Black Assault death trends analysis

Jon Minton

3 May 2016

## Introduction

This document will present a series of analyses relating to assault/homicide related death trends in young adult Black Non-Hispanic (BNH) males.

To start with, I will load various dependencies and set various options that allow existing R stripts to be rendered as a PDF document.

Now to start to load and filter the data:

dta <- read\_csv("data/tidied/icd\_8\_to\_10\_8fold.csv")  
  
dta

## Source: local data frame [30,144 x 8]  
##   
## icd\_class year age sex race  
## (chr) (int) (chr) (chr) (chr)  
## 1 icd\_08 1968 < 1 female black  
## 2 icd\_08 1968 < 1 female black  
## 3 icd\_08 1968 < 1 female black  
## 4 icd\_08 1968 < 1 female black  
## 5 icd\_08 1968 < 1 female black  
## 6 icd\_08 1968 < 1 female black  
## 7 icd\_08 1968 < 1 female black  
## 8 icd\_08 1968 < 1 female black  
## 9 icd\_08 1968 1-4 female black  
## 10 icd\_08 1968 1-4 female black  
## .. ... ... ... ... ...  
## Variables not shown: cause (chr), death\_count (int), population\_count  
## (dbl)

# Which age groups for which years?  
dta %>% xtabs(~ year + age, data = .)

## age  
## year < 1 1-4 10-14 15-19 20-24 25-34 35-44 45-54 5-9 55-64 65-74 75-84  
## 1968 48 48 48 48 48 48 48 48 48 48 48 48  
## 1969 48 48 48 48 48 48 48 48 48 48 48 48  
## 1970 48 48 48 48 48 48 48 48 48 48 48 48  
## 1971 48 48 48 48 48 48 48 48 48 48 48 48  
## 1972 48 48 48 48 48 48 48 48 48 48 48 48  
## 1973 48 48 48 48 48 48 48 48 48 48 48 48  
## 1974 48 48 48 48 48 48 48 48 48 48 48 48  
## 1975 48 48 48 48 48 48 48 48 48 48 48 48  
## 1976 48 48 48 48 48 48 48 48 48 48 48 48  
## 1977 48 48 48 48 48 48 48 48 48 48 48 48  
## 1978 48 48 48 48 48 48 48 48 48 48 48 48  
## 1979 48 48 48 48 48 48 48 48 48 48 48 48  
## 1980 48 48 48 48 48 48 48 48 48 48 48 48  
## 1981 48 48 48 48 48 48 48 48 48 48 48 48  
## 1982 48 48 48 48 48 48 48 48 48 48 48 48  
## 1983 48 48 48 48 48 48 48 48 48 48 48 48  
## 1984 48 48 48 48 48 48 48 48 48 48 48 48  
## 1985 48 48 48 48 48 48 48 48 48 48 48 48  
## 1986 48 48 48 48 48 48 48 48 48 48 48 48  
## 1987 48 48 48 48 48 48 48 48 48 48 48 48  
## 1988 48 48 48 48 48 48 48 48 48 48 48 48  
## 1989 48 48 48 48 48 48 48 48 48 48 48 48  
## 1990 48 48 48 48 48 48 48 48 48 48 48 48  
## 1991 48 48 48 48 48 48 48 48 48 48 48 48  
## 1992 48 48 48 48 48 48 48 48 48 48 48 48  
## 1993 48 48 48 48 48 48 48 48 48 48 48 48  
## 1994 48 48 48 48 48 48 48 48 48 48 48 48  
## 1995 48 48 48 48 48 48 48 48 48 48 48 48  
## 1996 48 48 48 48 48 48 48 48 48 48 48 48  
## 1997 48 48 48 48 48 48 48 48 48 48 48 48  
## 1998 48 48 48 48 48 48 48 48 48 48 48 48  
## 1999 64 64 64 64 64 64 64 64 64 64 64 64  
## 2000 64 64 64 64 64 64 64 64 64 64 64 64  
## 2001 64 64 64 64 64 64 64 64 64 64 64 64  
## 2002 64 64 64 64 64 64 64 64 64 64 64 64  
## 2003 64 64 64 64 64 64 64 64 64 64 64 64  
## 2004 64 64 64 64 64 64 64 64 64 64 64 64  
## 2005 64 64 64 64 64 64 64 64 64 64 64 64  
## 2006 64 64 64 64 64 64 64 64 64 64 64 64  
## 2007 64 64 64 64 64 64 64 64 64 64 64 64  
## 2008 64 64 64 64 64 64 64 64 64 64 64 64  
## 2009 64 64 64 64 64 64 64 64 64 64 64 64  
## 2010 64 64 64 64 64 64 64 64 64 64 64 64  
## 2011 64 64 64 64 64 64 64 64 64 64 64 64  
## 2012 64 64 64 64 64 64 64 64 64 64 64 64  
## 2013 64 64 64 64 64 64 64 64 64 64 64 64  
## 2014 64 64 64 64 64 64 64 64 64 64 64 64

