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.

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:plyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

Now to load and set up the data

```
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

```
print(data)
```

```
## Source: local data frame [3,600 x 6]
##
##    country  year   age   sex convicted total
##     (fctr) (int) (dbl) (chr)     (int) (int)
## 1      SCO  1989    16  male      3435 36212
## 2      SCO  1989    17  male      3988 38222
## 3      SCO  1989    18  male      4058 40295
## 4      SCO  1989    19  male      3887 40217
## 5      SCO  1989    20  male      3500 42004
## 6      SCO  1989    21  male      3272 42623
## 7      SCO  1989    22  male      2864 43389
## 8      SCO  1989    23  male      2477 41610
## 9      SCO  1989    24  male      2389 43254
## 10     SCO  1989    25  male      2223 42131
## ..     ...   ...   ...   ...       ...   ...
```

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
```

```
## Source: local data frame [3,600 x 4]
## Groups: year [24]
##
##     year   age    sex prop_convicted
##    (int) (dbl)  (chr)          (dbl)
## 1   1989    16 female    0.006599789
## 2   1989    17 female    0.007193770
## 3   1989    18 female    0.007160771
## 4   1989    19 female    0.006550290
## 5   1989    20 female    0.007144271
## 6   1989    21 female    0.006550290
## 7   1989    22 female    0.007061774
## 8   1989    23 female    0.005642819
## 9   1989    24 female    0.005659319
## 10  1989    25 female    0.004916843
## ..   ...   ...    ...            ...
```

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)
```

```
## [1] -21782.89 -22397.04 -23720.75
```

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)
```

```
## [1] -23718.75 -24845.18
```

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
```

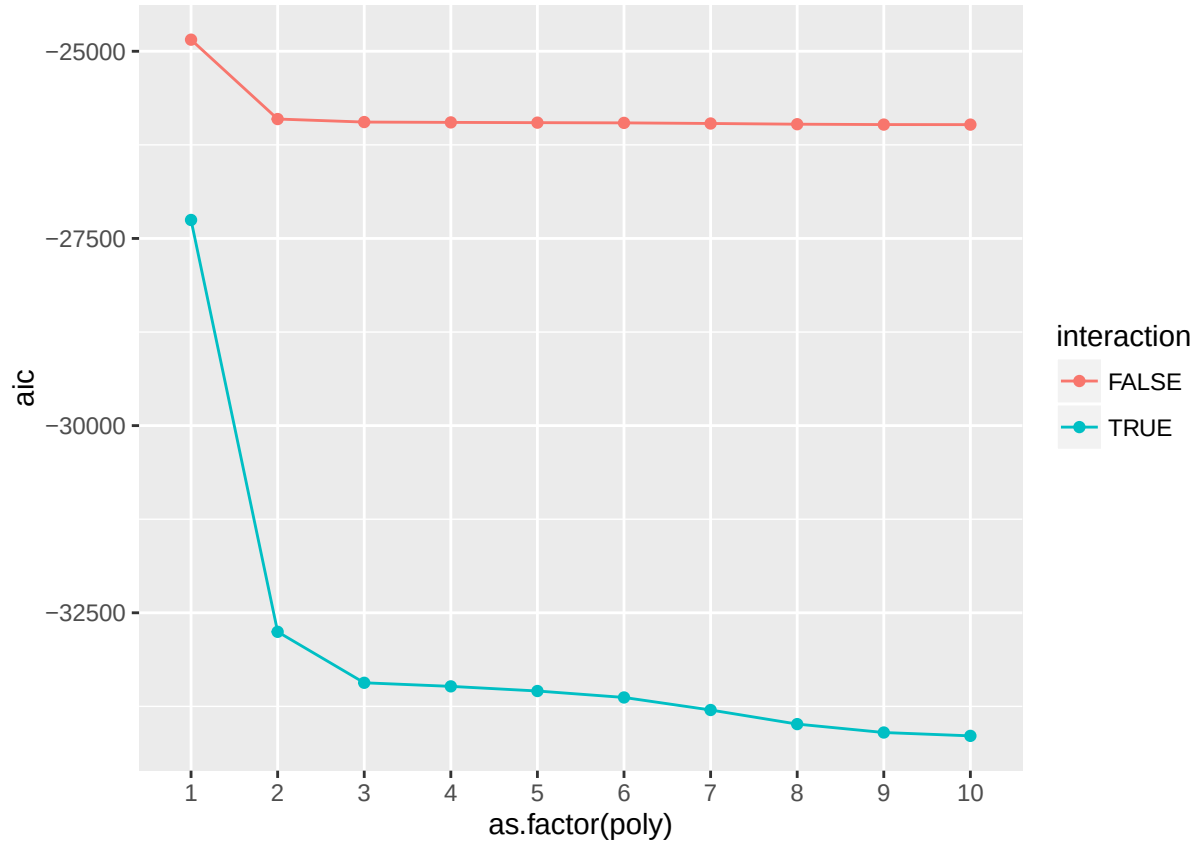
```
lm(prop_convicted ~ poly(age,8)  * sex, model_data) %>% AIC() -> a_age8sex
lm(prop_convicted ~ poly(age,9)  * sex, model_data) %>% AIC() -> a_age9sex
lm(prop_convicted ~ poly(age,10)  * sex, model_data) %>% AIC() -> a_age10sex
```

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),
                  aic =
                    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

                    )
)
```

