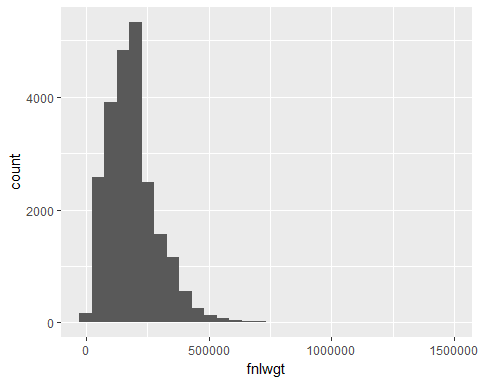
ggplot(data=census\_train, aes(x=age)) +  
 geom\_histogram()

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



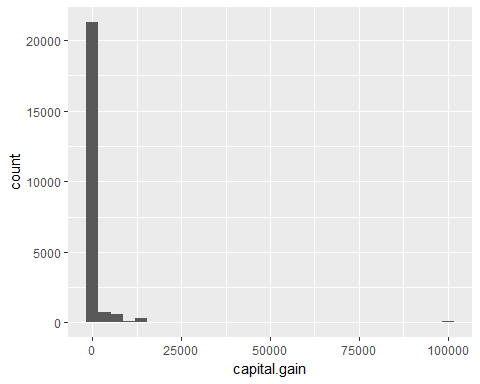
ggplot(data=census\_train, aes(x=fnlwgt)) +  
 geom\_histogram()

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



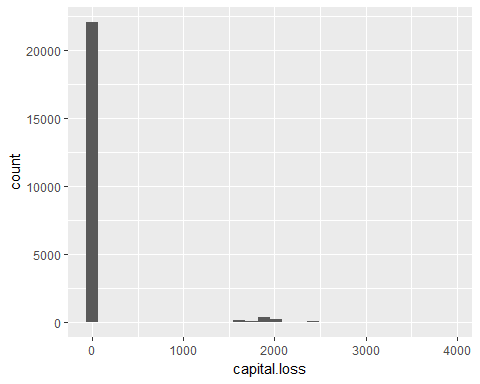
ggplot(data=census\_train, aes(x=capital.gain)) +  
 geom\_histogram()

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



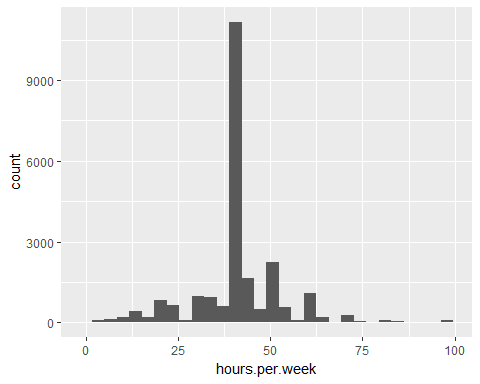
ggplot(data=census\_train, aes(x=capital.loss)) +  
 geom\_histogram()

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



ggplot(data=census\_train, aes(x=hours.per.week)) +  
 geom\_histogram()

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



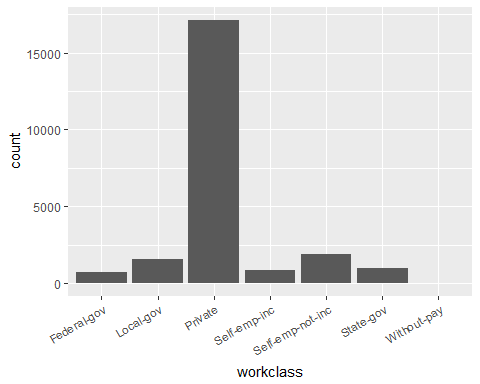
All values seem to make sense and be in a normal range. The only strange one is the maximum capital.gain value of 99999. It is nearly three times higher than the next highest value of 34095, but I am not convinced that it is a mistake or a missingness indicator. I have a feeling that it was censored for the sake of privacy. We will keep this value since it is a meaningful stand-in for subjects who make lots of money in capital gains.

Now we’ll look at bar graphs of the categorical predictors as well as the outcome variable income.

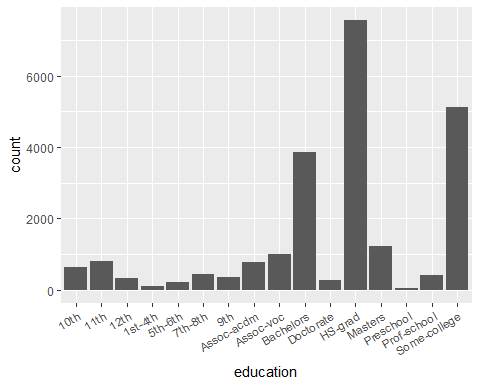
census\_train %>% select(workclass, education, marital.status, occupation,   
 race, sex, native.country, income) %>%  
 Hmisc::describe()

## .   
##   
## 8 Variables 23171 Observations  
## ---------------------------------------------------------------------------  
## workclass   
## n missing distinct   
## 23171 0 7   
##   
## Federal-gov (723, 0.031), Local-gov (1582, 0.068), Private (17136, 0.740),  
## Self-emp-inc (827, 0.036), Self-emp-not-inc (1916, 0.083), State-gov (977,  
## 0.042), Without-pay (10, 0.000)  
## ---------------------------------------------------------------------------  
## education   
## n missing distinct   
## 23171 0 16   
##   
## 10th (636, 0.027), 11th (802, 0.035), 12th (314, 0.014), 1st-4th (114,  
## 0.005), 5th-6th (225, 0.010), 7th-8th (434, 0.019), 9th (355, 0.015),  
## Assoc-acdm (772, 0.033), Assoc-voc (998, 0.043), Bachelors (3869, 0.167),  
## Doctorate (282, 0.012), HS-grad (7555, 0.326), Masters (1231, 0.053),  
## Preschool (34, 0.001), Prof-school (425, 0.018), Some-college (5125,  
## 0.221)  
## ---------------------------------------------------------------------------  
## marital.status   
## n missing distinct   
## 23171 0 7   
##   
## Divorced (3254, 0.140), Married-AF-spouse (17, 0.001), Married-civ-spouse  
## (10805, 0.466), Married-spouse-absent (288, 0.012), Never-married (7458,  
## 0.322), Separated (721, 0.031), Widowed (628, 0.027)  
## ---------------------------------------------------------------------------  
## occupation   
## n missing distinct   
## 23171 0 14   
##   
## Adm-clerical (2850, 0.123), Armed-Forces (6, 0.000), Craft-repair (3110,  
## 0.134), Exec-managerial (3088, 0.133), Farming-fishing (766, 0.033),  
## Handlers-cleaners (1036, 0.045), Machine-op-inspct (1504, 0.065),  
## Other-service (2478, 0.107), Priv-house-serv (108, 0.005), Prof-specialty  
## (3111, 0.134), Protective-serv (483, 0.021), Sales (2730, 0.118),  
## Tech-support (685, 0.030), Transport-moving (1216, 0.052)  
## ---------------------------------------------------------------------------  
## race   
## n missing distinct   
## 23171 0 5   
##   
## Value Amer-Indian-Eskimo Asian-Pac-Islander Black  
## Frequency 219 696 2131  
## Proportion 0.009 0.030 0.092  
##   
## Value Other White  
## Frequency 186 19939  
## Proportion 0.008 0.861  
## ---------------------------------------------------------------------------  
## sex   
## n missing distinct   
## 23171 0 2   
##   
## Value Female Male  
## Frequency 7466 15705  
## Proportion 0.322 0.678  
## ---------------------------------------------------------------------------  
## native.country   
## n missing distinct   
## 23171 0 41   
##   
## lowest : Cambodia Canada China Columbia Cuba   
## highest: Thailand Trinadad&Tobago United-States Vietnam Yugoslavia   
## ---------------------------------------------------------------------------  
## income   
## n missing distinct   
## 23171 0 2   
##   
## Value <=50K >50K  
## Frequency 17431 5740  
## Proportion 0.752 0.248  
## ---------------------------------------------------------------------------

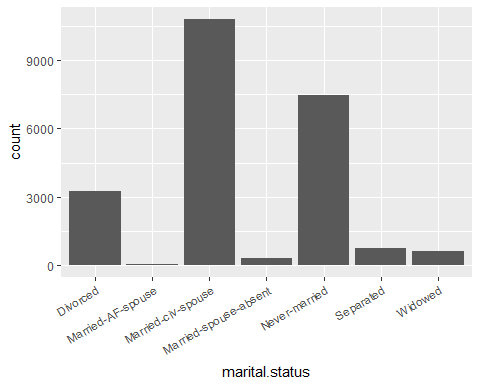
ggplot(data=census\_train, aes(x=workclass)) +  
 geom\_bar() +  
 theme(axis.text.x = element\_text(angle = 30, hjust = 1))



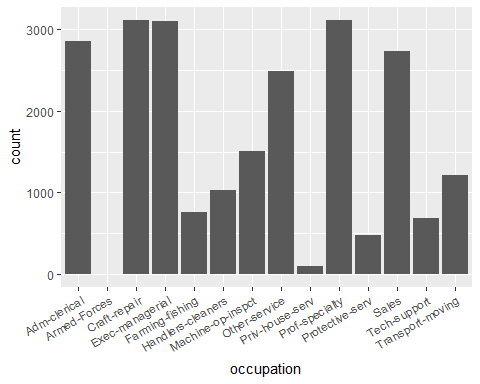
ggplot(data=census\_train, aes(x=education)) +  
 geom\_bar() +  
 theme(axis.text.x = element\_text(angle = 30, hjust = 1))



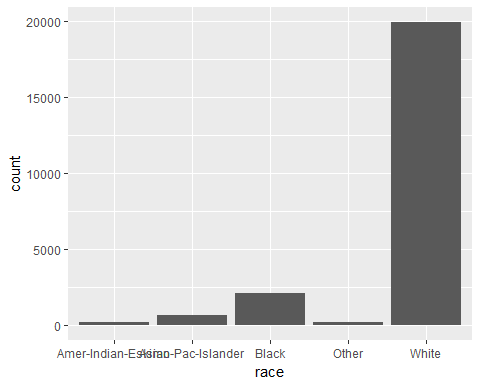
ggplot(data=census\_train, aes(x=marital.status)) +  
 geom\_bar() +  
 theme(axis.text.x = element\_text(angle = 30, hjust = 1))



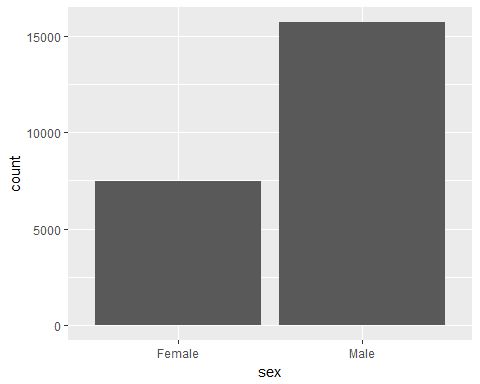
ggplot(data=census\_train, aes(x=occupation)) +  
 geom\_bar() +  
 theme(axis.text.x = element\_text(angle = 30, hjust = 1))



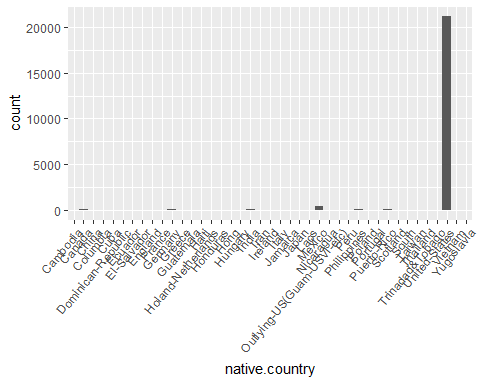
ggplot(data=census\_train, aes(x=race)) +  
 geom\_bar()



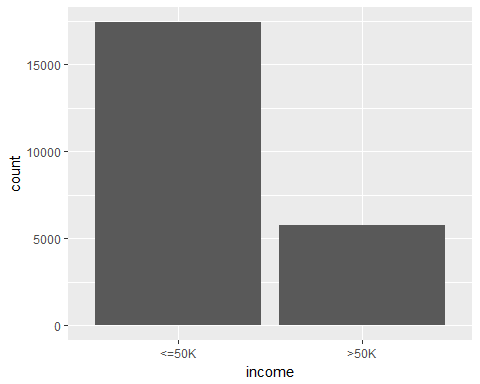
ggplot(data=census\_train, aes(x=sex)) +  
 geom\_bar()



ggplot(data=census\_train, aes(x=native.country)) +  
 geom\_bar() +  
 theme(axis.text.x = element\_text(angle = 50, hjust = 1))

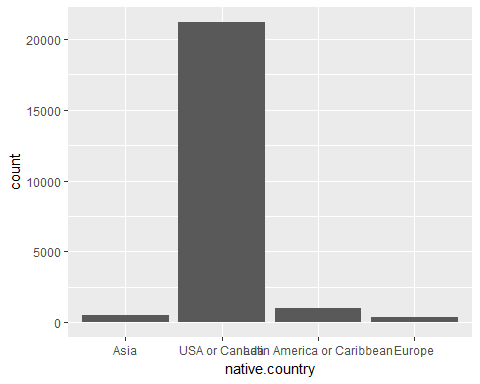


ggplot(data=census\_train, aes(x=income)) +  
 geom\_bar()