# What discrete types of death are available?  
  
dta %>% group\_by(cause) %>% tally

## Source: local data frame [8 x 2]  
##   
## cause n  
## (chr) (int)  
## 1 all\_cause 3768  
## 2 all\_external 3768  
## 3 assault 3768  
## 4 intentional\_or\_undetermined\_self\_harm 3768  
## 5 intentional\_self\_harm 3768  
## 6 non\_external 3768  
## 7 undetermined 3768  
## 8 vehicle 3768

We decided previously to look at the age group 15 to 34 years, to focus on BNH, and to look at deaths through assault and homicide. Additionally, we want to add a variable indicating for which years there was a Republican Presidency.

## Source: local data frame [94 x 5]  
## Groups: year [47]  
##   
## year sex death\_count population\_count rep  
## (int) (chr) (int) (dbl) (dbl)  
## 1 1968 female 710 3406000 1  
## 2 1968 male 3217 3027000 1  
## 3 1969 female 755 3538000 1  
## 4 1969 male 3566 3153000 1  
## 5 1970 female 760 3651372 0  
## 6 1970 male 3900 3267088 0  
## 7 1971 female 900 3840000 0  
## 8 1971 male 4514 3468000 0  
## 9 1972 female 956 4002000 0  
## 10 1972 male 4750 3640000 0  
## .. ... ... ... ... ...

Next, we produce another variant that allows the log mortality rate to vary as a function of time in years since each change in Presidency between parties, in a new dataframe called 'knotted'

dta\_subset %>% mutate(  
 k1970 = ifelse( year > 1970, year - 1970, 0), # JAVIER: IS THIS CORRECT?  
 k1978 = ifelse( year > 1978, year - 1978, 0),  
 k1982 = ifelse( year > 1982, year - 1982, 0),  
 k1994 = ifelse( year > 1994, year - 1994, 0),   
 k2002 = ifelse( year > 2002, year - 2002, 0),  
 k2010 = ifelse( year > 2010, year - 2010, 0)  
) %>%   
 mutate(  
 death\_rate = death\_count / population\_count,  
 lg10mr = log(death\_rate, 10)   
 ) -> knotted   
  
knotted

## Source: local data frame [94 x 13]  
## Groups: year [47]  
##   
## year sex death\_count population\_count rep k1970 k1978 k1982 k1994  
## (int) (chr) (int) (dbl) (dbl) (dbl) (dbl) (dbl) (dbl)  
## 1 1968 female 710 3406000 1 0 0 0 0  
## 2 1968 male 3217 3027000 1 0 0 0 0  
## 3 1969 female 755 3538000 1 0 0 0 0  
## 4 1969 male 3566 3153000 1 0 0 0 0  
## 5 1970 female 760 3651372 0 0 0 0 0  
## 6 1970 male 3900 3267088 0 0 0 0 0  
## 7 1971 female 900 3840000 0 1 0 0 0  
## 8 1971 male 4514 3468000 0 1 0 0 0  
## 9 1972 female 956 4002000 0 2 0 0 0  
## 10 1972 male 4750 3640000 0 2 0 0 0  
## .. ... ... ... ... ... ... ... ... ...  
## Variables not shown: k2002 (dbl), k2010 (dbl), death\_rate (dbl), lg10mr  
## (dbl)

## First analysis: model of log mort against knots

We first produce two separate models, one for each sex, of log mortality against the knots

models <- dlply( # This code runs the model  
 knotted,   
 .(sex),   
 function(x) lm(lg10mr ~ year + k1970 + k1978 + k1982 + k1994+ k2002+ k2010, data = x)  
)  
  