We can now produce a plot showing this relationship as follows

```
qplot(x = as.factor(poly), y = aic, group = interaction, colour = interaction, data = tmp) +
  geom_line()
```



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.

For now, let's look at some summary statistics for the third order model with sex interactions.

```
best_invariant_model <- lm(prop_convicted ~ poly(age,3)  * sex, model_data)
summary(best_invariant_model)
```

```
##
## Call:
## lm(formula = prop_convicted ~ poly(age, 3) * sex, data = model_data)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.036637 -0.000355  0.000039  0.000272  0.019948
##
## Coefficients:
##                          Estimate Std. Error  t value Pr(>|t|)
## (Intercept)             0.0020193  0.0000548   36.848   <2e-16 ***
## poly(age, 3)1          -0.1328940  0.0032880  -40.418   <2e-16 ***
## poly(age, 3)2           0.0570160  0.0032880   17.341   <2e-16 ***
## poly(age, 3)3           0.0041876  0.0032880    1.274    0.203
## sexmale                 0.0092948  0.0000775  119.935   <2e-16 ***
## poly(age, 3)1:sexmale -0.6427595  0.0046499 -138.230   <2e-16 ***
## poly(age, 3)2:sexmale  0.3510239  0.0046499   75.490   <2e-16 ***
## poly(age, 3)3:sexmale -0.0942976  0.0046499  -20.279   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.002325 on 3592 degrees of freedom
## Multiple R-squared:  0.9608, Adjusted R-squared:  0.9608
## F-statistic: 1.259e+04 on 7 and 3592 DF,  p-value: < 2.2e-16
```

We can see that almost all of the terms here are extremely statistically significant, with only the third order polynomial on age for females not significant at $p < 0.05$. Importantly, we can see that the multiple R-squared and adjusted R-squared scores are very high: around 0.96. This necessarily means that the absolute improvement in fit of a model which allows the age schedule to vary with age will time, compared to the time-invarient model, will necessarily be very small. However, the improvement in fit could still be worth it, as evaluated using something like AIC.

Building time varying alternative models

We are not interested in models that simply add period variables as extra variables. Instead, we are interested in models in which period is allowed to influence the polynomials of age against annual conviction share, i.e. is interacted with each of these terms. Two alternative specifications for doing this are produced below.

In the first, period is included as a simple dummy variable: 1 for year 2000 onwards, 0 otherwise. This is to represent the idea that the crime drop began in the 2000s, and before that the age-crime curve was invariant.

In the second model, time since 1989 (start of the time series) is added, meaning that it is assumed the level of change in the shape of the age-crime curve is constant throughout the time period. Again, this is of course a gross simplification.