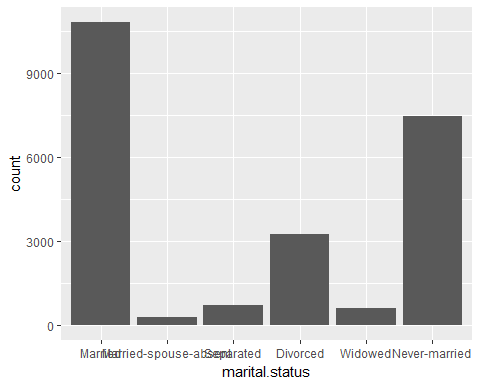
Concerning the native.country variable, the number of people native to the United States dwarfs all other possible values. We will collapse all non-US levels into an “Other” category.

census\_train <- census\_train %>% mutate(  
 native.country = fct\_recode(as.factor(native.country),  
 "USA or Canada" = "United-States", "USA or Canada" = "Canada",  
 "USA or Canada" = "Outlying-US(Guam-USVI-etc)",  
 "Europe" = "France", "Europe" = "Italy", "Europe" = "Poland",  
 "Europe" = "Scotland", "Europe" = "Germany", "Europe" = "Portugal",  
 "Europe" = "Yugoslavia", "Europe" = "England", "Europe" = "Greece",  
 "Europe" = "Holand-Netherlands", "Europe" = "Hungary",  
 "Europe" = "Ireland", "Asia" = "Cambodia", "Asia" = "India",  
 "Asia" = "Laos", "Asia" = "Thailand", "Asia" = "Vietnam",  
 "Asia" = "Hong", "Asia" = "Iran", "Asia" = "China",  
 "Asia" = "Japan", "Asia" = "Philippines", "Asia" = "Taiwan",  
 "Asia" = "South", "Latin America or Caribbean" = "Columbia",  
 "Latin America or Caribbean" = "Ecuador",  
 "Latin America or Caribbean" = "Guatemala",  
 "Latin America or Caribbean" = "Honduras",  
 "Latin America or Caribbean" = "Cuba",  
 "Latin America or Caribbean" = "El-Salvador",  
 "Latin America or Caribbean" = "Haiti",   
 "Latin America or Caribbean" = "Jamaica",  
 "Latin America or Caribbean" = "Mexico",  
 "Latin America or Caribbean" = "Peru",  
 "Latin America or Caribbean" = "Trinadad&Tobago",  
 "Latin America or Caribbean" = "Dominican-Republic",  
 "Latin America or Caribbean" = "Nicaragua",  
 "Latin America or Caribbean" = "Puerto-Rico"))  
  
ggplot(data=census\_train, aes(x=native.country)) +  
 geom\_bar()



Concerining the marital.status variable, we will collapse the trivial distinctions of Married-AF-spouse and Married-civ-spouse into a single category called Married.

census\_train <- census\_train %>% mutate(  
 marital.status = fct\_recode(as.factor(marital.status),  
 "Married" = "Married-AF-spouse", "Married" = "Married-civ-spouse"),  
 marital.status = fct\_relevel(marital.status, "Married", "Married-spouse-absent", "Separated",  
 "Divorced", "Widowed"))  
  
ggplot(data=census\_train, aes(x=marital.status)) +  
 geom\_bar()



Concerning workclass, below we will verify that all 10 individuals in the Without-pay category do not make more than $50,000, and then remove them since they are so few and obviously not relevant to the prediction task at hand (i.e., people who have never worked cannot make more than zero dollars).

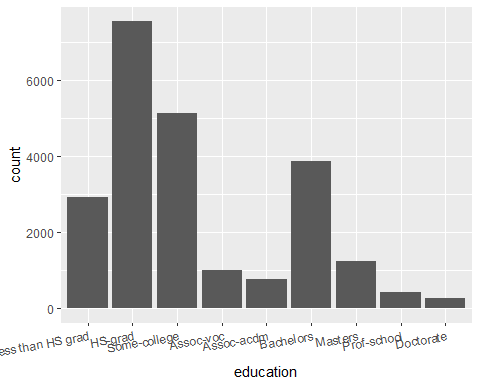
census\_train %>% filter(workclass == "Without-pay") %>% select(income)

## # A tibble: 10 x 1  
## income  
## <fct>   
## 1 <=50K   
## 2 <=50K   
## 3 <=50K   
## 4 <=50K   
## 5 <=50K   
## 6 <=50K   
## 7 <=50K   
## 8 <=50K   
## 9 <=50K   
## 10 <=50K

census\_train <- census\_train %>% filter(workclass != "Without-pay")  
census\_train <- census\_train %>% droplevels()

Concerning education, we will group together all subjects who seem not to have graduated high school or completed any college into a level called Less than HS grad.

census\_train <- census\_train %>% mutate(  
 education = fct\_collapse(as.factor(education),  
 "Less than HS grad" = c("Preschool", "1st-4th", "5th-6th", "7th-8th",  
 "9th", "10th", "11th", "12th")),  
 education = fct\_relevel(education, "Less than HS grad", "HS-grad", "Some-college", "Assoc-voc",  
 "Assoc-acdm", "Bachelors", "Masters", "Prof-school", "Doctorate"))  
ggplot(data=census\_train, aes(x=education)) +  
 geom\_bar() +  
 theme(axis.text.x = element\_text(angle = 10, hjust = 1))



We will now inspect the income value of the six subjects whose occupation value is Armed-Forces:

census\_train %>% filter(occupation == "Armed-Forces") %>% select(income)

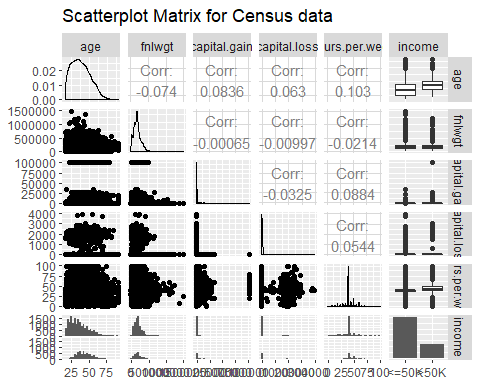
## # A tibble: 6 x 1  
## income  
## <fct>   
## 1 <=50K   
## 2 <=50K   
## 3 <=50K   
## 4 <=50K   
## 5 >50K   
## 6 <=50K

We see that they are not all the same, and because this level does not easily collapse into any other level in occupation, we will keep this level as-is.

We will now look at a scatterplot matrix of the quantitative predictors along with the outcome:

GGally::ggpairs(dplyr::select(census\_train, age, fnlwgt,  
 capital.gain,  
 capital.loss, hours.per.week, income),  
 title = "Scatterplot Matrix for Census data",  
 cardinality\_threshold = 41)

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



There do not seem to be any concerning trends.

Now we’re going to look at tables of each categorical variable with the income variable.

table(census\_train$workclass, census\_train$income)

##   
## <=50K >50K  
## Federal-gov 440 283  
## Local-gov 1125 457  
## Private 13391 3745  
## Self-emp-inc 367 460  
## Self-emp-not-inc 1375 541  
## State-gov 723 254

table(census\_train$education, census\_train$income)

##   
## <=50K >50K  
## Less than HS grad 2741 173  
## HS-grad 6308 1241  
## Some-college 4104 1018  
## Assoc-voc 733 265  
## Assoc-acdm 576 195  
## Bachelors 2259 1610  
## Masters 525 706  
## Prof-school 100 325  
## Doctorate 75 207

table(census\_train$marital.status, census\_train$income)

##   
## <=50K >50K  
## Married 5891 4924  
## Married-spouse-absent 265 22  
## Separated 666 55  
## Divorced 2907 347  
## Widowed 571 57  
## Never-married 7121 335

table(census\_train$occupation, census\_train$income)

##   
## <=50K >50K  
## Adm-clerical 2472 376  
## Armed-Forces 5 1  
## Craft-repair 2407 702  
## Exec-managerial 1595 1493  
## Farming-fishing 668 94  
## Handlers-cleaners 968 68  
## Machine-op-inspct 1327 176  
## Other-service 2379 98  
## Priv-house-serv 107 1  
## Prof-specialty 1730 1381  
## Protective-serv 327 156  
## Sales 1998 732  
## Tech-support 477 208  
## Transport-moving 961 254

table(census\_train$race, census\_train$income)

##   
## <=50K >50K  
## Amer-Indian-Eskimo 196 23  
## Asian-Pac-Islander 505 190  
## Black 1855 275  
## Other 171 15  
## White 14694 5237

table(census\_train$sex, census\_train$income)

##   
## <=50K >50K  
## Female 6610 853  
## Male 10811 4887

table(census\_train$native.country, census\_train$income)

##   
## <=50K >50K  
## Asia 369 165  
## USA or Canada 15823 5382  
## Latin America or Caribbean 959 84  
## Europe 270 109

No categorical variable seems determinative of income.

We now wish to build a model to predict whether or not a subject’s income is above $50,000 per year (presumably). In order to validate our model, we will further split census\_training into a training sample (80% of the data) and a query sample (20% of the data).

census\_train$id <- 1:nrow(census\_train)  
  
set.seed(0)  
census\_query <- census\_train %>% sample\_n(nrow(census\_train)/5)  
  
census\_train <- anti\_join(census\_train, census\_query, "id")

We use the kitchen sink model for logistic regression first.

log.reg1 = glm((income == ">50K") ~   
   
 # Quantitative  
 age + fnlwgt + capital.gain + capital.loss + hours.per.week +  
   
 # Qualitative  
 sex + race + marital.status + native.country + workclass +   
 education + occupation,  
 data=census\_train, family=binomial)

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

summary(log.reg1)

##   
## Call:  
## glm(formula = (income == ">50K") ~ age + fnlwgt + capital.gain +   
## capital.loss + hours.per.week + sex + race + marital.status +   
## native.country + workclass + education + occupation, family = binomial,   
## data = census\_train)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -4.3850 -0.5051 -0.2092 -0.0105 3.7144   
##   
## Coefficients:  
## Estimate Std. Error z value  
## (Intercept) -4.850e+00 4.286e-01 -11.318  
## age 2.664e-02 2.137e-03 12.466  
## fnlwgt 8.158e-07 2.216e-07 3.682  
## capital.gain 3.417e-04 1.385e-05 24.665  
## capital.loss 6.178e-04 4.874e-05 12.676  
## hours.per.week 3.005e-02 2.173e-03 13.831  
## sexMale 1.031e-01 6.900e-02 1.494  
## raceAsian-Pac-Islander 5.799e-01 3.558e-01 1.629  
## raceBlack 3.381e-01 3.068e-01 1.102  
## raceOther -8.056e-03 4.674e-01 -0.017  
## raceWhite 4.876e-01 2.928e-01 1.665  
## marital.statusMarried-spouse-absent -2.036e+00 2.878e-01 -7.074  
## marital.statusSeparated -2.116e+00 1.851e-01 -11.432  
## marital.statusDivorced -2.176e+00 8.736e-02 -24.911  
## marital.statusWidowed -2.354e+00 1.974e-01 -11.928  
## marital.statusNever-married -2.761e+00 8.450e-02 -32.677  
## native.countryUSA or Canada 2.681e-01 2.311e-01 1.160  
## native.countryLatin America or Caribbean -4.103e-01 2.810e-01 -1.460  
## native.countryEurope 3.461e-01 2.819e-01 1.228  
## workclassLocal-gov -6.858e-01 1.453e-01 -4.720  
## workclassPrivate -3.932e-01 1.211e-01 -3.248  
## workclassSelf-emp-inc -2.202e-01 1.601e-01 -1.375  
## workclassSelf-emp-not-inc -8.837e-01 1.423e-01 -6.210  
## workclassState-gov -7.899e-01 1.611e-01 -4.904  
## educationHS-grad 9.276e-01 1.075e-01 8.627  
## educationSome-college 1.254e+00 1.128e-01 11.115  
## educationAssoc-voc 1.410e+00 1.410e-01 10.002  
## educationAssoc-acdm 1.434e+00 1.565e-01 9.160  
## educationBachelors 2.056e+00 1.173e-01 17.527  
## educationMasters 2.478e+00 1.417e-01 17.491  
## educationProf-school 3.318e+00 2.100e-01 15.802  
## educationDoctorate 2.744e+00 2.301e-01 11.924  
## occupationArmed-Forces -9.757e+00 1.388e+02 -0.070  
## occupationCraft-repair -2.643e-02 1.007e-01 -0.262  
## occupationExec-managerial 7.008e-01 9.664e-02 7.252  
## occupationFarming-fishing -1.034e+00 1.770e-01 -5.842  
## occupationHandlers-cleaners -7.030e-01 1.786e-01 -3.935  
## occupationMachine-op-inspct -5.032e-01 1.329e-01 -3.787  
## occupationOther-service -9.172e-01 1.510e-01 -6.073  
## occupationPriv-house-serv -3.900e+00 2.434e+00 -1.602  
## occupationProf-specialty 3.867e-01 1.023e-01 3.779  
## occupationProtective-serv 5.777e-01 1.607e-01 3.595  
## occupationSales 1.419e-01 1.038e-01 1.366  
## occupationTech-support 5.646e-01 1.403e-01 4.023  
## occupationTransport-moving -7.982e-02 1.242e-01 -0.643  
## Pr(>|z|)   
## (Intercept) < 2e-16 \*\*\*  
## age < 2e-16 \*\*\*  
## fnlwgt 0.000232 \*\*\*  
## capital.gain < 2e-16 \*\*\*  
## capital.loss < 2e-16 \*\*\*  
## hours.per.week < 2e-16 \*\*\*  
## sexMale 0.135241   
## raceAsian-Pac-Islander 0.103208   
## raceBlack 0.270493   
## raceOther 0.986248   
## raceWhite 0.095832 .   
## marital.statusMarried-spouse-absent 1.50e-12 \*\*\*  
## marital.statusSeparated < 2e-16 \*\*\*  
## marital.statusDivorced < 2e-16 \*\*\*  
## marital.statusWidowed < 2e-16 \*\*\*  
## marital.statusNever-married < 2e-16 \*\*\*  
## native.countryUSA or Canada 0.245892   
## native.countryLatin America or Caribbean 0.144282   
## native.countryEurope 0.219475   
## workclassLocal-gov 2.36e-06 \*\*\*  
## workclassPrivate 0.001163 \*\*   
## workclassSelf-emp-inc 0.169060   
## workclassSelf-emp-not-inc 5.31e-10 \*\*\*  
## workclassState-gov 9.39e-07 \*\*\*  
## educationHS-grad < 2e-16 \*\*\*  
## educationSome-college < 2e-16 \*\*\*  
## educationAssoc-voc < 2e-16 \*\*\*  
## educationAssoc-acdm < 2e-16 \*\*\*  
## educationBachelors < 2e-16 \*\*\*  
## educationMasters < 2e-16 \*\*\*  
## educationProf-school < 2e-16 \*\*\*  
## educationDoctorate < 2e-16 \*\*\*  
## occupationArmed-Forces 0.943970   
## occupationCraft-repair 0.792946   
## occupationExec-managerial 4.12e-13 \*\*\*  
## occupationFarming-fishing 5.16e-09 \*\*\*  
## occupationHandlers-cleaners 8.31e-05 \*\*\*  
## occupationMachine-op-inspct 0.000153 \*\*\*  
## occupationOther-service 1.25e-09 \*\*\*  
## occupationPriv-house-serv 0.109106   
## occupationProf-specialty 0.000157 \*\*\*  
## occupationProtective-serv 0.000325 \*\*\*  
## occupationSales 0.171788   
## occupationTech-support 5.75e-05 \*\*\*  
## occupationTransport-moving 0.520468   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 20810 on 18528 degrees of freedom  
## Residual deviance: 12103 on 18484 degrees of freedom  
## AIC: 12193  
##   
## Number of Fisher Scoring iterations: 11