llply(models, summary) # This code prints the summary for each gender's model

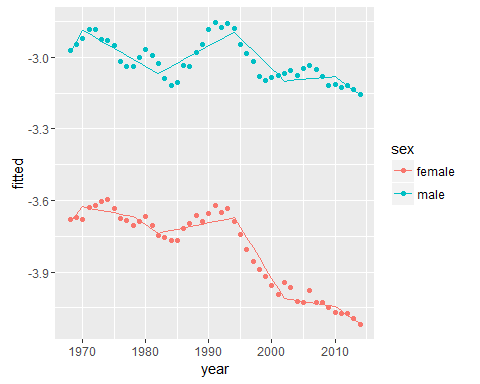
## $female  
##   
## Call:  
## lm(formula = lg10mr ~ year + k1970 + k1978 + k1982 + k1994 +   
## k2002 + k2010, data = x)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.057364 -0.023080 -0.002725 0.016194 0.067359   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -63.585285 39.904606 -1.593 0.11914   
## year 0.030436 0.020263 1.502 0.14114   
## k1970 -0.035382 0.022663 -1.561 0.12655   
## k1978 -0.012739 0.009954 -1.280 0.20817   
## k1982 0.023098 0.008270 2.793 0.00805 \*\*   
## k1994 -0.047528 0.004763 -9.979 2.72e-12 \*\*\*  
## k2002 0.037737 0.006058 6.229 2.49e-07 \*\*\*  
## k2010 -0.013505 0.011567 -1.168 0.25006   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.03572 on 39 degrees of freedom  
## Multiple R-squared: 0.9619, Adjusted R-squared: 0.9551   
## F-statistic: 140.8 on 7 and 39 DF, p-value: < 2.2e-16  
##   
##   
## $male  
##   
## Call:  
## lm(formula = lg10mr ~ year + k1970 + k1978 + k1982 + k1994 +   
## k2002 + k2010, data = x)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.082860 -0.035583 0.004911 0.033209 0.084179   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -9.416e+01 5.175e+01 -1.819 0.076540 .   
## year 4.633e-02 2.628e-02 1.763 0.085735 .   
## k1970 -6.140e-02 2.939e-02 -2.089 0.043271 \*   
## k1978 -3.581e-04 1.291e-02 -0.028 0.978010   
## k1982 3.006e-02 1.073e-02 2.803 0.007853 \*\*   
## k1994 -4.030e-02 6.177e-03 -6.525 9.7e-08 \*\*\*  
## k2002 2.803e-02 7.857e-03 3.568 0.000973 \*\*\*  
## k2010 -2.146e-02 1.500e-02 -1.431 0.160458   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.04632 on 39 degrees of freedom  
## Multiple R-squared: 0.7436, Adjusted R-squared: 0.6976   
## F-statistic: 16.16 on 7 and 39 DF, p-value: 9.242e-10

We might also want to compare the fitted against the actual values

tmp <- knotted %>%   
 select(year, sex, actual = lg10mr)  
  
tmp\_female <- tmp %>% filter(sex == "female")  
tmp\_male <- tmp %>% filter(sex == "male")  
  
tmp\_female$fitted <- models[["female"]][["fitted.values"]]  
tmp\_male$fitted <- models[["male"]][["fitted.values"]]  
  
fitted\_actual\_comparisons <- bind\_rows(tmp\_female, tmp\_male)  
  
rm(tmp, tmp\_female, tmp\_male)  
  
fitted\_actual\_comparisons

## Source: local data frame [94 x 4]  
##   
## year sex actual fitted  
## (int) (chr) (dbl) (dbl)  
## 1 1968 female -3.680986 -3.687740  
## 2 1969 female -3.670811 -3.657304  
## 3 1970 female -3.681642 -3.626868  
## 4 1971 female -3.630089 -3.631815  
## 5 1972 female -3.621819 -3.636762  
## 6 1973 female -3.603631 -3.641708  
## 7 1974 female -3.593696 -3.646655  
## 8 1975 female -3.634759 -3.651602  
## 9 1976 female -3.676343 -3.656548  
## 10 1977 female -3.684171 -3.661495  
## .. ... ... ... ...

fitted\_actual\_comparisons %>%   
 ggplot(., aes(x = year, group = sex, colour = sex)) +   
 geom\_line(aes(y = fitted)) +   
 geom\_point(aes(y = actual))

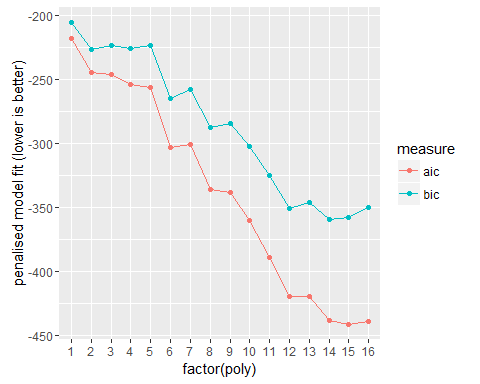


Interestingly, it seems from the figure above that the 'elbows' built into the linear model do not quite match the most likely points of change. In particular, mortality rates continued to deteriorate until around 1985 for both sexes, a log of around three years from change in regime to change in trend. Then, in the 1990s, changes in mortality rates LED change in regime by around four years. The change in 2000, by contrast, seems more synchronised with changes (and improvements) in death rate trends.