In the third model, time since 1989 (start of the time series) and time since 2000 (start of the supposed age-crime drop) are each allowed to interact with the other age/sex variables. This allows for the rate of change in the shape of the curve to be different after 2000 compared with before it. This third model in some ways could be seen as a hybrid of the first two models: the first which assumes a discrete change after 2000, the second which assumes a constant change in the shape over time. However, many alternative model specifications could be produced.

```
time_varying_interrupted_model <- 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 = .)

time_varying_constant_change_model <- model_data %>%
  ungroup() %>%
  mutate(year2 = year - min(year)) %>%
  lm(prop_convicted ~ poly(age,3)  * sex * year2, data = .)


time_varying_hybrid_model <- 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 = .)
```

Comparing models

We now have four models: our 'best' time-invariant model with third order polynomials of age interacted with sex; and three models of increasinging complexity which allow the shape of the age-crime curve to change over time. We also know that the time-invariant model has a very impressive fit with the empirical data, with an R-squared of 0.96, but that of course there is still room for improvement.

We begin by comparing the AIC of all four models against each other

```
AIC(
  best_invariant_model,
  time_varying_interrupted_model,
  time_varying_constant_change_model,
  time_varying_hybrid_model
  )
```

```
##                                   df       AIC
## best_invariant_model               9 -33434.80
## time_varying_interrupted_model    17 -34938.33
## time_varying_constant_change_model 17 -36371.61
## time_varying_hybrid_model         25 -36407.23
```

Each of these time-varying models has a better penalised model fit than the time-invariant model. However, the models which allow the the age-crime curve shape to change linearly over time outperform the second

model, in which the period is split into two simple periods (2000 and beyond; before 2000). Despite eight more parameters, the hybrid model still has a better penalised model fit, using AIC, than the constant change model. However, there are different parameters for penalising model complexity, and if each additional parameter were penalised more severely then the 'best' model may be different. BIC, an alternative to AIC, penalises additional parameters at a rate of log(n) per additional parameter, where n is the number of observations in the dataset. For this dataset that equates to 8.2 points per parameter, compared with 2 points per parameter for AIC. Therefore, it penalises more complex models more severely. Comparing BIC for the four above models gives the following:

```
BIC(
  best_invariant_model,
  time_varying_interrupted_model,
  time_varying_constant_change_model,
  time_varying_hybrid_model
  )
```

```
##                                     df       BIC
## best_invariant_model                 9 -33379.10
## time_varying_interrupted_model      17 -34833.12
## time_varying_constant_change_model 17 -36266.40
## time_varying_hybrid_model           25 -36252.52
```

We see here that BIC suggests the constant-change rather than the hybrid model is the best model to use. For completeness, and equivalence with the approach taken with polynomials of age, we might also wish to produce a series of models of polynomials of year. Some of these models are likely to contain many parameters, and because the model is already quite complex we will only consider up to five parameters.

```
tmp <- model_data %>% ungroup() %>% mutate(year2 = year - min(year))

t1_model <- lm(prop_convicted ~ poly(age,3)  * sex * poly(year2, 1), tmp)
t2_model <- lm(prop_convicted ~ poly(age,3)  * sex * poly(year2, 2), tmp)
t3_model <- lm(prop_convicted ~ poly(age,3)  * sex * poly(year2, 3), tmp)
t4_model <- lm(prop_convicted ~ poly(age,3)  * sex * poly(year2, 4), tmp)
t5_model <- lm(prop_convicted ~ poly(age,3)  * sex * poly(year2, 5), tmp)

AIC(best_invariant_model, t1_model, t2_model, t3_model, t4_model, t5_model)
```

```
##                      df       AIC
## best_invariant_model  9 -33434.80
## t1_model             17 -36371.61
## t2_model             25 -36415.75
## t3_model             33 -36631.75
## t4_model             41 -36617.14
## t5_model             49 -36607.69
```

```
BIC(best_invariant_model, t1_model, t2_model, t3_model, t4_model, t5_model)
```

```
##                      df       BIC
## best_invariant_model  9 -33379.10
## t1_model             17 -36266.40
## t2_model             25 -36261.03
## t3_model             33 -36427.52
## t4_model             41 -36363.41
## t5_model             49 -36304.45
```