log.reg.probabilities1 = predict(log.reg1, census\_query, type="response")  
log.reg.predictions1 = ifelse(log.reg.probabilities1 > 0.5, ">50K", "<=50K")  
mean(log.reg.predictions1 != census\_query$income)

## [1] 0.1429188

Now we’ll do LDA

lda1 = lda((income == ">50K") ~ age + workclass + fnlwgt + education +   
 marital.status + occupation + race + sex + capital.gain +   
 capital.loss + hours.per.week + native.country, data=census\_train)  
  
lda.predictions.text1 = predict(lda1, census\_query)  
lda.predictions1 = ifelse(lda.predictions.text1$class == "TRUE", ">50K", "<=50K")  
  
mean(lda.predictions1 != census\_query$income)

## [1] 0.1530656

train.X = cbind(census\_train$age, census\_train$workclass, census\_train$fnlwgt,   
 census\_train$education, census\_train$marital.status,   
 census\_train$occupation, census\_train$race, census\_train$sex,  
 census\_train$capital.gain, census\_train$capital.loss,  
 census\_train$hours.per.week, census\_train$native.country)  
query.X = cbind(census\_query$age, census\_query$workclass, census\_query$fnlwgt,   
 census\_query$education, census\_query$marital.status,   
 census\_query$occupation, census\_query$race, census\_query$sex,  
 census\_query$capital.gain, census\_query$capital.loss,  
 census\_query$hours.per.week, census\_query$native.country)  
  
set.seed(1)  
knn1 = knn(train.X, query.X, census\_train$income.num, k = 1)  
mean(knn1 != census\_query$income.num)

## [1] 0.2722366

This procedure has also yielded 18.2% test error. We’ll try a series of KNN procedures with different values for k: 3, 10, 50, and 100.

set.seed(3)  
knn3 = knn(train.X, query.X, census\_train$income.num, k = 3)  
mean(knn3 != census\_query$income.num)

## [1] 0.2428756

set.seed(5)  
knn5 = knn(train.X, query.X, census\_train$income.num, k = 5)  
mean(knn5 != census\_query$income.num)

## [1] 0.2266839

set.seed(10)  
knn10 = knn(train.X, query.X, census\_train$income.num, k = 10)  
mean(knn10 != census\_query$income.num)

## [1] 0.2091969

set.seed(50)  
knn50 = knn(train.X, query.X, census\_train$income.num, k = 50)  
mean(knn50 != census\_query$income.num)

## [1] 0.2100604

set.seed(100)  
knn100 = knn(train.X, query.X, census\_train$income.num, k = 100)  
mean(knn100 != census\_query$income.num)

## [1] 0.2182642

The KNN method has a much higher error rate in its classification when compared to logistic regression and LDA. We’ll look at those:

log.reg2 = glm((income == ">50K") ~   
   
 # Quantitative  
 rcs(age, 5) + rcs(fnlwgt, 5) + rcs(capital.gain, 5) +  
 rcs(capital.loss, 5) + rcs(hours.per.week, 5) +   
   
 # Qualitative  
 sex + race + marital.status + native.country + workclass +   
 education + occupation,  
 data=census\_train, family=binomial)  
summary(log.reg2)

##   
## Call:  
## glm(formula = (income == ">50K") ~ rcs(age, 5) + rcs(fnlwgt,   
## 5) + rcs(capital.gain, 5) + rcs(capital.loss, 5) + rcs(hours.per.week,   
## 5) + sex + race + marital.status + native.country + workclass +   
## education + occupation, family = binomial, data = census\_train)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -5.3143 -0.4869 -0.1847 -0.0022 4.0078   
##   
## Coefficients:  
## Estimate Std. Error z value  
## (Intercept) -1.169e+01 1.162e+00 -10.061  
## rcs(age, 5)age 2.691e-01 4.028e-02 6.680  
## rcs(age, 5)age' -8.991e-01 2.234e-01 -4.025  
## rcs(age, 5)age'' 2.037e+00 5.938e-01 3.430  
## rcs(age, 5)age''' -1.538e+00 5.251e-01 -2.928  
## rcs(fnlwgt, 5)fnlwgt -1.195e-07 1.710e-06 -0.070  
## rcs(fnlwgt, 5)fnlwgt' 1.110e-05 1.145e-05 0.969  
## rcs(fnlwgt, 5)fnlwgt'' -5.975e-05 4.893e-05 -1.221  
## rcs(fnlwgt, 5)fnlwgt''' 9.480e-05 6.686e-05 1.418  
## rcs(capital.gain, 5)capital.gain -8.689e-04 1.651e-04 -5.262  
## rcs(capital.gain, 5)capital.gain' 8.771e-01 1.624e-01 5.402  
## rcs(capital.gain, 5)capital.gain'' -1.504e+00 3.220e-01 -4.670  
## rcs(capital.gain, 5)capital.gain''' 6.918e-01 2.182e-01 3.171  
## rcs(capital.loss, 5)capital.loss -2.764e-03 1.039e-03 -2.659  
## rcs(capital.loss, 5)capital.loss' 7.332e-03 2.777e-03 2.641  
## rcs(capital.loss, 5)capital.loss'' -4.520e-02 3.960e-02 -1.142  
## rcs(capital.loss, 5)capital.loss''' -4.716e-02 3.788e-01 -0.124  
## rcs(hours.per.week, 5)hours.per.week 2.447e-02 7.982e-03 3.066  
## rcs(hours.per.week, 5)hours.per.week' 4.890e-02 1.556e-02 3.142  
## rcs(hours.per.week, 5)hours.per.week'' -6.421e-01 1.370e-01 -4.686  
## sexMale 4.592e-02 7.156e-02 0.642  
## raceAsian-Pac-Islander 7.898e-01 3.735e-01 2.115  
## raceBlack 4.025e-01 3.232e-01 1.245  
## raceOther 1.539e-01 4.925e-01 0.313  
## raceWhite 6.094e-01 3.092e-01 1.971  
## marital.statusMarried-spouse-absent -2.060e+00 2.906e-01 -7.087  
## marital.statusSeparated -2.214e+00 1.882e-01 -11.760  
## marital.statusDivorced -2.311e+00 9.012e-02 -25.645  
## marital.statusWidowed -2.103e+00 2.003e-01 -10.497  
## marital.statusNever-married -2.551e+00 8.999e-02 -28.345  
## native.countryUSA or Canada 3.254e-01 2.376e-01 1.369  
## native.countryLatin America or Caribbean -3.690e-01 2.872e-01 -1.285  
## native.countryEurope 4.347e-01 2.889e-01 1.505  
## workclassLocal-gov -6.762e-01 1.491e-01 -4.537  
## workclassPrivate -3.544e-01 1.243e-01 -2.851  
## workclassSelf-emp-inc -1.618e-01 1.646e-01 -0.983  
## workclassSelf-emp-not-inc -7.744e-01 1.461e-01 -5.301  
## workclassState-gov -7.635e-01 1.650e-01 -4.627  
## educationHS-grad 7.933e-01 1.093e-01 7.260  
## educationSome-college 1.129e+00 1.148e-01 9.828  
## educationAssoc-voc 1.240e+00 1.439e-01 8.620  
## educationAssoc-acdm 1.227e+00 1.603e-01 7.656  
## educationBachelors 1.869e+00 1.197e-01 15.620  
## educationMasters 2.243e+00 1.451e-01 15.458  
## educationProf-school 3.179e+00 2.123e-01 14.972  
## educationDoctorate 2.587e+00 2.333e-01 11.086  
## occupationArmed-Forces -9.878e+00 1.359e+02 -0.073  
## occupationCraft-repair -9.615e-02 1.035e-01 -0.929  
## occupationExec-managerial 6.613e-01 9.962e-02 6.637  
## occupationFarming-fishing -9.064e-01 1.802e-01 -5.031  
## occupationHandlers-cleaners -6.746e-01 1.833e-01 -3.681  
## occupationMachine-op-inspct -5.057e-01 1.352e-01 -3.739  
## occupationOther-service -8.711e-01 1.544e-01 -5.642  
## occupationPriv-house-serv -6.330e+00 9.605e+00 -0.659  
## occupationProf-specialty 3.574e-01 1.055e-01 3.388  
## occupationProtective-serv 6.608e-01 1.657e-01 3.989  
## occupationSales 1.573e-01 1.077e-01 1.461  
## occupationTech-support 5.868e-01 1.451e-01 4.044  
## occupationTransport-moving -1.132e-01 1.276e-01 -0.887  
## Pr(>|z|)   
## (Intercept) < 2e-16 \*\*\*  
## rcs(age, 5)age 2.39e-11 \*\*\*  
## rcs(age, 5)age' 5.70e-05 \*\*\*  
## rcs(age, 5)age'' 0.000604 \*\*\*  
## rcs(age, 5)age''' 0.003410 \*\*   
## rcs(fnlwgt, 5)fnlwgt 0.944279   
## rcs(fnlwgt, 5)fnlwgt' 0.332677   
## rcs(fnlwgt, 5)fnlwgt'' 0.222049   
## rcs(fnlwgt, 5)fnlwgt''' 0.156227   
## rcs(capital.gain, 5)capital.gain 1.42e-07 \*\*\*  
## rcs(capital.gain, 5)capital.gain' 6.58e-08 \*\*\*  
## rcs(capital.gain, 5)capital.gain'' 3.01e-06 \*\*\*  
## rcs(capital.gain, 5)capital.gain''' 0.001522 \*\*   
## rcs(capital.loss, 5)capital.loss 0.007837 \*\*   
## rcs(capital.loss, 5)capital.loss' 0.008274 \*\*   
## rcs(capital.loss, 5)capital.loss'' 0.253622   
## rcs(capital.loss, 5)capital.loss''' 0.900939   
## rcs(hours.per.week, 5)hours.per.week 0.002172 \*\*   
## rcs(hours.per.week, 5)hours.per.week' 0.001680 \*\*   
## rcs(hours.per.week, 5)hours.per.week'' 2.79e-06 \*\*\*  
## sexMale 0.521025   
## raceAsian-Pac-Islander 0.034432 \*   
## raceBlack 0.212967   
## raceOther 0.754657   
## raceWhite 0.048710 \*   
## marital.statusMarried-spouse-absent 1.37e-12 \*\*\*  
## marital.statusSeparated < 2e-16 \*\*\*  
## marital.statusDivorced < 2e-16 \*\*\*  
## marital.statusWidowed < 2e-16 \*\*\*  
## marital.statusNever-married < 2e-16 \*\*\*  
## native.countryUSA or Canada 0.170867   
## native.countryLatin America or Caribbean 0.198902   
## native.countryEurope 0.132394   
## workclassLocal-gov 5.71e-06 \*\*\*  
## workclassPrivate 0.004362 \*\*   
## workclassSelf-emp-inc 0.325410   
## workclassSelf-emp-not-inc 1.15e-07 \*\*\*  
## workclassState-gov 3.71e-06 \*\*\*  
## educationHS-grad 3.86e-13 \*\*\*  
## educationSome-college < 2e-16 \*\*\*  
## educationAssoc-voc < 2e-16 \*\*\*  
## educationAssoc-acdm 1.92e-14 \*\*\*  
## educationBachelors < 2e-16 \*\*\*  
## educationMasters < 2e-16 \*\*\*  
## educationProf-school < 2e-16 \*\*\*  
## educationDoctorate < 2e-16 \*\*\*  
## occupationArmed-Forces 0.942044   
## occupationCraft-repair 0.352952   
## occupationExec-managerial 3.19e-11 \*\*\*  
## occupationFarming-fishing 4.88e-07 \*\*\*  
## occupationHandlers-cleaners 0.000233 \*\*\*  
## occupationMachine-op-inspct 0.000184 \*\*\*  
## occupationOther-service 1.68e-08 \*\*\*  
## occupationPriv-house-serv 0.509846   
## occupationProf-specialty 0.000704 \*\*\*  
## occupationProtective-serv 6.64e-05 \*\*\*  
## occupationSales 0.144038   
## occupationTech-support 5.24e-05 \*\*\*  
## occupationTransport-moving 0.375155   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 20810 on 18528 degrees of freedom  
## Residual deviance: 11504 on 18470 degrees of freedom  
## AIC: 11622  
##   
## Number of Fisher Scoring iterations: 11

log.reg.probabilities2 = predict(log.reg2, census\_query, type="response")  
log.reg.predictions2 = ifelse(log.reg.probabilities2 > 0.5, ">50K", "<=50K")  
mean(log.reg.predictions2 != census\_query$income)