## Additional (interrim) exercise: polynomials with year

As an interrim modelling analysis, here are some comparative model fits using different numbers of polynomial on year. These are compared using AIC and BIC.

simple <- knotted %>%   
 select(year, sex, lg10mr) %>%   
 ungroup %>%   
 mutate(year2 = year - min(year)) # set to 1968 == 0  
  
fn <- function(k){  
 output <- lm(lg10mr ~ poly(year2, k) \* sex, data = simple)   
}  
  
tmp <- llply(1:16, fn)  
  
results <- data\_frame(  
 poly = 1:16,   
 aic = sapply(tmp, AIC),  
 bic = sapply(tmp, BIC)  
)  
  
results %>% gather(measure, value, aic, bic) %>%   
 ggplot(., aes(x = factor(poly), y = value, group = measure, colour = measure)) +   
 geom\_line() + geom\_point() +   
 labs(y = "penalised model fit (lower is better)")



rm(tmp, results)

This suggests that both 2nd order and 6th order polynomials seem reasonable fits, although as might be expected the fit tends to improve with the number of polynomials, even when adjusted (penalised) for model complexity using AIC or BIC. Interestingly, the penalised fit continues to improve up to around the 12th order, which could be telling as there are 47 years, which divided by 4, the number of years between presidential elections, equals 12 to the nearest whole number. Let's now look at the 2nd, 6th and 12th order polynomial fits against the actual data.

fit\_poly2 <- lm(lg10mr ~ poly(year2, 2) \* sex, data = simple)  
fit\_poly6 <- lm(lg10mr ~ poly(year2, 6) \* sex, data = simple)  
fit\_poly12 <- lm(lg10mr ~ poly(year2, 12) \* sex, data = simple)  
  
  
summary(fit\_poly2)

##   
## Call:  
## lm(formula = lg10mr ~ poly(year2, 2) \* sex, data = simple)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.14320 -0.03868 -0.00345 0.03711 0.14514   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -3.806640 0.009249 -411.553 < 2e-16 \*\*\*  
## poly(year2, 2)1 -1.436463 0.089677 -16.018 < 2e-16 \*\*\*  
## poly(year2, 2)2 -0.491404 0.089677 -5.480 4.00e-07 \*\*\*  
## sexmale 0.793313 0.013081 60.648 < 2e-16 \*\*\*  
## poly(year2, 2)1:sexmale 0.959522 0.126822 7.566 3.54e-11 \*\*\*  
## poly(year2, 2)2:sexmale 0.328112 0.126822 2.587 0.0113 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.06341 on 88 degrees of freedom  
## Multiple R-squared: 0.9785, Adjusted R-squared: 0.9772   
## F-statistic: 799.3 on 5 and 88 DF, p-value: < 2.2e-16

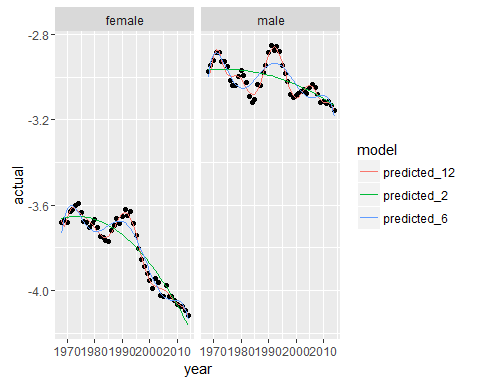
summary(fit\_poly6)

##   
## Call:  
## lm(formula = lg10mr ~ poly(year2, 6) \* sex, data = simple)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.094204 -0.027268 -0.000036 0.025689 0.083933   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -3.806640 0.006505 -585.167 < 2e-16 \*\*\*  
## poly(year2, 6)1 -1.436463 0.063070 -22.776 < 2e-16 \*\*\*  
## poly(year2, 6)2 -0.491404 0.063070 -7.791 2.07e-11 \*\*\*  
## poly(year2, 6)3 0.080165 0.063070 1.271 0.207398   
## poly(year2, 6)4 0.231517 0.063070 3.671 0.000435 \*\*\*  
## poly(year2, 6)5 0.123619 0.063070 1.960 0.053476 .   
## poly(year2, 6)6 -0.334642 0.063070 -5.306 9.76e-07 \*\*\*  
## sexmale 0.793313 0.009200 86.232 < 2e-16 \*\*\*  
## poly(year2, 6)1:sexmale 0.959522 0.089195 10.758 < 2e-16 \*\*\*  
## poly(year2, 6)2:sexmale 0.328112 0.089195 3.679 0.000424 \*\*\*  
## poly(year2, 6)3:sexmale -0.277083 0.089195 -3.106 0.002620 \*\*   
## poly(year2, 6)4:sexmale -0.084628 0.089195 -0.949 0.345579   
## poly(year2, 6)5:sexmale 0.039368 0.089195 0.441 0.660138   
## poly(year2, 6)6:sexmale -0.004841 0.089195 -0.054 0.956851   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.0446 on 80 degrees of freedom  
## Multiple R-squared: 0.9903, Adjusted R-squared: 0.9887   
## F-statistic: 629 on 13 and 80 DF, p-value: < 2.2e-16

summary(fit\_poly12)