Both AIC and BIC suggest a third order polynomial of age interacted with a third order polynomial of year since start of series (and sex) produces the best compromise between model fit and model complexity. Let's now look at some of the summary statistics for this third order polynomial model, as well as the t1_model introduced earlier.

```
summary(t1_model)
```

```
##
## Call:
## lm(formula = prop_convicted ~ poly(age, 3) * sex * poly(year2,
##     1), data = tmp)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0217018 -0.0003714  0.0000350  0.0003934  0.0097050
##
## Coefficients:
##                                       Estimate Std. Error  t value
## (Intercept)                           2.019e-03  3.640e-05   55.467
## poly(age, 3)1                        -1.329e-01  2.184e-03  -60.841
## poly(age, 3)2                         5.702e-02  2.184e-03   26.103
## poly(age, 3)3                         4.188e-03  2.184e-03    1.917
## sexmale                               9.295e-03  5.148e-05  180.539
## poly(year2, 1)                        8.850e-03  2.184e-03    4.052
## poly(age, 3)1:sexmale                -6.428e-01  3.089e-03 -208.078
## poly(age, 3)2:sexmale                 3.510e-01  3.089e-03  113.636
## poly(age, 3)3:sexmale                -9.430e-02  3.089e-03  -30.527
## poly(age, 3)1:poly(year2, 1)         -2.102e-01  1.311e-01   -1.604
## poly(age, 3)2:poly(year2, 1)         -5.712e-01  1.311e-01   -4.358
## poly(age, 3)3:poly(year2, 1)          7.400e-01  1.311e-01    5.646
## sexmale:poly(year2, 1)               -1.770e-02  3.089e-03   -5.730
## poly(age, 3)1:sexmale:poly(year2, 1)  3.389e+00  1.853e-01   18.284
## poly(age, 3)2:sexmale:poly(year2, 1) -5.451e+00  1.853e-01  -29.411
## poly(age, 3)3:sexmale:poly(year2, 1)  4.772e+00  1.853e-01   25.749
##                                      Pr(>|t|)
## (Intercept)                           < 2e-16 ***
## poly(age, 3)1                         < 2e-16 ***
## poly(age, 3)2                         < 2e-16 ***
## poly(age, 3)3                          0.0553 .
## sexmale                               < 2e-16 ***
## poly(year2, 1)                       5.19e-05 ***
## poly(age, 3)1:sexmale                 < 2e-16 ***
## poly(age, 3)2:sexmale                 < 2e-16 ***
## poly(age, 3)3:sexmale                 < 2e-16 ***
## poly(age, 3)1:poly(year2, 1)           0.1089
## poly(age, 3)2:poly(year2, 1)         1.35e-05 ***
## poly(age, 3)3:poly(year2, 1)         1.77e-08 ***
## sexmale:poly(year2, 1)               1.09e-08 ***
## poly(age, 3)1:sexmale:poly(year2, 1)  < 2e-16 ***
## poly(age, 3)2:sexmale:poly(year2, 1)  < 2e-16 ***
## poly(age, 3)3:sexmale:poly(year2, 1)  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 0.001545 on 3584 degrees of freedom
## Multiple R-squared:  0.9828, Adjusted R-squared:  0.9827
## F-statistic: 1.362e+04 on 15 and 3584 DF,  p-value: < 2.2e-16
```

```
summary(t3_model)
```

```
##
## Call:
## lm(formula = prop_convicted ~ poly(age, 3) * sex * poly(year2,
##     3), data = tmp)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0180169 -0.0003400  0.0000251  0.0003437  0.0094895
##
## Coefficients:
##                                        Estimate Std. Error  t value
## (Intercept)                            2.019e-03  3.504e-05   57.635
## poly(age, 3)1                         -1.329e-01  2.102e-03  -63.219
## poly(age, 3)2                          5.702e-02  2.102e-03   27.123
## poly(age, 3)3                          4.188e-03  2.102e-03    1.992
## sexmale                                9.295e-03  4.955e-05  187.594
## poly(year2, 3)1                        8.850e-03  2.102e-03    4.210
## poly(year2, 3)2                       -3.404e-03  2.102e-03   -1.619
## poly(year2, 3)3                       -1.790e-03  2.102e-03   -0.852
## poly(age, 3)1:sexmale                 -6.428e-01  2.973e-03 -216.210
## poly(age, 3)2:sexmale                  3.510e-01  2.973e-03  118.077
## poly(age, 3)3:sexmale                 -9.430e-02  2.973e-03  -31.720
## poly(age, 3)1:poly(year2, 3)1         -2.102e-01  1.261e-01   -1.666
## poly(age, 3)2:poly(year2, 3)1         -5.712e-01  1.261e-01   -4.528
## poly(age, 3)3:poly(year2, 3)1          7.400e-01  1.261e-01    5.867
## poly(age, 3)1:poly(year2, 3)2          3.545e-01  1.261e-01    2.811
## poly(age, 3)2:poly(year2, 3)2         -3.457e-01  1.261e-01   -2.741
## poly(age, 3)3:poly(year2, 3)2          1.670e-01  1.261e-01    1.324
## poly(age, 3)1:poly(year2, 3)3          1.236e-01  1.261e-01    0.980
## poly(age, 3)2:poly(year2, 3)3         -1.046e-01  1.261e-01   -0.829
## poly(age, 3)3:poly(year2, 3)3          1.499e-01  1.261e-01    1.189
## sexmale:poly(year2, 3)1               -1.770e-02  2.973e-03   -5.954
## sexmale:poly(year2, 3)2                6.809e-03  2.973e-03    2.290
## sexmale:poly(year2, 3)3                3.581e-03  2.973e-03    1.204
## poly(age, 3)1:sexmale:poly(year2, 3)1  3.389e+00  1.784e-01   18.999
## poly(age, 3)2:sexmale:poly(year2, 3)1 -5.451e+00  1.784e-01  -30.561
## poly(age, 3)3:sexmale:poly(year2, 3)1  4.772e+00  1.784e-01   26.756
## poly(age, 3)1:sexmale:poly(year2, 3)2 -1.702e-01  1.784e-01   -0.954
## poly(age, 3)2:sexmale:poly(year2, 3)2 -2.770e-01  1.784e-01   -1.553
## poly(age, 3)3:sexmale:poly(year2, 3)2  3.245e-01  1.784e-01    1.819
## poly(age, 3)1:sexmale:poly(year2, 3)3  1.350e-01  1.784e-01    0.757
## poly(age, 3)2:sexmale:poly(year2, 3)3 -8.951e-01  1.784e-01   -5.018
## poly(age, 3)3:sexmale:poly(year2, 3)3  1.475e+00  1.784e-01    8.269
##                                       Pr(>|t|)
## (Intercept)                            < 2e-16 ***
## poly(age, 3)1                          < 2e-16 ***
## poly(age, 3)2                          < 2e-16 ***
## poly(age, 3)3                          0.04644 *
```

```
## sexmale                                    < 2e-16 ***
## poly(year2, 3)1                           2.61e-05 ***
## poly(year2, 3)2                            0.10544
## poly(year2, 3)3                            0.39445
## poly(age, 3)1:sexmale                      < 2e-16 ***
## poly(age, 3)2:sexmale                      < 2e-16 ***
## poly(age, 3)3:sexmale                      < 2e-16 ***
## poly(age, 3)1:poly(year2, 3)1              0.09576 .
## poly(age, 3)2:poly(year2, 3)1             6.13e-06 ***
## poly(age, 3)3:poly(year2, 3)1             4.84e-09 ***
## poly(age, 3)1:poly(year2, 3)2              0.00497 **
## poly(age, 3)2:poly(year2, 3)2              0.00616 **
## poly(age, 3)3:poly(year2, 3)2              0.18560
## poly(age, 3)1:poly(year2, 3)3              0.32711
## poly(age, 3)2:poly(year2, 3)3              0.40690
## poly(age, 3)3:poly(year2, 3)3              0.23466
## sexmale:poly(year2, 3)1                   2.87e-09 ***
## sexmale:poly(year2, 3)2                    0.02207 *
## sexmale:poly(year2, 3)3                    0.22850
## poly(age, 3)1:sexmale:poly(year2, 3)1      < 2e-16 ***
## poly(age, 3)2:sexmale:poly(year2, 3)1      < 2e-16 ***
## poly(age, 3)3:sexmale:poly(year2, 3)1      < 2e-16 ***
## poly(age, 3)1:sexmale:poly(year2, 3)2      0.34002
## poly(age, 3)2:sexmale:poly(year2, 3)2      0.12052
## poly(age, 3)3:sexmale:poly(year2, 3)2      0.06894 .
## poly(age, 3)1:sexmale:poly(year2, 3)3      0.44902
## poly(age, 3)2:sexmale:poly(year2, 3)3     5.48e-07 ***
## poly(age, 3)3:sexmale:poly(year2, 3)3      < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.001486 on 3568 degrees of freedom
## Multiple R-squared:  0.9841, Adjusted R-squared:  0.984
## F-statistic:  7123 on 31 and 3568 DF,  p-value: < 2.2e-16
```

