Comparison of models

Jon Minton

30 April 2016

## Introduction

This document will compare a series of models which predict the share of all convictions, in any given year, committed by different age/sex groups. Firsly, a series of models are developed which predict the share of convictions in any given year given age and sex only. The best of these models will be our reference 'best time invariant' model, and will be a formal representation of the age-crime (gender) curve.

The hypothesis that the age crime curve is not invariant will be formally assessed by exploring whether meaningful improvements in the model fit can be achieved by including period-based terms which interact with age terms.

## Load packages and data

rm(list=ls())  
  
  
library(plyr)  
library(stringr)  
library(tidyr)  
library(dplyr)  
  
  
library(RColorBrewer)  
library(ggplot2)  
  
data <- read.csv("../data/real/scotland\_all.csv") %>%  
 tbl\_df  
  
names(data) <- c(  
 "country",  
 "year",  
 "age",  
 "sex",  
 "convicted",  
 "total"  
)  
  
data$sex <- tolower(data$sex)  
data$age <- revalue(data$age, c("90 & over" = "90"))  
data$age <- as.numeric(as.character(data$age))

We first want to see what the data look like

```{r, eval = T}

data ```

As we are interested primarily in whether the shape of the age-crime curve has changed over time, we calculate for each year the proportion of all convictions associated with each age/sex combination

model\_data <- data %>%  
 group\_by(year) %>%  
 mutate(prop\_convicted = convicted / sum(convicted)) %>%  
 select(year, age, sex, prop\_convicted) %>%  
 arrange(year, sex, age)  
  
model\_data

We now start to fit an increasingly complicated series of linear regression models which regression the proportion convicted against other terms. For each of these, we extract the AIC, a measure of model fit penalised by model complexity (i.e. number of terms), and save these AIC scores as a new output.

lm(prop\_convicted ~ year, model\_data) %>% AIC() -> a\_year  
lm(prop\_convicted ~ sex, model\_data) %>% AIC() -> a\_sex  
lm(prop\_convicted ~ age, model\_data) %>% AIC() -> a\_age  
  
c(a\_year, a\_sex, a\_age)

Of these three simplest models, the age model has the lowest AIC, even though the sex model involves fewer parameters. All further models therefore contain age at the very least.

We now look at models containing age + at least one other set of terms. To start with, we look at adding additional terms without interactions.

lm(prop\_convicted ~ age + year, model\_data) %>% AIC() -> a\_age\_year  
lm(prop\_convicted ~ age + sex, model\_data) %>% AIC() -> a\_age\_sex  
  
c(a\_age\_year, a\_age\_sex)

Of these two models, age + sex beats age + year, so all further models will now contain age and sex at a minimum.

An important feature of the age-crime curve is its skewednes/nonlinearity. To represent this as terms in a linear regression model we will look at adding varying numbers of polynomials to the model. We compare between 1 and 10 polynomials as follows:

Firstly, we look at models which include between 1 and 10 polynomials of age, along with sex, but without interactions between them:

lm(prop\_convicted ~ poly(age,1) + sex, model\_data) %>% AIC() -> a\_age1\_sex  
lm(prop\_convicted ~ poly(age,2) + sex, model\_data) %>% AIC() -> a\_age2\_sex  
lm(prop\_convicted ~ poly(age,3) + sex, model\_data) %>% AIC() -> a\_age3\_sex  
lm(prop\_convicted ~ poly(age,4) + sex, model\_data) %>% AIC() -> a\_age4\_sex  
lm(prop\_convicted ~ poly(age,5) + sex, model\_data) %>% AIC() -> a\_age5\_sex  
lm(prop\_convicted ~ poly(age,6) + sex, model\_data) %>% AIC() -> a\_age6\_sex  
lm(prop\_convicted ~ poly(age,7) + sex, model\_data) %>% AIC() -> a\_age7\_sex  
lm(prop\_convicted ~ poly(age,8) + sex, model\_data) %>% AIC() -> a\_age8\_sex  
lm(prop\_convicted ~ poly(age,9) + sex, model\_data) %>% AIC() -> a\_age9\_sex  
lm(prop\_convicted ~ poly(age,10) + sex, model\_data) %>% AIC() -> a\_age10\_sex