##   
## Call:  
## lm(formula = lg10mr ~ poly(year2, 12) \* sex, data = simple)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.044113 -0.013861 -0.002095 0.012439 0.041378   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -3.806640 0.003339 -1139.966 < 2e-16 \*\*\*  
## poly(year2, 12)1 -1.436463 0.032375 -44.369 < 2e-16 \*\*\*  
## poly(year2, 12)2 -0.491404 0.032375 -15.178 < 2e-16 \*\*\*  
## poly(year2, 12)3 0.080165 0.032375 2.476 0.015777 \*   
## poly(year2, 12)4 0.231517 0.032375 7.151 7.58e-10 \*\*\*  
## poly(year2, 12)5 0.123619 0.032375 3.818 0.000293 \*\*\*  
## poly(year2, 12)6 -0.334642 0.032375 -10.336 1.39e-15 \*\*\*  
## poly(year2, 12)7 -0.071283 0.032375 -2.202 0.031076 \*   
## poly(year2, 12)8 0.236441 0.032375 7.303 4.02e-10 \*\*\*  
## poly(year2, 12)9 0.027542 0.032375 0.851 0.397920   
## poly(year2, 12)10 -0.106171 0.032375 -3.279 0.001642 \*\*   
## poly(year2, 12)11 -0.092121 0.032375 -2.845 0.005857 \*\*   
## poly(year2, 12)12 0.080967 0.032375 2.501 0.014803 \*   
## sexmale 0.793313 0.004722 167.988 < 2e-16 \*\*\*  
## poly(year2, 12)1:sexmale 0.959522 0.045786 20.957 < 2e-16 \*\*\*  
## poly(year2, 12)2:sexmale 0.328112 0.045786 7.166 7.11e-10 \*\*\*  
## poly(year2, 12)3:sexmale -0.277083 0.045786 -6.052 6.90e-08 \*\*\*  
## poly(year2, 12)4:sexmale -0.084628 0.045786 -1.848 0.068898 .   
## poly(year2, 12)5:sexmale 0.039368 0.045786 0.860 0.392905   
## poly(year2, 12)6:sexmale -0.004841 0.045786 -0.106 0.916104   
## poly(year2, 12)7:sexmale 0.060825 0.045786 1.328 0.188459   
## poly(year2, 12)8:sexmale -0.010134 0.045786 -0.221 0.825493   
## poly(year2, 12)9:sexmale 0.081116 0.045786 1.772 0.080933 .   
## poly(year2, 12)10:sexmale -0.084485 0.045786 -1.845 0.069358 .   
## poly(year2, 12)11:sexmale -0.092149 0.045786 -2.013 0.048117 \*   
## poly(year2, 12)12:sexmale 0.078601 0.045786 1.717 0.090585 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.02289 on 68 degrees of freedom  
## Multiple R-squared: 0.9978, Adjusted R-squared: 0.997   
## F-statistic: 1251 on 25 and 68 DF, p-value: < 2.2e-16

comparisons <- simple %>%   
 select(year, sex, actual = lg10mr)   
  
comparisons$predicted\_2 <- fit\_poly2$fitted.values  
comparisons$predicted\_6 <- fit\_poly6$fitted.values  
comparisons$predicted\_12 <- fit\_poly12$fitted.values  
  
comparisons <- comparisons %>% gather(model, prediction, predicted\_2, predicted\_6, predicted\_12)  
  
comparisons %>%   
 ggplot(., aes(x = year)) +  
 geom\_point(aes(y = actual)) +   
 geom\_line(aes(y = prediction, group = model, colour = model)) +   
 facet\_wrap(~ sex)



We can see from the above that the 12th order polynomial has a close fit to the data, including possible four year cycles in the rates which may be indicative of the electoral cycle influencing the trends. This intuition is explored further in the following exercise

## Detrending exercises

Given the exploratory analyses above, I will now look at the residuals on Republican associated with the 12th order polynomial, corresponding approximately to a 4 year cycle, as well as 6th order polynomial, corresponding approximately to an 8 year cycle.