## [1] 0.1379534

We have improved just a little. It seems like the splines on the fnlwgt didn’t do much, nor did many of the interactions.

log.reg3 = glm((income == ">50K") ~   
   
 # Quantitative  
 rcs(age, 5) + rcs(fnlwgt, 3) + rcs(capital.gain, 5) +  
 rcs(capital.loss, 3) + rcs(hours.per.week, 5) +   
   
 # Qualitative  
 sex \* race + marital.status + workclass +   
 education + occupation +  
   
 # Interactions  
 marital.status %ia% age + workclass %ia% age +  
 education %ia% age + hours.per.week %ia% workclass +   
 hours.per.week %ia% occupation + age %ia% hours.per.week,  
 data=census\_train, family=binomial)  
summary(log.reg3)

##   
## Call:  
## glm(formula = (income == ">50K") ~ rcs(age, 5) + rcs(fnlwgt,   
## 3) + rcs(capital.gain, 5) + rcs(capital.loss, 3) + rcs(hours.per.week,   
## 5) + sex \* race + marital.status + workclass + education +   
## occupation + marital.status %ia% age + workclass %ia% age +   
## education %ia% age + hours.per.week %ia% workclass + hours.per.week %ia%   
## occupation + age %ia% hours.per.week, family = binomial,   
## data = census\_train)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -5.3132 -0.4871 -0.1752 -0.0025 3.7372   
##   
## Coefficients:  
## Estimate  
## (Intercept) -1.150e+01  
## rcs(age, 5)age 2.783e-01  
## rcs(age, 5)age' -8.413e-01  
## rcs(age, 5)age'' 1.917e+00  
## rcs(age, 5)age''' -1.502e+00  
## rcs(fnlwgt, 3)fnlwgt 6.593e-07  
## rcs(fnlwgt, 3)fnlwgt' 1.062e-07  
## rcs(capital.gain, 5)capital.gain -8.746e-04  
## rcs(capital.gain, 5)capital.gain' 8.870e-01  
## rcs(capital.gain, 5)capital.gain'' -1.521e+00  
## rcs(capital.gain, 5)capital.gain''' 6.995e-01  
## rcs(capital.loss, 3)capital.loss -9.680e-05  
## rcs(capital.loss, 3)capital.loss' 1.002e-03  
## rcs(hours.per.week, 5)hours.per.week 4.051e-02  
## rcs(hours.per.week, 5)hours.per.week' 3.959e-02  
## rcs(hours.per.week, 5)hours.per.week'' -5.973e-01  
## sexMale -1.283e+00  
## raceAsian-Pac-Islander -1.567e-01  
## raceBlack -7.271e-01  
## raceOther -6.257e-01  
## raceWhite -3.037e-01  
## marital.statusMarried-spouse-absent -6.592e+00  
## marital.statusSeparated -3.302e+00  
## marital.statusDivorced -2.391e+00  
## marital.statusWidowed -5.712e+00  
## marital.statusNever-married -2.896e+00  
## workclassLocal-gov 1.166e+00  
## workclassPrivate 3.782e-01  
## workclassSelf-emp-inc -7.340e-02  
## workclassSelf-emp-not-inc 1.196e+00  
## workclassState-gov -1.944e+00  
## educationHS-grad 2.928e-01  
## educationSome-college 8.950e-01  
## educationAssoc-voc 1.192e+00  
## educationAssoc-acdm 1.446e+00  
## educationBachelors 1.970e+00  
## educationMasters 2.610e+00  
## educationProf-school 4.102e+00  
## educationDoctorate 3.334e+00  
## occupationArmed-Forces -9.729e+00  
## occupationCraft-repair -3.883e-01  
## occupationExec-managerial 6.708e-01  
## occupationFarming-fishing -1.825e+00  
## occupationHandlers-cleaners -1.194e+00  
## occupationMachine-op-inspct -8.558e-01  
## occupationOther-service -1.233e+00  
## occupationPriv-house-serv -5.068e+00  
## occupationProf-specialty 1.336e+00  
## occupationProtective-serv 4.862e-01  
## occupationSales 1.743e-01  
## occupationTech-support 8.694e-01  
## occupationTransport-moving -4.346e-01  
## marital.status %ia% agemarital.status=Married-spouse-absent \* age 9.528e-02  
## marital.status %ia% agemarital.status=Separated \* age 2.445e-02  
## marital.status %ia% agemarital.status=Divorced \* age 1.551e-03  
## marital.status %ia% agemarital.status=Widowed \* age 6.379e-02  
## marital.status %ia% agemarital.status=Never-married \* age 8.242e-03  
## workclass %ia% ageworkclass=Local-gov \* age -4.495e-02  
## workclass %ia% ageworkclass=Private \* age -1.915e-02  
## workclass %ia% ageworkclass=Self-emp-inc \* age 7.876e-03  
## workclass %ia% ageworkclass=Self-emp-not-inc \* age -3.318e-02  
## workclass %ia% ageworkclass=State-gov \* age 3.684e-03  
## education %ia% ageeducation=HS-grad \* age 1.276e-02  
## education %ia% ageeducation=Some-college \* age 6.524e-03  
## education %ia% ageeducation=Assoc-voc \* age 2.348e-03  
## education %ia% ageeducation=Assoc-acdm \* age -4.232e-03  
## education %ia% ageeducation=Bachelors \* age -8.358e-04  
## education %ia% ageeducation=Masters \* age -6.311e-03  
## education %ia% ageeducation=Prof-school \* age -1.779e-02  
## education %ia% ageeducation=Doctorate \* age -1.375e-02  
## hours.per.week %ia% workclasshours.per.week \* workclass=Local-gov 4.547e-03  
## hours.per.week %ia% workclasshours.per.week \* workclass=Private 3.724e-03  
## hours.per.week %ia% workclasshours.per.week \* workclass=Self-emp-inc -8.906e-03  
## hours.per.week %ia% workclasshours.per.week \* workclass=Self-emp-not-inc -9.004e-03  
## hours.per.week %ia% workclasshours.per.week \* workclass=State-gov 2.581e-02  
## hours.per.week %ia% occupationhours.per.week \* occupation=Armed-Forces 3.826e-03  
## hours.per.week %ia% occupationhours.per.week \* occupation=Craft-repair 6.391e-03  
## hours.per.week %ia% occupationhours.per.week \* occupation=Exec-managerial -4.879e-04  
## hours.per.week %ia% occupationhours.per.week \* occupation=Farming-fishing 1.716e-02  
## hours.per.week %ia% occupationhours.per.week \* occupation=Handlers-cleaners 1.173e-02  
## hours.per.week %ia% occupationhours.per.week \* occupation=Machine-op-inspct 7.430e-03  
## hours.per.week %ia% occupationhours.per.week \* occupation=Other-service 7.779e-03  
## hours.per.week %ia% occupationhours.per.week \* occupation=Priv-house-serv -2.936e-02  
## hours.per.week %ia% occupationhours.per.week \* occupation=Prof-specialty -2.342e-02  
## hours.per.week %ia% occupationhours.per.week \* occupation=Protective-serv 2.174e-03  
## hours.per.week %ia% occupationhours.per.week \* occupation=Sales -6.149e-04  
## hours.per.week %ia% occupationhours.per.week \* occupation=Tech-support -6.502e-03  
## hours.per.week %ia% occupationhours.per.week \* occupation=Transport-moving 6.681e-03  
## age %ia% hours.per.week -1.696e-04  
## sexMale:raceAsian-Pac-Islander 1.111e+00  
## sexMale:raceBlack 1.680e+00  
## sexMale:raceOther 8.564e-01  
## sexMale:raceWhite 1.330e+00  
## Std. Error  
## (Intercept) 1.621e+00  
## rcs(age, 5)age 4.482e-02  
## rcs(age, 5)age' 2.295e-01  
## rcs(age, 5)age'' 6.083e-01  
## rcs(age, 5)age''' 5.370e-01  
## rcs(fnlwgt, 3)fnlwgt 5.880e-07  
## rcs(fnlwgt, 3)fnlwgt' 6.629e-07  
## rcs(capital.gain, 5)capital.gain 1.661e-04  
## rcs(capital.gain, 5)capital.gain' 1.634e-01  
## rcs(capital.gain, 5)capital.gain'' 3.240e-01  
## rcs(capital.gain, 5)capital.gain''' 2.193e-01  
## rcs(capital.loss, 3)capital.loss 1.673e-04  
## rcs(capital.loss, 3)capital.loss' 2.331e-04  
## rcs(hours.per.week, 5)hours.per.week 2.079e-02  
## rcs(hours.per.week, 5)hours.per.week' 1.630e-02  
## rcs(hours.per.week, 5)hours.per.week'' 1.414e-01  
## sexMale 6.176e-01  
## raceAsian-Pac-Islander 5.619e-01  
## raceBlack 5.065e-01  
## raceOther 8.898e-01  
## raceWhite 4.790e-01  
## marital.statusMarried-spouse-absent 1.622e+00  
## marital.statusSeparated 9.109e-01  
## marital.statusDivorced 4.451e-01  
## marital.statusWidowed 1.198e+00  
## marital.statusNever-married 3.482e-01  
## workclassLocal-gov 1.022e+00  
## workclassPrivate 8.793e-01  
## workclassSelf-emp-inc 1.120e+00  
## workclassSelf-emp-not-inc 9.703e-01  
## workclassState-gov 1.149e+00  
## educationHS-grad 4.567e-01  
## educationSome-college 4.720e-01  
## educationAssoc-voc 6.108e-01  
## educationAssoc-acdm 6.886e-01  
## educationBachelors 4.722e-01  
## educationMasters 5.814e-01  
## educationProf-school 7.859e-01  
## educationDoctorate 9.259e-01  
## occupationArmed-Forces 8.460e+02  
## occupationCraft-repair 5.167e-01  
## occupationExec-managerial 4.800e-01  
## occupationFarming-fishing 7.106e-01  
## occupationHandlers-cleaners 8.872e-01  
## occupationMachine-op-inspct 7.161e-01  
## occupationOther-service 6.480e-01  
## occupationPriv-house-serv 3.862e+01  
## occupationProf-specialty 4.638e-01  
## occupationProtective-serv 6.941e-01  
## occupationSales 5.049e-01  
## occupationTech-support 6.693e-01  
## occupationTransport-moving 5.548e-01  
## marital.status %ia% agemarital.status=Married-spouse-absent \* age 3.251e-02  
## marital.status %ia% agemarital.status=Separated \* age 2.041e-02  
## marital.status %ia% agemarital.status=Divorced \* age 9.829e-03  
## marital.status %ia% agemarital.status=Widowed \* age 2.017e-02  
## marital.status %ia% agemarital.status=Never-married \* age 8.993e-03  
## workclass %ia% ageworkclass=Local-gov \* age 1.426e-02  
## workclass %ia% ageworkclass=Private \* age 1.184e-02  
## workclass %ia% ageworkclass=Self-emp-inc \* age 1.518e-02  
## workclass %ia% ageworkclass=Self-emp-not-inc \* age 1.344e-02  
## workclass %ia% ageworkclass=State-gov \* age 1.578e-02  
## education %ia% ageeducation=HS-grad \* age 9.624e-03  
## education %ia% ageeducation=Some-college \* age 1.004e-02  
## education %ia% ageeducation=Assoc-voc \* age 1.364e-02  
## education %ia% ageeducation=Assoc-acdm \* age 1.584e-02  
## education %ia% ageeducation=Bachelors \* age 1.007e-02  
## education %ia% ageeducation=Masters \* age 1.236e-02  
## education %ia% ageeducation=Prof-school \* age 1.701e-02  
## education %ia% ageeducation=Doctorate \* age 1.931e-02  
## hours.per.week %ia% workclasshours.per.week \* workclass=Local-gov 1.608e-02  
## hours.per.week %ia% workclasshours.per.week \* workclass=Private 1.397e-02  
## hours.per.week %ia% workclasshours.per.week \* workclass=Self-emp-inc 1.616e-02  
## hours.per.week %ia% workclasshours.per.week \* workclass=Self-emp-not-inc 1.476e-02  
## hours.per.week %ia% workclasshours.per.week \* workclass=State-gov 1.844e-02  
## hours.per.week %ia% occupationhours.per.week \* occupation=Armed-Forces 1.847e+01  
## hours.per.week %ia% occupationhours.per.week \* occupation=Craft-repair 1.212e-02  
## hours.per.week %ia% occupationhours.per.week \* occupation=Exec-managerial 1.126e-02  
## hours.per.week %ia% occupationhours.per.week \* occupation=Farming-fishing 1.434e-02  
## hours.per.week %ia% occupationhours.per.week \* occupation=Handlers-cleaners 2.036e-02  
## hours.per.week %ia% occupationhours.per.week \* occupation=Machine-op-inspct 1.658e-02  
## hours.per.week %ia% occupationhours.per.week \* occupation=Other-service 1.516e-02  
## hours.per.week %ia% occupationhours.per.week \* occupation=Priv-house-serv 1.075e+00  
## hours.per.week %ia% occupationhours.per.week \* occupation=Prof-specialty 1.104e-02  
## hours.per.week %ia% occupationhours.per.week \* occupation=Protective-serv 1.574e-02  
## hours.per.week %ia% occupationhours.per.week \* occupation=Sales 1.174e-02  
## hours.per.week %ia% occupationhours.per.week \* occupation=Tech-support 1.594e-02  
## hours.per.week %ia% occupationhours.per.week \* occupation=Transport-moving 1.252e-02  
## age %ia% hours.per.week 2.054e-04  
## sexMale:raceAsian-Pac-Islander 7.010e-01  
## sexMale:raceBlack 6.505e-01  
## sexMale:raceOther 1.062e+00  
## sexMale:raceWhite 6.191e-01  
## z value  
## (Intercept) -7.090  
## rcs(age, 5)age 6.209  
## rcs(age, 5)age' -3.665  
## rcs(age, 5)age'' 3.151  
## rcs(age, 5)age''' -2.798  
## rcs(fnlwgt, 3)fnlwgt 1.121  
## rcs(fnlwgt, 3)fnlwgt' 0.160  
## rcs(capital.gain, 5)capital.gain -5.265  
## rcs(capital.gain, 5)capital.gain' 5.429  
## rcs(capital.gain, 5)capital.gain'' -4.695  
## rcs(capital.gain, 5)capital.gain''' 3.189  
## rcs(capital.loss, 3)capital.loss -0.579  
## rcs(capital.loss, 3)capital.loss' 4.298  
## rcs(hours.per.week, 5)hours.per.week 1.949  
## rcs(hours.per.week, 5)hours.per.week' 2.428  
## rcs(hours.per.week, 5)hours.per.week'' -4.223  
## sexMale -2.077  
## raceAsian-Pac-Islander -0.279  
## raceBlack -1.436  
## raceOther -0.703  
## raceWhite -0.634  
## marital.statusMarried-spouse-absent -4.063  
## marital.statusSeparated -3.624  
## marital.statusDivorced -5.372  
## marital.statusWidowed -4.767  
## marital.statusNever-married -8.316  
## workclassLocal-gov 1.141  
## workclassPrivate 0.430  
## workclassSelf-emp-inc -0.066  
## workclassSelf-emp-not-inc 1.233  
## workclassState-gov -1.692  
## educationHS-grad 0.641  
## educationSome-college 1.896  
## educationAssoc-voc 1.952  
## educationAssoc-acdm 2.100  
## educationBachelors 4.172  
## educationMasters 4.489  
## educationProf-school 5.220  
## educationDoctorate 3.601  
## occupationArmed-Forces -0.012  
## occupationCraft-repair -0.751  
## occupationExec-managerial 1.397  
## occupationFarming-fishing -2.568  
## occupationHandlers-cleaners -1.346  
## occupationMachine-op-inspct -1.195  
## occupationOther-service -1.902  
## occupationPriv-house-serv -0.131  
## occupationProf-specialty 2.881  
## occupationProtective-serv 0.701  
## occupationSales 0.345  
## occupationTech-support 1.299  
## occupationTransport-moving -0.783  
## marital.status %ia% agemarital.status=Married-spouse-absent \* age 2.931  
## marital.status %ia% agemarital.status=Separated \* age 1.198  
## marital.status %ia% agemarital.status=Divorced \* age 0.158  
## marital.status %ia% agemarital.status=Widowed \* age 3.163  
## marital.status %ia% agemarital.status=Never-married \* age 0.916  
## workclass %ia% ageworkclass=Local-gov \* age -3.151  
## workclass %ia% ageworkclass=Private \* age -1.618  
## workclass %ia% ageworkclass=Self-emp-inc \* age 0.519  
## workclass %ia% ageworkclass=Self-emp-not-inc \* age -2.469  
## workclass %ia% ageworkclass=State-gov \* age 0.233  
## education %ia% ageeducation=HS-grad \* age 1.326  
## education %ia% ageeducation=Some-college \* age 0.650  
## education %ia% ageeducation=Assoc-voc \* age 0.172  
## education %ia% ageeducation=Assoc-acdm \* age -0.267  
## education %ia% ageeducation=Bachelors \* age -0.083  
## education %ia% ageeducation=Masters \* age -0.510  
## education %ia% ageeducation=Prof-school \* age -1.046  
## education %ia% ageeducation=Doctorate \* age -0.712  
## hours.per.week %ia% workclasshours.per.week \* workclass=Local-gov 0.283  
## hours.per.week %ia% workclasshours.per.week \* workclass=Private 0.267  
## hours.per.week %ia% workclasshours.per.week \* workclass=Self-emp-inc -0.551  
## hours.per.week %ia% workclasshours.per.week \* workclass=Self-emp-not-inc -0.610  
## hours.per.week %ia% workclasshours.per.week \* workclass=State-gov 1.400  
## hours.per.week %ia% occupationhours.per.week \* occupation=Armed-Forces 0.000  
## hours.per.week %ia% occupationhours.per.week \* occupation=Craft-repair 0.527  
## hours.per.week %ia% occupationhours.per.week \* occupation=Exec-managerial -0.043  
## hours.per.week %ia% occupationhours.per.week \* occupation=Farming-fishing 1.197  
## hours.per.week %ia% occupationhours.per.week \* occupation=Handlers-cleaners 0.576  
## hours.per.week %ia% occupationhours.per.week \* occupation=Machine-op-inspct 0.448  
## hours.per.week %ia% occupationhours.per.week \* occupation=Other-service 0.513  
## hours.per.week %ia% occupationhours.per.week \* occupation=Priv-house-serv -0.027  
## hours.per.week %ia% occupationhours.per.week \* occupation=Prof-specialty -2.122  
## hours.per.week %ia% occupationhours.per.week \* occupation=Protective-serv 0.138  
## hours.per.week %ia% occupationhours.per.week \* occupation=Sales -0.052  
## hours.per.week %ia% occupationhours.per.week \* occupation=Tech-support -0.408  
## hours.per.week %ia% occupationhours.per.week \* occupation=Transport-moving 0.534  
## age %ia% hours.per.week -0.826  
## sexMale:raceAsian-Pac-Islander 1.585  
## sexMale:raceBlack 2.582  
## sexMale:raceOther 0.806  
## sexMale:raceWhite 2.148  
## Pr(>|z|)  
## (Intercept) 1.34e-12  
## rcs(age, 5)age 5.33e-10  
## rcs(age, 5)age' 0.000247  
## rcs(age, 5)age'' 0.001630  
## rcs(age, 5)age''' 0.005146  
## rcs(fnlwgt, 3)fnlwgt 0.262180  
## rcs(fnlwgt, 3)fnlwgt' 0.872654  
## rcs(capital.gain, 5)capital.gain 1.40e-07  
## rcs(capital.gain, 5)capital.gain' 5.66e-08  
## rcs(capital.gain, 5)capital.gain'' 2.67e-06  
## rcs(capital.gain, 5)capital.gain''' 0.001428  
## rcs(capital.loss, 3)capital.loss 0.562784  
## rcs(capital.loss, 3)capital.loss' 1.72e-05  
## rcs(hours.per.week, 5)hours.per.week 0.051316  
## rcs(hours.per.week, 5)hours.per.week' 0.015181  
## rcs(hours.per.week, 5)hours.per.week'' 2.41e-05  
## sexMale 0.037778  
## raceAsian-Pac-Islander 0.780375  
## raceBlack 0.151124  
## raceOther 0.481913  
## raceWhite 0.525976  
## marital.statusMarried-spouse-absent 4.84e-05  
## marital.statusSeparated 0.000290  
## marital.statusDivorced 7.81e-08  
## marital.statusWidowed 1.87e-06  
## marital.statusNever-married < 2e-16  
## workclassLocal-gov 0.254042  
## workclassPrivate 0.667124  
## workclassSelf-emp-inc 0.947760  
## workclassSelf-emp-not-inc 0.217604  
## workclassState-gov 0.090724  
## educationHS-grad 0.521417  
## educationSome-college 0.057948  
## educationAssoc-voc 0.050967  
## educationAssoc-acdm 0.035710  
## educationBachelors 3.02e-05  
## educationMasters 7.15e-06  
## educationProf-school 1.79e-07  
## educationDoctorate 0.000318  
## occupationArmed-Forces 0.990824  
## occupationCraft-repair 0.452419  
## occupationExec-managerial 0.162265  
## occupationFarming-fishing 0.010216  
## occupationHandlers-cleaners 0.178420  
## occupationMachine-op-inspct 0.232072  
## occupationOther-service 0.057131  
## occupationPriv-house-serv 0.895595  
## occupationProf-specialty 0.003961  
## occupationProtective-serv 0.483598  
## occupationSales 0.729882  
## occupationTech-support 0.193977  
## occupationTransport-moving 0.433405  
## marital.status %ia% agemarital.status=Married-spouse-absent \* age 0.003379  
## marital.status %ia% agemarital.status=Separated \* age 0.230984  
## marital.status %ia% agemarital.status=Divorced \* age 0.874606  
## marital.status %ia% agemarital.status=Widowed \* age 0.001560  
## marital.status %ia% agemarital.status=Never-married \* age 0.359409  
## workclass %ia% ageworkclass=Local-gov \* age 0.001625  
## workclass %ia% ageworkclass=Private \* age 0.105687  
## workclass %ia% ageworkclass=Self-emp-inc \* age 0.603815  
## workclass %ia% ageworkclass=Self-emp-not-inc \* age 0.013542  
## workclass %ia% ageworkclass=State-gov \* age 0.815378  
## education %ia% ageeducation=HS-grad \* age 0.184778  
## education %ia% ageeducation=Some-college \* age 0.515788  
## education %ia% ageeducation=Assoc-voc \* age 0.863309  
## education %ia% ageeducation=Assoc-acdm \* age 0.789346  
## education %ia% ageeducation=Bachelors \* age 0.933839  
## education %ia% ageeducation=Masters \* age 0.609735  
## education %ia% ageeducation=Prof-school \* age 0.295597  
## education %ia% ageeducation=Doctorate \* age 0.476517  
## hours.per.week %ia% workclasshours.per.week \* workclass=Local-gov 0.777332  
## hours.per.week %ia% workclasshours.per.week \* workclass=Private 0.789836  
## hours.per.week %ia% workclasshours.per.week \* workclass=Self-emp-inc 0.581624  
## hours.per.week %ia% workclasshours.per.week \* workclass=Self-emp-not-inc 0.541742  
## hours.per.week %ia% workclasshours.per.week \* workclass=State-gov 0.161566  
## hours.per.week %ia% occupationhours.per.week \* occupation=Armed-Forces 0.999835  
## hours.per.week %ia% occupationhours.per.week \* occupation=Craft-repair 0.597891  
## hours.per.week %ia% occupationhours.per.week \* occupation=Exec-managerial 0.965432  
## hours.per.week %ia% occupationhours.per.week \* occupation=Farming-fishing 0.231373  
## hours.per.week %ia% occupationhours.per.week \* occupation=Handlers-cleaners 0.564320  
## hours.per.week %ia% occupationhours.per.week \* occupation=Machine-op-inspct 0.654099  
## hours.per.week %ia% occupationhours.per.week \* occupation=Other-service 0.607869  
## hours.per.week %ia% occupationhours.per.week \* occupation=Priv-house-serv 0.978207  
## hours.per.week %ia% occupationhours.per.week \* occupation=Prof-specialty 0.033824  
## hours.per.week %ia% occupationhours.per.week \* occupation=Protective-serv 0.890141  
## hours.per.week %ia% occupationhours.per.week \* occupation=Sales 0.958224  
## hours.per.week %ia% occupationhours.per.week \* occupation=Tech-support 0.683342  
## hours.per.week %ia% occupationhours.per.week \* occupation=Transport-moving 0.593637  
## age %ia% hours.per.week 0.409023  
## sexMale:raceAsian-Pac-Islander 0.112933  
## sexMale:raceBlack 0.009817  
## sexMale:raceOther 0.420121  
## sexMale:raceWhite 0.031695  
##   
## (Intercept) \*\*\*  
## rcs(age, 5)age \*\*\*  
## rcs(age, 5)age' \*\*\*  
## rcs(age, 5)age'' \*\*   
## rcs(age, 5)age''' \*\*   
## rcs(fnlwgt, 3)fnlwgt   
## rcs(fnlwgt, 3)fnlwgt'   
## rcs(capital.gain, 5)capital.gain \*\*\*  
## rcs(capital.gain, 5)capital.gain' \*\*\*  
## rcs(capital.gain, 5)capital.gain'' \*\*\*  
## rcs(capital.gain, 5)capital.gain''' \*\*   
## rcs(capital.loss, 3)capital.loss   
## rcs(capital.loss, 3)capital.loss' \*\*\*  
## rcs(hours.per.week, 5)hours.per.week .   
## rcs(hours.per.week, 5)hours.per.week' \*   
## rcs(hours.per.week, 5)hours.per.week'' \*\*\*  
## sexMale \*   
## raceAsian-Pac-Islander   
## raceBlack   
## raceOther   
## raceWhite   
## marital.statusMarried-spouse-absent \*\*\*  
## marital.statusSeparated \*\*\*  
## marital.statusDivorced \*\*\*  
## marital.statusWidowed \*\*\*  
## marital.statusNever-married \*\*\*  
## workclassLocal-gov   
## workclassPrivate   
## workclassSelf-emp-inc   
## workclassSelf-emp-not-inc   
## workclassState-gov .   
## educationHS-grad   
## educationSome-college .   
## educationAssoc-voc .   
## educationAssoc-acdm \*   
## educationBachelors \*\*\*  
## educationMasters \*\*\*  
## educationProf-school \*\*\*  
## educationDoctorate \*\*\*  
## occupationArmed-Forces   
## occupationCraft-repair   
## occupationExec-managerial   
## occupationFarming-fishing \*   
## occupationHandlers-cleaners   
## occupationMachine-op-inspct   
## occupationOther-service .   
## occupationPriv-house-serv   
## occupationProf-specialty \*\*   
## occupationProtective-serv   
## occupationSales   
## occupationTech-support   
## occupationTransport-moving   
## marital.status %ia% agemarital.status=Married-spouse-absent \* age \*\*   
## marital.status %ia% agemarital.status=Separated \* age   
## marital.status %ia% agemarital.status=Divorced \* age   
## marital.status %ia% agemarital.status=Widowed \* age \*\*   
## marital.status %ia% agemarital.status=Never-married \* age   
## workclass %ia% ageworkclass=Local-gov \* age \*\*   
## workclass %ia% ageworkclass=Private \* age   
## workclass %ia% ageworkclass=Self-emp-inc \* age   
## workclass %ia% ageworkclass=Self-emp-not-inc \* age \*   
## workclass %ia% ageworkclass=State-gov \* age   
## education %ia% ageeducation=HS-grad \* age   
## education %ia% ageeducation=Some-college \* age   
## education %ia% ageeducation=Assoc-voc \* age   
## education %ia% ageeducation=Assoc-acdm \* age   
## education %ia% ageeducation=Bachelors \* age   
## education %ia% ageeducation=Masters \* age   
## education %ia% ageeducation=Prof-school \* age   
## education %ia% ageeducation=Doctorate \* age   
## hours.per.week %ia% workclasshours.per.week \* workclass=Local-gov   
## hours.per.week %ia% workclasshours.per.week \* workclass=Private   
## hours.per.week %ia% workclasshours.per.week \* workclass=Self-emp-inc   
## hours.per.week %ia% workclasshours.per.week \* workclass=Self-emp-not-inc   
## hours.per.week %ia% workclasshours.per.week \* workclass=State-gov   
## hours.per.week %ia% occupationhours.per.week \* occupation=Armed-Forces   
## hours.per.week %ia% occupationhours.per.week \* occupation=Craft-repair   
## hours.per.week %ia% occupationhours.per.week \* occupation=Exec-managerial   
## hours.per.week %ia% occupationhours.per.week \* occupation=Farming-fishing   
## hours.per.week %ia% occupationhours.per.week \* occupation=Handlers-cleaners   
## hours.per.week %ia% occupationhours.per.week \* occupation=Machine-op-inspct   
## hours.per.week %ia% occupationhours.per.week \* occupation=Other-service   
## hours.per.week %ia% occupationhours.per.week \* occupation=Priv-house-serv   
## hours.per.week %ia% occupationhours.per.week \* occupation=Prof-specialty \*   
## hours.per.week %ia% occupationhours.per.week \* occupation=Protective-serv   
## hours.per.week %ia% occupationhours.per.week \* occupation=Sales   
## hours.per.week %ia% occupationhours.per.week \* occupation=Tech-support   
## hours.per.week %ia% occupationhours.per.week \* occupation=Transport-moving   
## age %ia% hours.per.week   
## sexMale:raceAsian-Pac-Islander   
## sexMale:raceBlack \*\*   
## sexMale:raceOther   
## sexMale:raceWhite \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 20810 on 18528 degrees of freedom  
## Residual deviance: 11457 on 18436 degrees of freedom  
## AIC: 11643  
##   
## Number of Fisher Scoring iterations: 11

