QMSS GR5016 - Lab1

Mathieu Sauterey - UNI: mjs2364

September 29, 2018

Conduct a trend analysis of some variable of interest. Graph it and try different functional forms. Look for subgroup variation across time, too. Extra credit if you consider other variables as a means of explaining the trend. Explain all of your results.

Analysis across the whole population

I am interested in the trends of American's approval of physician-assisted suicide over time. Respondents were specifically asked "When a person has a disease that cannot be cured, do you think doctors should be allowed by law to end the patient's life by some painless means if the patient and his family request it?". The results are included in the column *letdie*1 of the GSS data table. We are also going to conduct subgroup study by comparing older respondents with younger respondents as we believe that age is a siginificant driver of support for assisted suicide.

```
#loads all necessary packages
library(data.table)
library(plyr)
library(ggplot2)
library(car)
```

```
## Warning: package 'car' was built under R version 3.4.3
```

```
# Pulls the features that we want to study
vars <- c("letdie1", "year", "age")

# Prompts to chosse file (we select trends-gss.csv)
sub = data.table::fread(
  file.choose(),
  sep = ",",
  select = vars)
```

```
##
Read 0.0% of 57061 rows
Read 57061 rows and 3 (of 297) columns from 0.041 GB file in 00:00:03
```

```
## the analysis will only consider non-missing data
sub <- na.omit(sub)

## recode "letdie1" so it is 1 if you want assisted suicide legalized
sub$legalsuicide = ifelse(sub$letdie1==1, 1,0)

## if you are 70 and older, you are "old" for the purposes of this analysis
sub$old = ifelse(sub$age>=70, 1,0)

## just the square of year
sub$yearsq = sub$year*sub$year
```

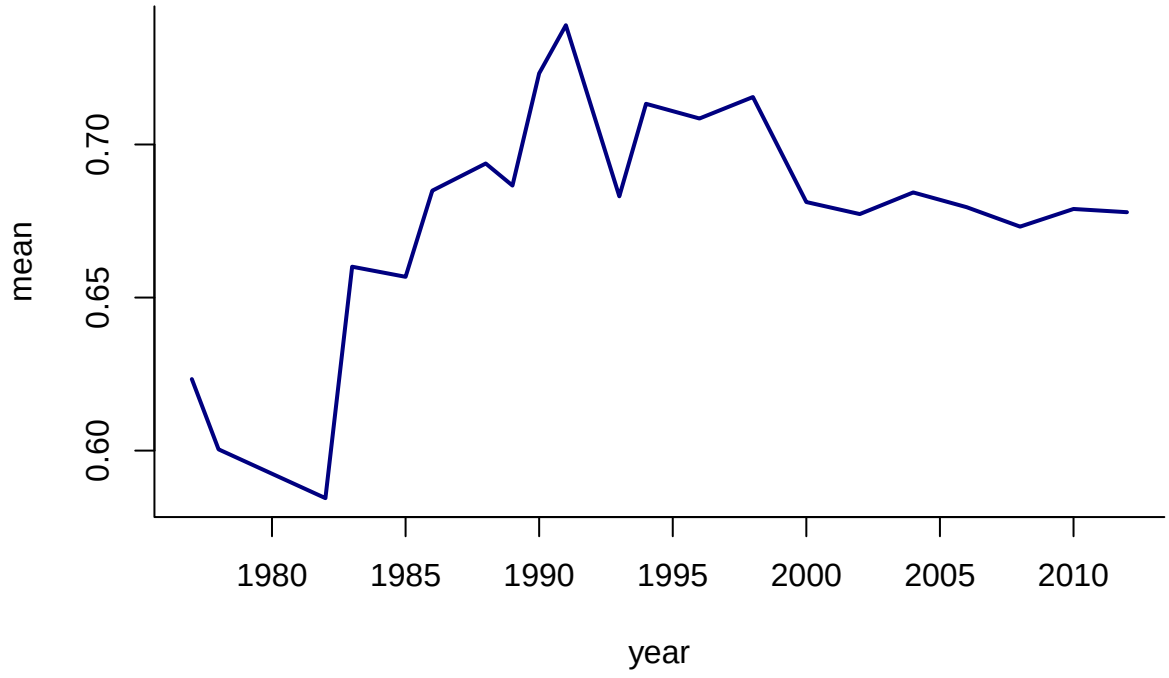
```r
by.year <- ddply(sub, # data frame to use
                 "year", # variable
                 summarise, # function to use
                 mean = mean(legalsuicide)) # create new variable "mean"

table(sub$legalsuicide)*100/length(sub$legalsuicide)
```

```
##
##        0        1
## 32.47857 67.52143
```