# Exercise 2 --------------------------------------------------------------  
  
# Those trends are bumpy. But to run at least a first approximation, let's do two detrending exercises.  
#   
#   
# E1. Let's generate a 6th order polynomial function of time.   
# Create 6 variables, where t1=year-1968, t2=t1^2, t3=t1^3... t6=t1^6. Then run the following model:  
#   
# y = a + b1\*t + b2\*t2 +... b6\*t6 + e [2]  
#   
# recover the residuals of y (again, for each race-sex (4), age-group (adults), and mortality cause (3)).   
# Then run this regression:  
#   
# res\_y = a + b1\*rep + e [3]  
#   
# where rep is the Republican dummy. Recover the fitted values and plot them.  
#   
  
model\_poly\_6 <- dta\_subset %>%  
 ungroup() %>%  
 mutate(death\_rate = death\_count / population\_count,   
 lg10mr = log(death\_rate, 10),  
 year2 = year - min(year)  
 ) %>%   
 lm(lg10mr ~ poly(year, 6) \* sex, data = .)  
  
model\_poly\_12 <- dta\_subset %>%  
 ungroup() %>%  
 mutate(death\_rate = death\_count / population\_count,   
 lg10mr = log(death\_rate, 10),  
 year2 = year - min(year)  
 ) %>%   
 lm(lg10mr ~ poly(year, 12) \* sex, data = .)  
  
AIC(model\_poly\_6, model\_poly\_12)

## df AIC  
## model\_poly\_6 15 -303.0932  
## model\_poly\_12 27 -419.7389

BIC(model\_poly\_6, model\_poly\_12)

## df BIC  
## model\_poly\_6 15 -264.9438  
## model\_poly\_12 27 -351.0699

# Interestingly AIC and BIC now disagree: AIC suggests 12th order, BIC suggests   
# 6th order  
  
# Now to return to doing separate models for males and females separately in the   
# analyses also using residuals  
  
do\_both\_stages <- function(x, k){  
 stage1 <- lm(lg10mr ~ poly(year, k), data =x)  
   
 residuals <- stage1$residuals  
   
 df <- data.frame(res\_y = residuals, rep = x$rep)  
   
 stage2 <- lm(res\_y ~ rep, data = df)  
   
 output <- list(stage1 = stage1, stage2 = stage2)  
 }  
  
both\_stages\_6 <- dta\_subset %>%   
 mutate(death\_rate = death\_count / population\_count,   
 lg10mr = log(death\_rate, 10)) %>%   
 dlply(., .(sex), do\_both\_stages, k = 6)  
  
both\_stages\_12 <- dta\_subset %>%   
 mutate(death\_rate = death\_count / population\_count,   
 lg10mr = log(death\_rate, 10)) %>%   
 dlply(., .(sex), do\_both\_stages, k = 12)  
  
AIC(both\_stages\_6$male$stage1, both\_stages\_12$male$stage1)

## df AIC  
## both\_stages\_6$male$stage1 8 -140.6138  
## both\_stages\_12$male$stage1 14 -206.2726

BIC(both\_stages\_6$male$stage1, both\_stages\_12$male$stage1)

## df BIC  
## both\_stages\_6$male$stage1 8 -125.8126  
## both\_stages\_12$male$stage1 14 -180.3705

AIC(both\_stages\_6$female$stage1, both\_stages\_12$female$stage1)

## df AIC  
## both\_stages\_6$female$stage1 8 -163.1570  
## both\_stages\_12$female$stage1 14 -211.6183

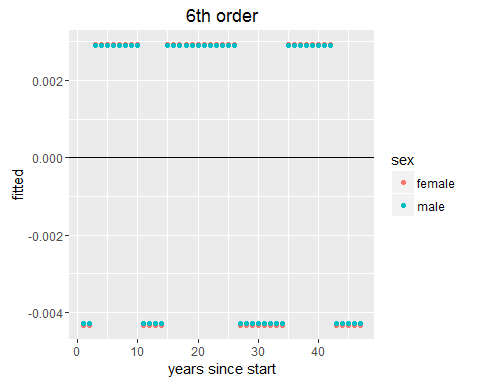
BIC(both\_stages\_6$female$stage1, both\_stages\_12$female$stage1)

## df BIC  
## both\_stages\_6$female$stage1 8 -148.3558  
## both\_stages\_12$female$stage1 14 -185.7162

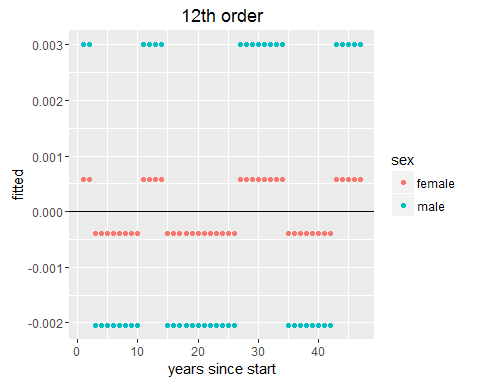
Interestingly, with separate models for males and females, the AIC and BIC are now consistent in terms of relative preference, and suggest the 12th order model has a better penalised model fit than the 6th order model.