log.reg.probabilities3 = predict(log.reg3, census\_query, type="response")  
log.reg.predictions3 = ifelse(log.reg.probabilities3 > 0.5, ">50K", "<=50K")  
mean(log.reg.predictions3 != census\_query$income)

## [1] 0.1381693

log.reg4 = glm((income == ">50K") ~   
   
 # Quantitative  
 rcs(age, 5) + fnlwgt + rcs(capital.gain, 5) +  
 rcs(capital.loss, 5) + rcs(hours.per.week, 5) +   
   
 # Qualitative  
 sex \* race + marital.status + workclass +   
 education + occupation +  
   
 # Interactions  
 marital.status %ia% age + workclass %ia% age +  
 education %ia% age + hours.per.week %ia% workclass +   
 hours.per.week %ia% occupation + age %ia% hours.per.week,  
 data=census\_train, family=binomial)  
summary(log.reg4)

##   
## Call:  
## glm(formula = (income == ">50K") ~ rcs(age, 5) + fnlwgt + rcs(capital.gain,   
## 5) + rcs(capital.loss, 5) + rcs(hours.per.week, 5) + sex \*   
## race + marital.status + workclass + education + occupation +   
## marital.status %ia% age + workclass %ia% age + education %ia%   
## age + hours.per.week %ia% workclass + hours.per.week %ia%   
## occupation + age %ia% hours.per.week, family = binomial,   
## data = census\_train)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -5.3140 -0.4859 -0.1759 -0.0025 3.7385   
##   
## Coefficients:  
## Estimate  
## (Intercept) -1.170e+01  
## rcs(age, 5)age 2.835e-01  
## rcs(age, 5)age' -8.605e-01  
## rcs(age, 5)age'' 1.965e+00  
## rcs(age, 5)age''' -1.539e+00  
## fnlwgt 7.605e-07  
## rcs(capital.gain, 5)capital.gain -8.730e-04  
## rcs(capital.gain, 5)capital.gain' 8.850e-01  
## rcs(capital.gain, 5)capital.gain'' -1.517e+00  
## rcs(capital.gain, 5)capital.gain''' 6.976e-01  
## rcs(capital.loss, 5)capital.loss -2.655e-03  
## rcs(capital.loss, 5)capital.loss' 7.033e-03  
## rcs(capital.loss, 5)capital.loss'' -4.039e-02  
## rcs(capital.loss, 5)capital.loss''' -9.824e-02  
## rcs(hours.per.week, 5)hours.per.week 4.123e-02  
## rcs(hours.per.week, 5)hours.per.week' 3.980e-02  
## rcs(hours.per.week, 5)hours.per.week'' -6.011e-01  
## sexMale -1.280e+00  
## raceAsian-Pac-Islander -1.621e-01  
## raceBlack -7.263e-01  
## raceOther -5.912e-01  
## raceWhite -2.886e-01  
## marital.statusMarried-spouse-absent -6.629e+00  
## marital.statusSeparated -3.292e+00  
## marital.statusDivorced -2.418e+00  
## marital.statusWidowed -5.688e+00  
## marital.statusNever-married -2.905e+00  
## workclassLocal-gov 1.256e+00  
## workclassPrivate 4.542e-01  
## workclassSelf-emp-inc 1.954e-02  
## workclassSelf-emp-not-inc 1.242e+00  
## workclassState-gov -1.877e+00  
## educationHS-grad 2.976e-01  
## educationSome-college 9.243e-01  
## educationAssoc-voc 1.192e+00  
## educationAssoc-acdm 1.447e+00  
## educationBachelors 1.975e+00  
## educationMasters 2.604e+00  
## educationProf-school 4.076e+00  
## educationDoctorate 3.300e+00  
## occupationArmed-Forces -9.692e+00  
## occupationCraft-repair -3.807e-01  
## occupationExec-managerial 6.811e-01  
## occupationFarming-fishing -1.836e+00  
## occupationHandlers-cleaners -1.185e+00  
## occupationMachine-op-inspct -8.245e-01  
## occupationOther-service -1.215e+00  
## occupationPriv-house-serv -5.045e+00  
## occupationProf-specialty 1.337e+00  
## occupationProtective-serv 4.862e-01  
## occupationSales 1.960e-01  
## occupationTech-support 8.193e-01  
## occupationTransport-moving -4.292e-01  
## marital.status %ia% agemarital.status=Married-spouse-absent \* age 9.620e-02  
## marital.status %ia% agemarital.status=Separated \* age 2.455e-02  
## marital.status %ia% agemarital.status=Divorced \* age 2.803e-03  
## marital.status %ia% agemarital.status=Widowed \* age 6.334e-02  
## marital.status %ia% agemarital.status=Never-married \* age 9.069e-03  
## workclass %ia% ageworkclass=Local-gov \* age -4.612e-02  
## workclass %ia% ageworkclass=Private \* age -2.047e-02  
## workclass %ia% ageworkclass=Self-emp-inc \* age 6.486e-03  
## workclass %ia% ageworkclass=Self-emp-not-inc \* age -3.437e-02  
## workclass %ia% ageworkclass=State-gov \* age 2.840e-03  
## education %ia% ageeducation=HS-grad \* age 1.227e-02  
## education %ia% ageeducation=Some-college \* age 5.639e-03  
## education %ia% ageeducation=Assoc-voc \* age 2.009e-03  
## education %ia% ageeducation=Assoc-acdm \* age -4.702e-03  
## education %ia% ageeducation=Bachelors \* age -1.324e-03  
## education %ia% ageeducation=Masters \* age -6.807e-03  
## education %ia% ageeducation=Prof-school \* age -1.736e-02  
## education %ia% ageeducation=Doctorate \* age -1.344e-02  
## hours.per.week %ia% workclasshours.per.week \* workclass=Local-gov 3.479e-03  
## hours.per.week %ia% workclasshours.per.week \* workclass=Private 3.188e-03  
## hours.per.week %ia% workclasshours.per.week \* workclass=Self-emp-inc -9.662e-03  
## hours.per.week %ia% workclasshours.per.week \* workclass=Self-emp-not-inc -9.129e-03  
## hours.per.week %ia% workclasshours.per.week \* workclass=State-gov 2.512e-02  
## hours.per.week %ia% occupationhours.per.week \* occupation=Armed-Forces 2.904e-03  
## hours.per.week %ia% occupationhours.per.week \* occupation=Craft-repair 6.314e-03  
## hours.per.week %ia% occupationhours.per.week \* occupation=Exec-managerial -6.842e-04  
## hours.per.week %ia% occupationhours.per.week \* occupation=Farming-fishing 1.750e-02  
## hours.per.week %ia% occupationhours.per.week \* occupation=Handlers-cleaners 1.159e-02  
## hours.per.week %ia% occupationhours.per.week \* occupation=Machine-op-inspct 6.892e-03  
## hours.per.week %ia% occupationhours.per.week \* occupation=Other-service 7.745e-03  
## hours.per.week %ia% occupationhours.per.week \* occupation=Priv-house-serv -2.978e-02  
## hours.per.week %ia% occupationhours.per.week \* occupation=Prof-specialty -2.346e-02  
## hours.per.week %ia% occupationhours.per.week \* occupation=Protective-serv 2.229e-03  
## hours.per.week %ia% occupationhours.per.week \* occupation=Sales -1.106e-03  
## hours.per.week %ia% occupationhours.per.week \* occupation=Tech-support -5.348e-03  
## hours.per.week %ia% occupationhours.per.week \* occupation=Transport-moving 6.640e-03  
## age %ia% hours.per.week -1.699e-04  
## sexMale:raceAsian-Pac-Islander 1.116e+00  
## sexMale:raceBlack 1.669e+00  
## sexMale:raceOther 7.999e-01  
## sexMale:raceWhite 1.318e+00  
## Std. Error  
## (Intercept) 1.623e+00  
## rcs(age, 5)age 4.492e-02  
## rcs(age, 5)age' 2.302e-01  
## rcs(age, 5)age'' 6.099e-01  
## rcs(age, 5)age''' 5.382e-01  
## fnlwgt 2.290e-07  
## rcs(capital.gain, 5)capital.gain 1.658e-04  
## rcs(capital.gain, 5)capital.gain' 1.631e-01  
## rcs(capital.gain, 5)capital.gain'' 3.234e-01  
## rcs(capital.gain, 5)capital.gain''' 2.190e-01  
## rcs(capital.loss, 5)capital.loss 1.035e-03  
## rcs(capital.loss, 5)capital.loss' 2.770e-03  
## rcs(capital.loss, 5)capital.loss'' 3.968e-02  
## rcs(capital.loss, 5)capital.loss''' 3.819e-01  
## rcs(hours.per.week, 5)hours.per.week 2.082e-02  
## rcs(hours.per.week, 5)hours.per.week' 1.634e-02  
## rcs(hours.per.week, 5)hours.per.week'' 1.417e-01  
## sexMale 6.179e-01  
## raceAsian-Pac-Islander 5.613e-01  
## raceBlack 5.055e-01  
## raceOther 8.906e-01  
## raceWhite 4.779e-01  
## marital.statusMarried-spouse-absent 1.614e+00  
## marital.statusSeparated 9.054e-01  
## marital.statusDivorced 4.420e-01  
## marital.statusWidowed 1.203e+00  
## marital.statusNever-married 3.474e-01  
## workclassLocal-gov 1.024e+00  
## workclassPrivate 8.810e-01  
## workclassSelf-emp-inc 1.122e+00  
## workclassSelf-emp-not-inc 9.723e-01  
## workclassState-gov 1.152e+00  
## educationHS-grad 4.553e-01  
## educationSome-college 4.708e-01  
## educationAssoc-voc 6.105e-01  
## educationAssoc-acdm 6.891e-01  
## educationBachelors 4.709e-01  
## educationMasters 5.813e-01  
## educationProf-school 7.850e-01  
## educationDoctorate 9.275e-01  
## occupationArmed-Forces 8.476e+02  
## occupationCraft-repair 5.166e-01  
## occupationExec-managerial 4.797e-01  
## occupationFarming-fishing 7.090e-01  
## occupationHandlers-cleaners 8.883e-01  
## occupationMachine-op-inspct 7.160e-01  
## occupationOther-service 6.461e-01  
## occupationPriv-house-serv 3.834e+01  
## occupationProf-specialty 4.638e-01  
## occupationProtective-serv 6.940e-01  
## occupationSales 5.054e-01  
## occupationTech-support 6.733e-01  
## occupationTransport-moving 5.560e-01  
## marital.status %ia% agemarital.status=Married-spouse-absent \* age 3.234e-02  
## marital.status %ia% agemarital.status=Separated \* age 2.033e-02  
## marital.status %ia% agemarital.status=Divorced \* age 9.751e-03  
## marital.status %ia% agemarital.status=Widowed \* age 2.025e-02  
## marital.status %ia% agemarital.status=Never-married \* age 8.969e-03  
## workclass %ia% ageworkclass=Local-gov \* age 1.425e-02  
## workclass %ia% ageworkclass=Private \* age 1.181e-02  
## workclass %ia% ageworkclass=Self-emp-inc \* age 1.516e-02  
## workclass %ia% ageworkclass=Self-emp-not-inc \* age 1.342e-02  
## workclass %ia% ageworkclass=State-gov \* age 1.578e-02  
## education %ia% ageeducation=HS-grad \* age 9.591e-03  
## education %ia% ageeducation=Some-college \* age 1.001e-02  
## education %ia% ageeducation=Assoc-voc \* age 1.363e-02  
## education %ia% ageeducation=Assoc-acdm \* age 1.585e-02  
## education %ia% ageeducation=Bachelors \* age 1.004e-02  
## education %ia% ageeducation=Masters \* age 1.236e-02  
## education %ia% ageeducation=Prof-school \* age 1.699e-02  
## education %ia% ageeducation=Doctorate \* age 1.934e-02  
## hours.per.week %ia% workclasshours.per.week \* workclass=Local-gov 1.613e-02  
## hours.per.week %ia% workclasshours.per.week \* workclass=Private 1.403e-02  
## hours.per.week %ia% workclasshours.per.week \* workclass=Self-emp-inc 1.621e-02  
## hours.per.week %ia% workclasshours.per.week \* workclass=Self-emp-not-inc 1.481e-02  
## hours.per.week %ia% workclasshours.per.week \* workclass=State-gov 1.849e-02  
## hours.per.week %ia% occupationhours.per.week \* occupation=Armed-Forces 1.851e+01  
## hours.per.week %ia% occupationhours.per.week \* occupation=Craft-repair 1.212e-02  
## hours.per.week %ia% occupationhours.per.week \* occupation=Exec-managerial 1.125e-02  
## hours.per.week %ia% occupationhours.per.week \* occupation=Farming-fishing 1.430e-02  
## hours.per.week %ia% occupationhours.per.week \* occupation=Handlers-cleaners 2.039e-02  
## hours.per.week %ia% occupationhours.per.week \* occupation=Machine-op-inspct 1.659e-02  
## hours.per.week %ia% occupationhours.per.week \* occupation=Other-service 1.512e-02  
## hours.per.week %ia% occupationhours.per.week \* occupation=Priv-house-serv 1.067e+00  
## hours.per.week %ia% occupationhours.per.week \* occupation=Prof-specialty 1.104e-02  
## hours.per.week %ia% occupationhours.per.week \* occupation=Protective-serv 1.573e-02  
## hours.per.week %ia% occupationhours.per.week \* occupation=Sales 1.175e-02  
## hours.per.week %ia% occupationhours.per.week \* occupation=Tech-support 1.603e-02  
## hours.per.week %ia% occupationhours.per.week \* occupation=Transport-moving 1.254e-02  
## age %ia% hours.per.week 2.058e-04  
## sexMale:raceAsian-Pac-Islander 7.018e-01  
## sexMale:raceBlack 6.509e-01  
## sexMale:raceOther 1.065e+00  
## sexMale:raceWhite 6.195e-01  
## z value  
## (Intercept) -7.205  
## rcs(age, 5)age 6.312  
## rcs(age, 5)age' -3.739  
## rcs(age, 5)age'' 3.222  
## rcs(age, 5)age''' -2.860  
## fnlwgt 3.321  
## rcs(capital.gain, 5)capital.gain -5.264  
## rcs(capital.gain, 5)capital.gain' 5.426  
## rcs(capital.gain, 5)capital.gain'' -4.692  
## rcs(capital.gain, 5)capital.gain''' 3.185  
## rcs(capital.loss, 5)capital.loss -2.565  
## rcs(capital.loss, 5)capital.loss' 2.539  
## rcs(capital.loss, 5)capital.loss'' -1.018  
## rcs(capital.loss, 5)capital.loss''' -0.257  
## rcs(hours.per.week, 5)hours.per.week 1.980  
## rcs(hours.per.week, 5)hours.per.week' 2.437  
## rcs(hours.per.week, 5)hours.per.week'' -4.242  
## sexMale -2.072  
## raceAsian-Pac-Islander -0.289  
## raceBlack -1.437  
## raceOther -0.664  
## raceWhite -0.604  
## marital.statusMarried-spouse-absent -4.107  
## marital.statusSeparated -3.635  
## marital.statusDivorced -5.470  
## marital.statusWidowed -4.727  
## marital.statusNever-married -8.364  
## workclassLocal-gov 1.226  
## workclassPrivate 0.516  
## workclassSelf-emp-inc 0.017  
## workclassSelf-emp-not-inc 1.277  
## workclassState-gov -1.630  
## educationHS-grad 0.654  
## educationSome-college 1.963  
## educationAssoc-voc 1.952  
## educationAssoc-acdm 2.099  
## educationBachelors 4.195  
## educationMasters 4.479  
## educationProf-school 