Next, we will produce a series of models, from which we extract the AIC, containing interactions between each of the age polynomials and sex. This means that the schedule of relative propensity to conviction and age is allowed to be different for males and females.

lm(prop\_convicted ~ poly(age,1) \* sex, model\_data) %>% AIC() -> a\_age1sex  
lm(prop\_convicted ~ poly(age,2) \* sex, model\_data) %>% AIC() -> a\_age2sex  
lm(prop\_convicted ~ poly(age,3) \* sex, model\_data) %>% AIC() -> a\_age3sex  
lm(prop\_convicted ~ poly(age,4) \* sex, model\_data) %>% AIC() -> a\_age4sex  
lm(prop\_convicted ~ poly(age,5) \* sex, model\_data) %>% AIC() -> a\_age5sex  
lm(prop\_convicted ~ poly(age,6) \* sex, model\_data) %>% AIC() -> a\_age6sex  
lm(prop\_convicted ~ poly(age,7) \* sex, model\_data) %>% AIC() -> a\_age7sex  
lm(prop\_convicted ~ poly(age,8) \* sex, model\_data) %>% AIC() -> a\_age8sex  
lm(prop\_convicted ~ poly(age,9) \* sex, model\_data) %>% AIC() -> a\_age9sex  
lm(prop\_convicted ~ poly(age,10) \* sex, model\_data) %>% AIC() -> a\_age10sex

As this has produced a lot of model fits to compare against each other, and we want to understand better the relationship between the number of polynomials and the quality of the model fit, we package the AIC values together in a dataframe to visualise the results better

tmp <- data.frame(poly = 1:10,  
 interaction = rep(c(F, T), each = 10),  
 model = rep(c("aic", "bic"), each = 20),  
 fit =  
 c(  
 a\_age1\_sex, a\_age2\_sex, a\_age3\_sex, a\_age4\_sex, a\_age5\_sex,  
 a\_age6\_sex, a\_age7\_sex, a\_age8\_sex, a\_age9\_sex, a\_age10\_sex,  
 a\_age1sex, a\_age2sex, a\_age3sex, a\_age4sex, a\_age5sex,  
 a\_age6sex, a\_age7sex, a\_age8sex, a\_age9sex, a\_age10sex,  
 b\_age1\_sex, b\_age2\_sex, b\_age3\_sex, b\_age4\_sex, b\_age5\_sex,  
 b\_age6\_sex, b\_age7\_sex, b\_age8\_sex, b\_age9\_sex, b\_age10\_sex,  
 b\_age1sex, b\_age2sex, b\_age3sex, b\_age4sex, b\_age5sex,  
 b\_age6sex, b\_age7sex, b\_age8sex, b\_age9sex, b\_age10sex  
  
  
 )  
)

We can now produce a plot showing this relationship as follows

qplot(x = as.factor(poly), y = fit, group = interaction, colour = interaction, shape = model, data = tmp)

From this we can conclude firstly that the models with interactions between age and sex tend to greatly outperform the non-interaction models, and so interactions should be included. We can conclude secondly that penalised model fit continues to improve with each additional polynomial added. However, most of the improvement in fit comes only from the first three polynomials: age, age squared, and age cubed. Although lower AIC is better, slightly simpler models tend to be a bit easier to interpret, and so as a compromise we select the model with interactions and a third order polynomial as our best model to represent the hypothesis that the age-crime curve is time-invariant.

# best\_invariant\_model <- lm(prop\_convicted ~ poly(age,3) \* sex, model\_data)  
#   
#   
# model\_with\_year <- update(best\_invariant\_model, . ~ . + I(year - 1989))  
#   
# model\_interrupted\_complex <- model\_data %>%  
# ungroup() %>%  
# mutate(year2 = year - 1989) %>%  
# mutate(year3 = ifelse(year - 2000 < 0, 0, year - 2000)) %>%  
# lm(prop\_convicted ~ poly(age,3) \* sex + year2 + year3, data = .)  
#   
#   
#   
# model\_interrupted\_interracted <- model\_data %>%  
# ungroup() %>%  
# mutate(year2 = year - min(year)) %>%  
# mutate(post\_drop = ifelse(year >= 2000, TRUE, FALSE)) %>%  
# lm(prop\_convicted ~ poly(age,3) \* sex \* post\_drop, data = .)  
#   
# model\_yearandinterrupted\_interracted <- model\_data %>%  
# ungroup() %>%  
# mutate(year2 = year - min(year)) %>%  
# mutate(year3 = ifelse(year >= 2000, year - 2000, 0)) %>%  
# lm(prop\_convicted ~ (poly(age,3) \* sex )\* (year2 + year3), data = .)  
#   
# ```  
#   
# ## Including Plots  
#   
# You can also embed plots, for example:  
#   
# ```{r pressure, echo=FALSE}  
# plot(pressure)  
# ```  
#   
# Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.