Let's now compare the residuals against rep for each of the four models (male, female, 12th order, 6th order)

ldply(both\_stages\_6, function(x) {x[["stage2"]][["fitted.values"]]}) %>%   
 tbl\_df %>%   
 gather(key = obs, value = fitted, -sex) %>%   
 mutate(obs = as.numeric(obs)) %>%   
 arrange(sex) %>%   
 ggplot(.) +   
 geom\_point(aes(x = obs, y = fitted, colour = sex, group = sex)) +   
 geom\_hline(aes(yintercept = 0)) + labs(x = "years since start", title = "6th order")



ldply(both\_stages\_12, function(x) {x[["stage2"]][["fitted.values"]]}) %>%   
 tbl\_df %>%   
 gather(key = obs, value = fitted, -sex) %>%   
 mutate(obs = as.numeric(obs)) %>%   
 arrange(sex) %>%   
 ggplot(.) +   
 geom\_point(aes(x = obs, y = fitted, colour = sex, group = sex)) +   
 geom\_hline(aes(yintercept = 0)) + labs(x = "years since start", title = "12th order")



So, what is the comparison between males and females, and 6th order and 12th order polynomial, showing? To me, it appears to be that, with the model that better represents the 4 year cycle (i.e. 12th order poly), the effect of presidency is much stronger on males than females.

## HP filter-based models

We will turn to filter-based models, which separate values over time into cycle and trend components. The smoothing parameters used in the filter affect the values of both components, and so two different smoothing parameters will be considered, 6.25 and 100.00.

# Exercise 2.2 ------------------------------------------------------------  
  
# E2. Detrend y using the Hodrick-Prescott (HP) filter (there should be a R package that does that).   
# There are two conventional values for the smoothing parameter for annual data: 6.25 and 100.   
# Please use both since they usually do vary the detrending a lot.   
# Once you detrended y using HP, then run equation 3, where res\_y will be the   
# detrended values of y using HP, and plot the fitted values.  
  
# Need to first convert the data to a timeseries format  
  
# actually two separate dataseries - one for males and one for females  
  
fn <- function(x){  
 y <- data.frame(year = as.Date(as.character(x$year), format = "%Y"), value = x$death\_rate)  
 output <- xts(y$value, y$year)  
 return(output)  
}  
  
dta\_list <- dta\_subset %>%   
 mutate(death\_rate = death\_count / population\_count) %>%   
 mutate(death\_rate = log(death\_rate, 10)) %>%   
 dlply(., .(sex), fn)  
  
  
dta\_list\_detrended\_100 <- llply(dta\_list, hpfilter, freq = 100)  
dta\_list\_detrended\_6\_25 <- llply(dta\_list, hpfilter, freq = 6.25)  
  
  
  
fn2 <- function(x){  
 md\_part <- paste0("-", str\_split(today(), "\\-", n = 2)[[1]][[2]]) # the filter above adds the month and the day to years,  
 # this needs to be removed using the lines above (from lubridate) and below  
 year <- dimnames(x$trend)[1][[1]] %>% str\_replace\_all(., md\_part, "") %>% as.numeric()  
 output <- data.frame(year = year, cycle = x$cycle, trend = x$trend[,1])  
 return(output)  
}  
  
  
dta\_df\_detrended\_100 <- ldply(dta\_list\_detrended\_100, fn2) %>%   
 tbl\_df %>% rename(sex = .id)  
  
dta\_df\_detrended\_6\_25 <- ldply(dta\_list\_detrended\_6\_25, fn2) %>%   
 tbl\_df %>% rename(sex = .id)  
  
  
  
  
dta\_df\_detrended\_100 <- dta\_subset %>%   
 select(year, rep) %>%   
 distinct %>%   
 right\_join(dta\_df\_detrended\_100)

## Joining by: "year"

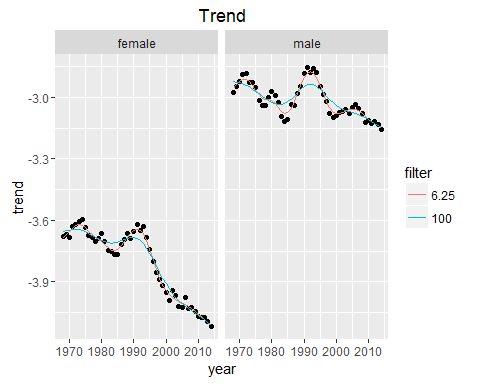
dta\_df\_detrended\_6\_25 <- dta\_subset %>%   
 select(year, rep) %>%   
 distinct %>%   
 right\_join(dta\_df\_detrended\_6\_25)