5.192  
## educationDoctorate 3.557  
## occupationArmed-Forces -0.011  
## occupationCraft-repair -0.737  
## occupationExec-managerial 1.420  
## occupationFarming-fishing -2.590  
## occupationHandlers-cleaners -1.334  
## occupationMachine-op-inspct -1.151  
## occupationOther-service -1.880  
## occupationPriv-house-serv -0.132  
## occupationProf-specialty 2.882  
## occupationProtective-serv 0.701  
## occupationSales 0.388  
## occupationTech-support 1.217  
## occupationTransport-moving -0.772  
## marital.status %ia% agemarital.status=Married-spouse-absent \* age 2.975  
## marital.status %ia% agemarital.status=Separated \* age 1.208  
## marital.status %ia% agemarital.status=Divorced \* age 0.287  
## marital.status %ia% agemarital.status=Widowed \* age 3.128  
## marital.status %ia% agemarital.status=Never-married \* age 1.011  
## workclass %ia% ageworkclass=Local-gov \* age -3.236  
## workclass %ia% ageworkclass=Private \* age -1.734  
## workclass %ia% ageworkclass=Self-emp-inc \* age 0.428  
## workclass %ia% ageworkclass=Self-emp-not-inc \* age -2.561  
## workclass %ia% ageworkclass=State-gov \* age 0.180  
## education %ia% ageeducation=HS-grad \* age 1.280  
## education %ia% ageeducation=Some-college \* age 0.563  
## education %ia% ageeducation=Assoc-voc \* age 0.147  
## education %ia% ageeducation=Assoc-acdm \* age -0.297  
## education %ia% ageeducation=Bachelors \* age -0.132  
## education %ia% ageeducation=Masters \* age -0.551  
## education %ia% ageeducation=Prof-school \* age -1.022  
## education %ia% ageeducation=Doctorate \* age -0.695  
## hours.per.week %ia% workclasshours.per.week \* workclass=Local-gov 0.216  
## hours.per.week %ia% workclasshours.per.week \* workclass=Private 0.227  
## hours.per.week %ia% workclasshours.per.week \* workclass=Self-emp-inc -0.596  
## hours.per.week %ia% workclasshours.per.week \* workclass=Self-emp-not-inc -0.616  
## hours.per.week %ia% workclasshours.per.week \* workclass=State-gov 1.358  
## hours.per.week %ia% occupationhours.per.week \* occupation=Armed-Forces 0.000  
## hours.per.week %ia% occupationhours.per.week \* occupation=Craft-repair 0.521  
## hours.per.week %ia% occupationhours.per.week \* occupation=Exec-managerial -0.061  
## hours.per.week %ia% occupationhours.per.week \* occupation=Farming-fishing 1.223  
## hours.per.week %ia% occupationhours.per.week \* occupation=Handlers-cleaners 0.569  
## hours.per.week %ia% occupationhours.per.week \* occupation=Machine-op-inspct 0.415  
## hours.per.week %ia% occupationhours.per.week \* occupation=Other-service 0.512  
## hours.per.week %ia% occupationhours.per.week \* occupation=Priv-house-serv -0.028  
## hours.per.week %ia% occupationhours.per.week \* occupation=Prof-specialty -2.125  
## hours.per.week %ia% occupationhours.per.week \* occupation=Protective-serv 0.142  
## hours.per.week %ia% occupationhours.per.week \* occupation=Sales -0.094  
## hours.per.week %ia% occupationhours.per.week \* occupation=Tech-support -0.334  
## hours.per.week %ia% occupationhours.per.week \* occupation=Transport-moving 0.529  
## age %ia% hours.per.week -0.826  
## sexMale:raceAsian-Pac-Islander 1.590  
## sexMale:raceBlack 2.564  
## sexMale:raceOther 0.751  
## sexMale:raceWhite 2.127  
## Pr(>|z|)  
## (Intercept) 5.82e-13  
## rcs(age, 5)age 2.76e-10  
## rcs(age, 5)age' 0.000185  
## rcs(age, 5)age'' 0.001272  
## rcs(age, 5)age''' 0.004232  
## fnlwgt 0.000897  
## rcs(capital.gain, 5)capital.gain 1.41e-07  
## rcs(capital.gain, 5)capital.gain' 5.76e-08  
## rcs(capital.gain, 5)capital.gain'' 2.71e-06  
## rcs(capital.gain, 5)capital.gain''' 0.001446  
## rcs(capital.loss, 5)capital.loss 0.010323  
## rcs(capital.loss, 5)capital.loss' 0.011104  
## rcs(capital.loss, 5)capital.loss'' 0.308747  
## rcs(capital.loss, 5)capital.loss''' 0.796993  
## rcs(hours.per.week, 5)hours.per.week 0.047655  
## rcs(hours.per.week, 5)hours.per.week' 0.014821  
## rcs(hours.per.week, 5)hours.per.week'' 2.22e-05  
## sexMale 0.038270  
## raceAsian-Pac-Islander 0.772792  
## raceBlack 0.150785  
## raceOther 0.506803  
## raceWhite 0.545930  
## marital.statusMarried-spouse-absent 4.01e-05  
## marital.statusSeparated 0.000278  
## marital.statusDivorced 4.50e-08  
## marital.statusWidowed 2.28e-06  
## marital.statusNever-married < 2e-16  
## workclassLocal-gov 0.220017  
## workclassPrivate 0.606185  
## workclassSelf-emp-inc 0.986106  
## workclassSelf-emp-not-inc 0.201614  
## workclassState-gov 0.103158  
## educationHS-grad 0.513406  
## educationSome-college 0.049603  
## educationAssoc-voc 0.050948  
## educationAssoc-acdm 0.035794  
## educationBachelors 2.73e-05  
## educationMasters 7.49e-06  
## educationProf-school 2.08e-07  
## educationDoctorate 0.000374  
## occupationArmed-Forces 0.990876  
## occupationCraft-repair 0.461159  
## occupationExec-managerial 0.155621  
## occupationFarming-fishing 0.009610  
## occupationHandlers-cleaners 0.182325  
## occupationMachine-op-inspct 0.249557  
## occupationOther-service 0.060089  
## occupationPriv-house-serv 0.895319  
## occupationProf-specialty 0.003948  
## occupationProtective-serv 0.483563  
## occupationSales 0.698083  
## occupationTech-support 0.223651  
## occupationTransport-moving 0.440224  
## marital.status %ia% agemarital.status=Married-spouse-absent \* age 0.002934  
## marital.status %ia% agemarital.status=Separated \* age 0.227157  
## marital.status %ia% agemarital.status=Divorced \* age 0.773786  
## marital.status %ia% agemarital.status=Widowed \* age 0.001763  
## marital.status %ia% agemarital.status=Never-married \* age 0.311945  
## workclass %ia% ageworkclass=Local-gov \* age 0.001211  
## workclass %ia% ageworkclass=Private \* age 0.082974  
## workclass %ia% ageworkclass=Self-emp-inc \* age 0.668698  
## workclass %ia% ageworkclass=Self-emp-not-inc \* age 0.010451  
## workclass %ia% ageworkclass=State-gov \* age 0.857200  
## education %ia% ageeducation=HS-grad \* age 0.200686  
## education %ia% ageeducation=Some-college \* age 0.573259  
## education %ia% ageeducation=Assoc-voc \* age 0.882869  
## education %ia% ageeducation=Assoc-acdm \* age 0.766681  
## education %ia% ageeducation=Bachelors \* age 0.895035  
## education %ia% ageeducation=Masters \* age 0.581730  
## education %ia% ageeducation=Prof-school \* age 0.306807  
## education %ia% ageeducation=Doctorate \* age 0.487149  
## hours.per.week %ia% workclasshours.per.week \* workclass=Local-gov 0.829280  
## hours.per.week %ia% workclasshours.per.week \* workclass=Private 0.820202  
## hours.per.week %ia% workclasshours.per.week \* workclass=Self-emp-inc 0.551078  
## hours.per.week %ia% workclasshours.per.week \* workclass=Self-emp-not-inc 0.537579  
## hours.per.week %ia% workclasshours.per.week \* workclass=State-gov 0.174317  
## hours.per.week %ia% occupationhours.per.week \* occupation=Armed-Forces 0.999875  
## hours.per.week %ia% occupationhours.per.week \* occupation=Craft-repair 0.602282  
## hours.per.week %ia% occupationhours.per.week \* occupation=Exec-managerial 0.951510  
## hours.per.week %ia% occupationhours.per.week \* occupation=Farming-fishing 0.221247  
## hours.per.week %ia% occupationhours.per.week \* occupation=Handlers-cleaners 0.569671  
## hours.per.week %ia% occupationhours.per.week \* occupation=Machine-op-inspct 0.677777  
## hours.per.week %ia% occupationhours.per.week \* occupation=Other-service 0.608457  
## hours.per.week %ia% occupationhours.per.week \* occupation=Priv-house-serv 0.977731  
## hours.per.week %ia% occupationhours.per.week \* occupation=Prof-specialty 0.033558  
## hours.per.week %ia% occupationhours.per.week \* occupation=Protective-serv 0.887320  
## hours.per.week %ia% occupationhours.per.week \* occupation=Sales 0.924981  
## hours.per.week %ia% occupationhours.per.week \* occupation=Tech-support 0.738734  
## hours.per.week %ia% occupationhours.per.week \* occupation=Transport-moving 0.596475  
## age %ia% hours.per.week 0.408940  
## sexMale:raceAsian-Pac-Islander 0.111925  
## sexMale:raceBlack 0.010356  
## sexMale:raceOther 0.452479  
## sexMale:raceWhite 0.033434  
##   
## (Intercept) \*\*\*  
## rcs(age, 5)age \*\*\*  
## rcs(age, 5)age' \*\*\*  
## rcs(age, 5)age'' \*\*   
## rcs(age, 5)age''' \*\*   
## fnlwgt \*\*\*  
## rcs(capital.gain, 5)capital.gain \*\*\*  
## rcs(capital.gain, 5)capital.gain' \*\*\*  
## rcs(capital.gain, 5)capital.gain'' \*\*\*  
## rcs(capital.gain, 5)capital.gain''' \*\*   
## rcs(capital.loss, 5)capital.loss \*   
## rcs(capital.loss, 5)capital.loss' \*   
## rcs(capital.loss, 5)capital.loss''   
## rcs(capital.loss, 5)capital.loss'''   
## rcs(hours.per.week, 5)hours.per.week \*   
## rcs(hours.per.week, 5)hours.per.week' \*   
## rcs(hours.per.week, 5)hours.per.week'' \*\*\*  
## sexMale \*   
## raceAsian-Pac-Islander   
## raceBlack   
## raceOther   
## raceWhite   
## marital.statusMarried-spouse-absent \*\*\*  
## marital.statusSeparated \*\*\*  
## marital.statusDivorced \*\*\*  
## marital.statusWidowed \*\*\*  
## marital.statusNever-married \*\*\*  
## workclassLocal-gov   
## workclassPrivate   
## workclassSelf-emp-inc   
## workclassSelf-emp-not-inc   
## workclassState-gov   
## educationHS-grad   
## educationSome-college \*   
## educationAssoc-voc .   
## educationAssoc-acdm \*   
## educationBachelors \*\*\*  
## educationMasters \*\*\*  
## educationProf-school \*\*\*  
## educationDoctorate \*\*\*  
## occupationArmed-Forces   
## occupationCraft-repair   
## occupationExec-managerial   
## occupationFarming-fishing \*\*   
## occupationHandlers-cleaners   
## occupationMachine-op-inspct   
## occupationOther-service .   
## occupationPriv-house-serv   
## occupationProf-specialty \*\*   
## occupationProtective-serv   
## occupationSales   
## occupationTech-support   
## occupationTransport-moving   
## marital.status %ia% agemarital.status=Married-spouse-absent \* age \*\*   
## marital.status %ia% agemarital.status=Separated \* age   
## marital.status %ia% agemarital.status=Divorced \* age   
## marital.status %ia% agemarital.status=Widowed \* age \*\*   
## marital.status %ia% agemarital.status=Never-married \* age   
## workclass %ia% ageworkclass=Local-gov \* age \*\*   
## workclass %ia% ageworkclass=Private \* age .   
## workclass %ia% ageworkclass=Self-emp-inc \* age   
## workclass %ia% ageworkclass=Self-emp-not-inc \* age \*   
## workclass %ia% ageworkclass=State-gov \* age   
## education %ia% ageeducation=HS-grad \* age   
## education %ia% ageeducation=Some-college \* age   
## education %ia% ageeducation=Assoc-voc \* age   
## education %ia% ageeducation=Assoc-acdm \* age   
## education %ia% ageeducation=Bachelors \* age   
## education %ia% ageeducation=Masters \* age   
## education %ia% ageeducation=Prof-school \* age   
## education %ia% ageeducation=Doctorate \* age   
## hours.per.week %ia% workclasshours.per.week \* workclass=Local-gov   
## hours.per.week %ia% workclasshours.per.week \* workclass=Private   
## hours.per.week %ia% workclasshours.per.week \* workclass=Self-emp-inc   
## hours.per.week %ia% workclasshours.per.week \* workclass=Self-emp-not-inc   
## hours.per.week %ia% workclasshours.per.week \* workclass=State-gov   
## hours.per.week %ia% occupationhours.per.week \* occupation=Armed-Forces   
## hours.per.week %ia% occupationhours.per.week \* occupation=Craft-repair   
## hours.per.week %ia% occupationhours.per.week \* occupation=Exec-managerial   
## hours.per.week %ia% occupationhours.per.week \* occupation=Farming-fishing   
## hours.per.week %ia% occupationhours.per.week \* occupation=Handlers-cleaners   
## hours.per.week %ia% occupationhours.per.week \* occupation=Machine-op-inspct   
## hours.per.week %ia% occupationhours.per.week \* occupation=Other-service   
## hours.per.week %ia% occupationhours.per.week \* occupation=Priv-house-serv   
## hours.per.week %ia% occupationhours.per.week \* occupation=Prof-specialty \*   
## hours.per.week %ia% occupationhours.per.week \* occupation=Protective-serv   
## hours.per.week %ia% occupationhours.per.week \* occupation=Sales   
## hours.per.week %ia% occupationhours.per.week \* occupation=Tech-support   
## hours.per.week %ia% occupationhours.per.week \* occupation=Transport-moving   
## age %ia% hours.per.week   
## sexMale:raceAsian-Pac-Islander   
## sexMale:raceBlack \*   
## sexMale:raceOther   
## sexMale:raceWhite \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 20810 on 18528 degrees of freedom  
## Residual deviance: 11427 on 18435 degrees of freedom  
## AIC: 11615  
##   
## Number of Fisher Scoring iterations: 11