Although the third order polynomial of year model has slightly better R-squared and adjusted R-squared than the first order version, the improvement in fit is marginal: with Multiple R-squared increasing from 0.983 to 0.984. As the first order version is also easier to present, we might want to use this as our 'best' model for representing the hypothesis that the shape of the age-crime curve is time variant rather than time invariant.

Iterrim conclusion

Regardless of which time-varying model we use, all models tend to outperform the time-invariant model in terms of AIC and BIC. However, we should also bear in mind that the time-invariant model has a very high model fit, with R-squared values of around 0.96. There is therefore evidence in support of both positions: people arguing that the age-crime curve is invariant over time can point to the very high model fit of a time-invariant mode; people arguing that the age-crime curve is varying over time can point to the superior penalised model fits of models which allow the shape of the age-crime curve to vary with time, largely regardless of how these time-varying models are operationalised.

Predictions

Something we might want to to is compare predicted values against actual values for specific years, using different model specifications. For brevity, we will look at how the models compare with actual results for the years 1999, 2000 and 2010.

Plotting predicted against actual values shows something important about the models which include third order polynomials against age: although they capture the shape of the decline in conviction share with age after the age of peak conviction, they do not capture the first part of the curve, i.e. the rapid rise in conviction propensity from the age of 16 to 17 or 18 years (i.e. age of peak conviction). A number of additional models are also calculated and compared below. The model 'time invariant complex' does not allow the curve to vary with time, but does use a 10th order polynomial with age, so should be able to represent this left hand side of the figure. The mode' t3 complex' uses the third order polynomial of time interacted with the 10th order polynomial of age, so is the most complicated and flexible of all models presented here. The AIC and BIC of these models are also presented.