## Joining by: "year"

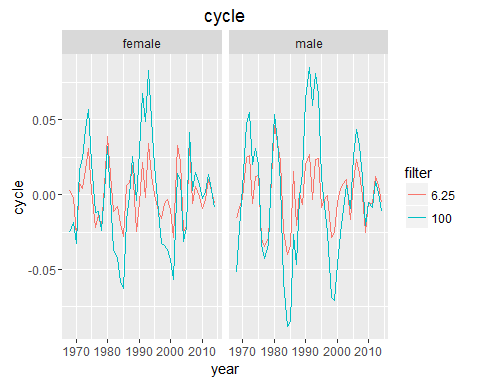
dta\_df\_detrended\_100 <- dta\_df\_detrended\_100 %>% mutate(filter = 100)  
dta\_df\_detrended\_6\_25 <- dta\_df\_detrended\_6\_25 %>% mutate(filter = 6.25)  
  
  
  
dta\_df\_detrended <- bind\_rows(dta\_df\_detrended\_100, dta\_df\_detrended\_6\_25)  
dta\_df\_detrended <- dta\_df\_detrended %>%   
 left\_join(simple) %>%   
 select( -year2) %>%   
 rename(actual = lg10mr) %>%   
 mutate(filter = factor(filter))

## Joining by: c("year", "sex")

dta\_df\_detrended %>%   
 ggplot(., aes(x = year)) +  
 geom\_point(aes(y = actual)) +  
 geom\_line(aes(y = trend, colour = filter, group = filter)) +   
 facet\_wrap(~ sex) +   
 labs(y = "trend", title = "Trend")



dta\_df\_detrended %>%   
 ggplot(., aes(x = year)) +  
 geom\_line(aes(y = cycle, colour = filter, group = filter)) +   
 facet\_wrap(~ sex) +   
 labs(y = "cycle", title = "cycle")



Of the two variants of the model, the parameter 6.25 appears to provide a much closer fit, even though it does not appear to be as close as a 12th order polynomial.

We will now consider the relationship between cycle and rep

dta\_df\_detrended %>% filter(sex == "female", filter == 100) %>% lm(cycle ~ rep, .) %>% summary

##   
## Call:  
## lm(formula = cycle ~ rep, data = .)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.06912 -0.02343 0.00048 0.01848 0.07643   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 0.006223 0.006268 0.993 0.326  
## rep -0.015395 0.009858 -1.562 0.125  
##   
## Residual standard error: 0.03316 on 45 degrees of freedom  
## Multiple R-squared: 0.05141, Adjusted R-squared: 0.03033   
## F-statistic: 2.439 on 1 and 45 DF, p-value: 0.1254

dta\_df\_detrended %>% filter(sex == "male", filter == 100) %>% lm(cycle ~ rep, .) %>% summary

##   
## Call:  
## lm(formula = cycle ~ rep, data = .)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.094622 -0.025577 0.001311 0.024168 0.078109   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 0.006613 0.008002 0.826 0.413  
## rep -0.016359 0.012586 -1.300 0.200  
##   
## Residual standard error: 0.04234 on 45 degrees of freedom  
## Multiple R-squared: 0.03619, Adjusted R-squared: 0.01477   
## F-statistic: 1.69 on 1 and 45 DF, p-value: 0.2003

dta\_df\_detrended %>% filter(sex == "female", filter == 6.25) %>% lm(cycle ~ rep, .) %>% summary

##   
## Call:  
## lm(formula = cycle ~ rep, data = .)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.029438 -0.010910 -0.001459 0.009451 0.040256   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 0.001111 0.003414 0.325 0.746  
## rep -0.002749 0.005369 -0.512 0.611  
##   
## Residual standard error: 0.01806 on 45 degrees of freedom  
## Multiple R-squared: 0.00579, Adjusted R-squared: -0.0163   
## F-statistic: 0.2621 on 1 and 45 DF, p-value: 0.6112

dta\_df\_detrended %>% filter(sex == "male", filter == 6.25) %>% lm(cycle ~ rep, .) %>% summary

##   
## Call:  
## lm(formula = cycle ~ rep, data = .)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.040060 -0.012184 -0.000859 0.013419 0.046210   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 0.0001775 0.0038605 0.046 0.964  
## rep -0.0004390 0.0060719 -0.072 0.943  
##   
## Residual standard error: 0.02043 on 45 degrees of freedom  
## Multiple R-squared: 0.0001161, Adjusted R-squared: -0.0221   
## F-statistic: 0.005227 on 1 and 45 DF, p-value: 0.9427

In each of the above regressions the coefficient on rep appears to be in the right direction. However in no case, for either sex or for either filter, is the coefficient on rep statistically significant.

## Suggestions for additional analysis

* More systematically explore the effect of other filter values
* Try to extract something like AIC or BIC if possible for different filter-based models
* Add more terms to model predicting cycle, for example socioeconomic variables