log.reg.probabilities4 = predict(log.reg4, census\_query, type="response")  
log.reg.predictions4 = ifelse(log.reg.probabilities4 > 0.5, ">50K", "<=50K")  
mean(log.reg.predictions4 != census\_query$income)

## [1] 0.1401123

This gives us the best AIC so far. THe error rate is 14%

Now we’ll do LDA with those variables just to investigate the error.

lda2 = lda((income == ">50K") ~   
   
 # Quantitative  
 rcs(age, 5) + fnlwgt + rcs(capital.gain, 5) +  
 rcs(capital.loss, 5) + rcs(hours.per.week, 5) +   
   
 # Qualitative  
 sex \* race + marital.status + workclass +   
 education + occupation +  
   
 # Interactions  
 marital.status %ia% age + workclass %ia% age +  
 education %ia% age + hours.per.week %ia% workclass +   
 hours.per.week %ia% occupation + age %ia% hours.per.week, data=census\_train)  
lda.predictions.text2 = predict(lda2, census\_query)  
lda.predictions2 = ifelse(lda.predictions.text2$class == "TRUE", ">50K", "<=50K")  
  
mean(lda.predictions2 != census\_query$income)

## [1] 0.1433506

We don’t get as good test error. We’ll go with the fourth logistic regression model.

We’re ready to test with the official test data, but first we have to format it.

census\_test = read\_csv("census\_test.csv")

## Parsed with column specification:  
## cols(  
## age = col\_integer(),  
## workclass = col\_character(),  
## fnlwgt = col\_integer(),  
## education = col\_character(),  
## `education-num` = col\_integer(),  
## `marital-status` = col\_character(),  
## occupation = col\_character(),  
## relationship = col\_character(),  
## race = col\_character(),  
## sex = col\_character(),  
## `capital-gain` = col\_integer(),  
## `capital-loss` = col\_integer(),  
## `hours-per-week` = col\_integer(),  
## `native-country` = col\_character(),  
## income = col\_character()  
## )

census\_test[census\_test == "?"] <- NA  
census\_test <- census\_test %>%   
 na.omit %>%   
 mutate(workclass = as.factor(workclass), education = as.factor(education),  
 marital.status = as.factor(`marital-status`), occupation = as.factor(occupation),  
 race = as.factor(race), sex = as.factor(sex), capital.gain = `capital-gain`,  
 capital.loss = `capital-loss`, hours.per.week = `hours-per-week`,  
 native.country = as.factor(`native-country`), income = as.factor(income),  
 income.num = ifelse(income==">50K", 1, 0)) %>%   
 select(age, workclass, fnlwgt, education, marital.status, occupation, race, sex,   
 capital.gain, capital.loss, hours.per.week, native.country, income, income.num) %>%   
 mutate(native.country = fct\_recode(as.factor(native.country),  
 "USA or Canada" = "United-States", "USA or Canada" = "Canada",  
 "USA or Canada" = "Outlying-US(Guam-USVI-etc)",  
 "Europe" = "France", "Europe" = "Italy", "Europe" = "Poland",  
 "Europe" = "Scotland", "Europe" = "Germany", "Europe" = "Portugal",  
 "Europe" = "Yugoslavia", "Europe" = "England", "Europe" = "Greece",  
 "Europe" = "Holand-Netherlands", "Europe" = "Hungary",  
 "Europe" = "Ireland", "Asia" = "Cambodia", "Asia" = "India",  
 "Asia" = "Laos", "Asia" = "Thailand", "Asia" = "Vietnam",  
 "Asia" = "Hong", "Asia" = "Iran", "Asia" = "China",  
 "Asia" = "Japan", "Asia" = "Philippines", "Asia" = "Taiwan",  
 "Asia" = "South", "Latin America or Caribbean" = "Columbia",  
 "Latin America or Caribbean" = "Ecuador",  
 "Latin America or Caribbean" = "Guatemala",  
 "Latin America or Caribbean" = "Honduras",  
 "Latin America or Caribbean" = "Cuba",  
 "Latin America or Caribbean" = "El-Salvador",  
 "Latin America or Caribbean" = "Haiti",   
 "Latin America or Caribbean" = "Jamaica",  
 "Latin America or Caribbean" = "Mexico",  
 "Latin America or Caribbean" = "Peru",  
 "Latin America or Caribbean" = "Trinadad&Tobago",  
 "Latin America or Caribbean" = "Dominican-Republic",  
 "Latin America or Caribbean" = "Nicaragua",  
 "Latin America or Caribbean" = "Puerto-Rico"),  
 marital.status = fct\_recode(as.factor(marital.status),  
 "Married" = "Married-AF-spouse", "Married" = "Married-civ-spouse"),  
 marital.status = fct\_relevel(marital.status, "Married", "Married-spouse-absent", "Separated",  
 "Divorced", "Widowed")) %>%   
 filter(workclass != "Without-pay") %>%  
 droplevels() %>%  
 mutate(education = fct\_collapse(as.factor(education),  
 "Less than HS grad" = c("Preschool", "1st-4th", "5th-6th", "7th-8th",  
 "9th", "10th", "11th", "12th")),  
 education = fct\_relevel(education, "Less than HS grad", "HS-grad", "Some-college", "Assoc-voc",  
 "Assoc-acdm", "Bachelors", "Masters", "Prof-school", "Doctorate"))

## Warning: Unknown levels in `f`: Holand-Netherlands

census\_test$id <- 1:nrow(census\_test)

log.reg.probabilities.test = predict(log.reg4, census\_test, type="response")  
log.reg.predictions.test = ifelse(log.reg.probabilities.test > 0.5, ">50K", "<=50K")  
mean(log.reg.predictions.test != census\_test$income)

## [1] 0.1535709

The final test error is 15.4 percent.