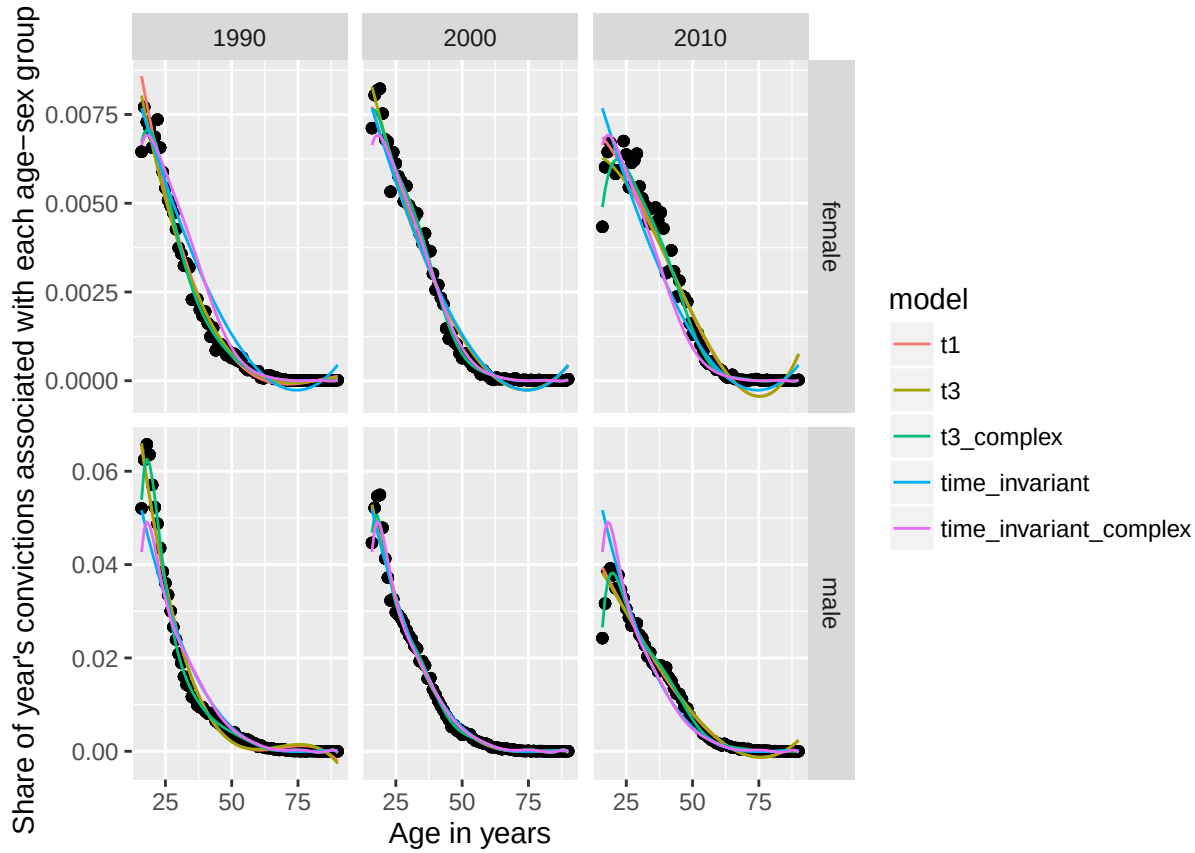
```
predicted_actual_values <- model_data

# I also want to compare all these models against 10th order poly

model_10_poly_invariant <- lm(prop_convicted ~ poly(age, 10) * sex, data = model_data)
model_10_poly_variant <- model_data %>% ungroup() %>% mutate(year2 = year - min(year)) %>% lm(prop_convi



predicted_actual_values <- data.frame(
  predicted_actual_values,
  predicted_time_invariant = best_invariant_model$fitted.values,
  predicted_time_invariant_complex = model_10_poly_invariant$fitted.values,
  predicted_t1 = t1_model$fitted.values,
  predicted_t3 = t3_model$fitted.values,
  predicted_t3_complex = model_10_poly_variant$fitted.values
) %>% tbl_df %>%
  select(sex, year, age, actual = prop_convicted, contains("predicted")) %>%
  gather(key = "model", value = "prediction", contains("predicted")) %>%
  mutate(model = str_replace(model, "^predicted\\_", ""))

AIC(
  best_invariant_model,
  model_10_poly_invariant,
  t1_model,
  t3_model,
  model_10_poly_variant
)
```

```
##                           df       AIC
## best_invariant_model       9 -33434.80
## model_10_poly_invariant   23 -34143.59
## t1_model                  17 -36371.61
## t3_model                  33 -36631.75
## model_10_poly_variant     89 -42461.17
```

```
BIC(
  best_invariant_model,
  model_10_poly_invariant,
  t1_model,
  t3_model,
  model_10_poly_variant
)
```

```
##                           df       BIC
## best_invariant_model       9 -33379.10
## model_10_poly_invariant   23 -34001.25
## t1_model                  17 -36266.40
## t3_model                  33 -36427.52
## model_10_poly_variant     89 -41910.38
```

```
predicted_actual_values %>%
  filter(year %in% c(1990, 2000, 2010)) %>%
  mutate(year = factor(year)) %>%
  ggplot(., aes(x = age)) +
  facet_grid(sex ~ year, scales = "free_y") +
  geom_point(aes(y = actual)) +
  geom_line(aes(y = prediction, group = model, colour = model)) +
  labs(y = "Share of year's convictions associated with each age-sex group", x = "Age in years")
```



Discussion

As both the graphs comparing fitted against actual values, and the calculations of penalised model fit using AIC and BIC demonstrate, a considerable improvement in the quality of the fit can be found with models that allow both polynomials of year and age, and for interactions between age, year and sex. However, models using more parameters are both harder to explain and more liable to be 'over-fit', meaning they may be less effective for projections than simpler models using fewer parameters. It may be possible to assess this further by seeing, for example, how effective different models fit on data up to (say) 2000 are at predicting shares after 2000.

Regarding the main hypothesis being evaluated - whether the age-crime curve is variant or invariant over time - there is reasonable evidence that the shape of the age-crime curve has changed over time, and so that including period terms produces substantial improvement in model fit. However it is also important not to forget that the fit of almost all models tends to be very high. For exampe, the simpler time-invariant model has an adjusted R squared value of 0.961, compared with an adjusted R squared of 0.997 for the most complex of the models, so the actual improvement in model fit, at least over the current period, has been relatively modest.