471 Midterm

Nik Krieger

2018-03-09

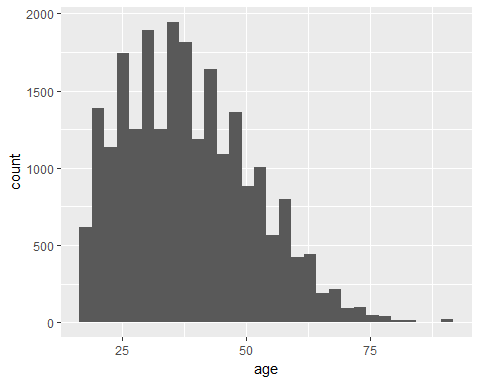
**Exploratory Data Analysis**

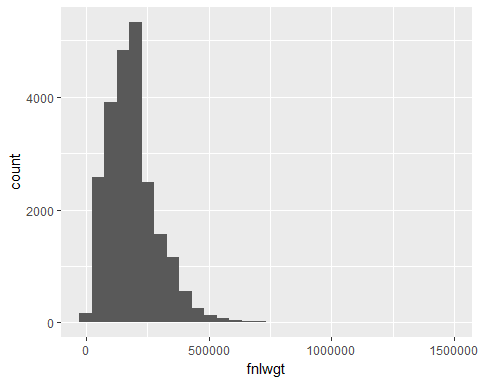
First, we employed Hmisc::describe() to check for missingness. The function reported no missing data, but the categorical variables workclass, occupation, and native.country each contain a level named ?. We flagged those values as missing data and then check to see how much missingness there was. There turned out to be 1829 missing values, which is about 7.316% of the data. We omitted these values.

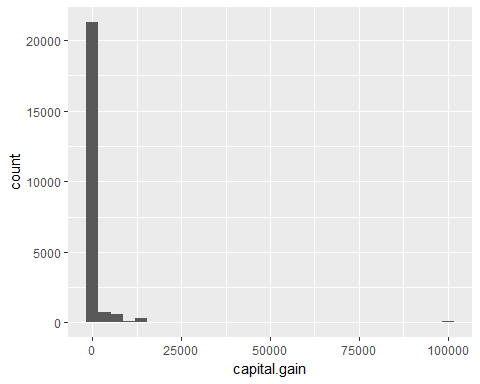
A cross-table of education and education-num revealed that they contain the same data in different formats, I ignored education-num.

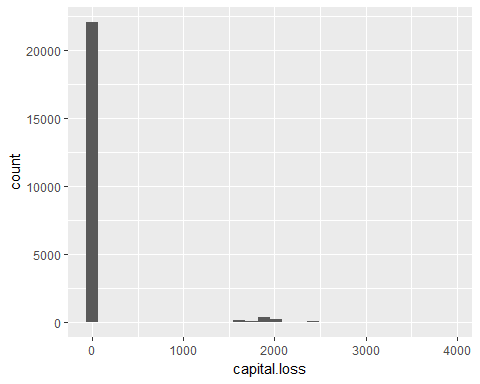
A cross-table of marital-status and relationship indicated some kind of redundancy. Since marital-status is intuitive and relationship is decidedly not, we dropped relationship.

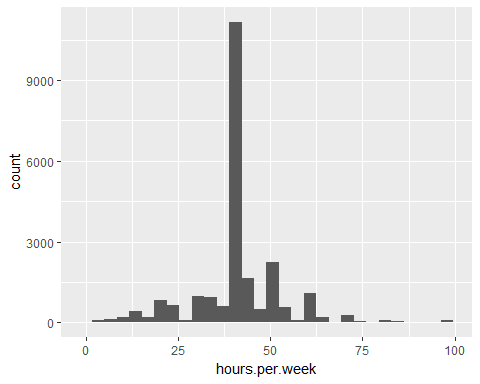
We performed numerical summaries and created histograms of the quantitative predictors. Histograms are shown below:





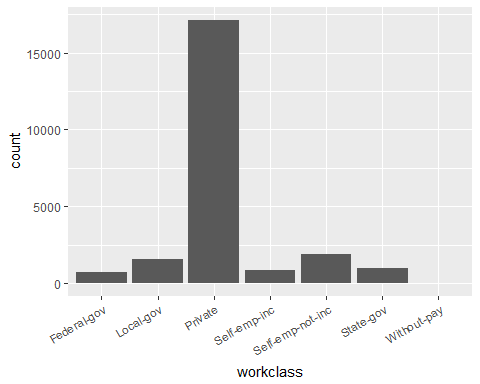


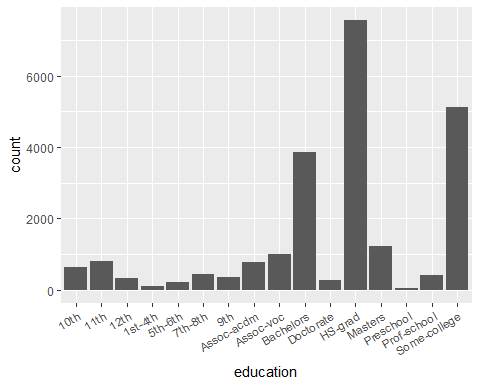


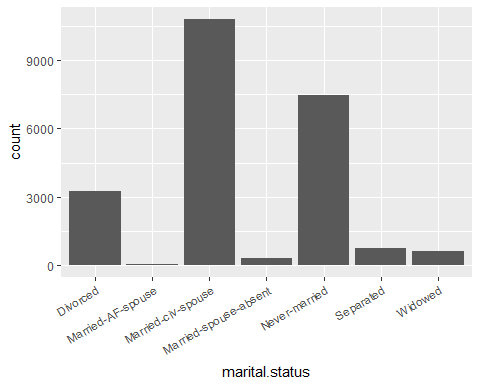


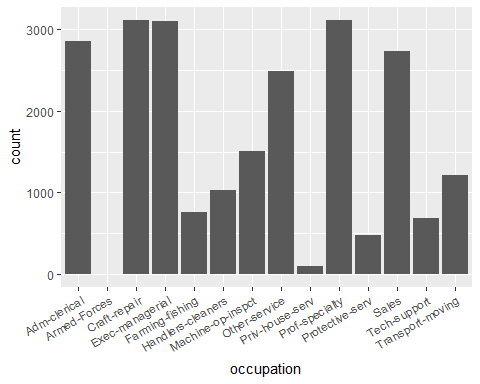
All values seemed to make sense and be in a normal range. The only strange one was the maximum capital.gain value of 99999. It was nearly three times higher than the next highest value of 34095, but I was not convinced that it is a mistake or a missingness indicator. I have a feeling that it was censored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

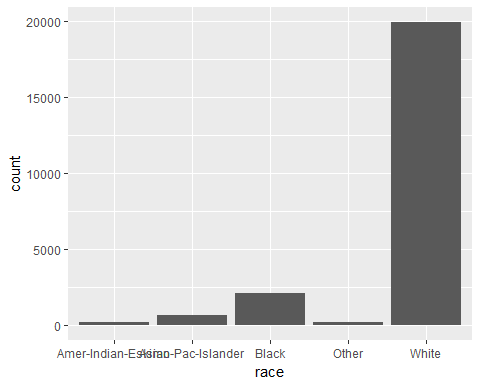
Next we looked at bar graphs of the categorical predictors as well as the outcome variable income.

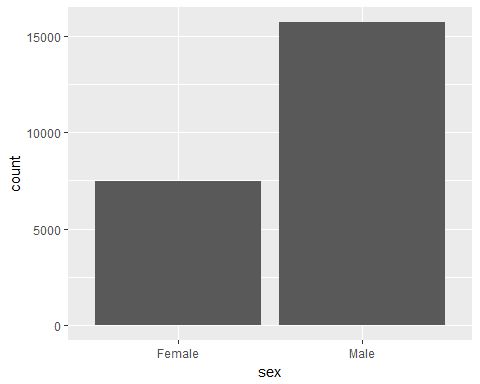


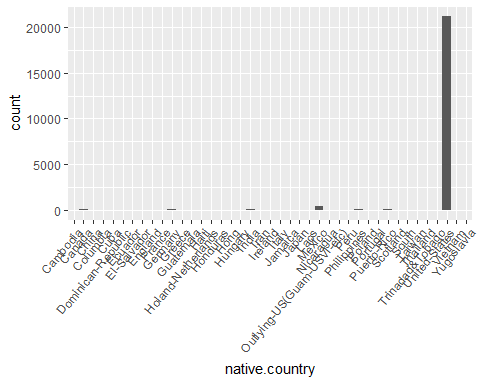


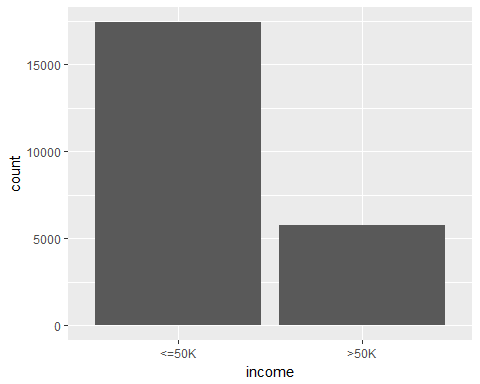












We collapsed various smaller levels of a few of the factors when possible.

We then looked at a scatterplot matrix of the quantitative predictors along with the outcome, and we looked at tables of the qualitative predictors and the outcome.

There did not seem to be any concerning trends for the quantitative predictors nor determinative relationships among the quantitative variables.

**Creating the Models**

In order to validate our model, we further split census\_training into a training sample (80% of the data) and a query sample (20% of the data).

We used 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)

It had an AIC of:

## AIC: 12193

and a query test error of

## [1] 0.1429188

Next we made an LDA kitchen sink model:

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)  
  
This produced a query test error of 15.3%

Then we tried KNN models with k = 1, 3, 5, 10, 50, and 100, which produced errors of:

## [k=1] 0.2722366

## [k=3] 0.2428756

## [k=5] 0.2266839

## [k=10] 0.2091969

## [k=50] 0.2100604

## [k=100] 0.2182642

These error rates were much higher when compared to logistic regression and LDA.

We decided to go with the logistic regression and fine-tune it.

Firstly, we used cubic splines with 5 nodes for all five quantitative predictors:

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)

This produced an AIC of:

## AIC: 11622

and a test error of

## [1] 0.1379534

So we improved a little. We decided to reduce the number of nodes for fnlwgt and capital.loss from 5 to 3 and added several interaction terms:

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)

This produced an AIC of:

## AIC: 11643

and a test error of

## [1] 0.1381693

We put the cubic splines back on capital.loss and removed some of the more complicated interaction terms that had not seemed to produce much significance.

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)

We got the lowest AIC yet at:

## AIC: 11615

The error rate was

## [1] 0.1401123

For a last stab at a better model, we used the same combination of predictors above with an LDA model:

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)

This produced a test error of:

## [1] 0.1433506

This is not as good as the logistic regression model. We chose the fourth logistic regression model log.reg4 as the final model.

**Testing the Test Data**  
We finally test our model on the test data in census\_test.csv:

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. Not bad!

Here’s that final model one more time:

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)