Across all years and participants, 32.47% do not support physician-assisted suicide will 67.52% support it.

```r
# plot the trend
plot(by.year, type = "l", lwd = 2, col = "navyblue", bty = "l")
```



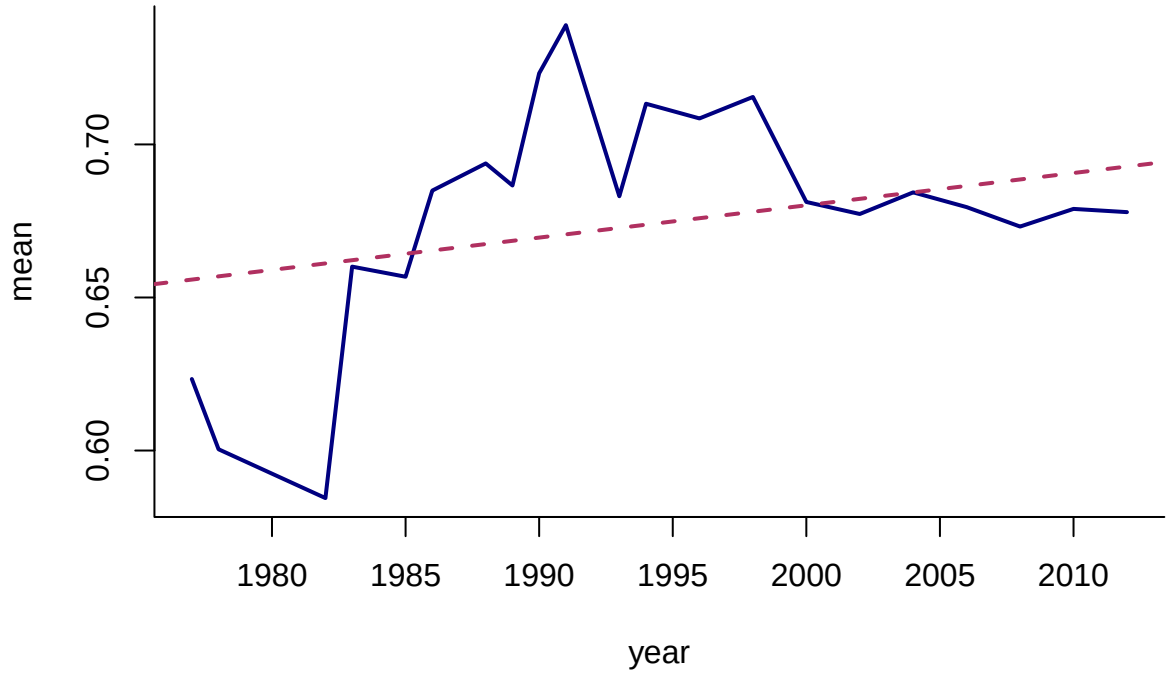
We observed that between 1977 and 2012, support for assisted suicide has approximately increased by 10% based on a population of respondents that includes all ages. More specifically, we observe that there was a sudden hike in support during 1982 followed by more moderate and also more noisy increase until 2000. From there on until 2012 the level of support has flattened at around 68%.

To capture this increase over time, we first try a plain vanilla linear model where year is entered as a numerical linear term:

```r
options(scipen=999)

# plot the trend
plot(by.year, type = "l", lwd = 2, col = "navyblue", bty = "l")
```

```r
# add a fitted line
with(by.year, abline(line(year, mean), col = "maroon", lwd = 2, lty = 2))
```




```r
#simple linear regression
lm1 = lm(legalsuicide ~ year, sub)
summary(lm1)
```

```
##
## Call:
## lm(formula = legalsuicide ~ year, data = sub)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.7108 -0.6551  0.3122  0.3334  0.3564
##
## Coefficients:
##               Estimate Std. Error t value         Pr(>|t|)
## (Intercept) -3.1548571  0.5467184  -5.771 0.00000000798535 ***
## year         0.0019213  0.0002743   7.006 0.00000000000251 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4679 on 27998 degrees of freedom
## Multiple R-squared:  0.00175,    Adjusted R-squared:  0.001714
## F-statistic: 49.08 on 1 and 27998 DF,  p-value: 0.000000000002514
```

As expected, the coefficient assigned to the term *year* is positive and statistically significant (p-value is nearly zero). This means that the support for assisted suicide on average grows by 0.2% per year. However we notice that the adjusted R^2 is only 0.0017 which means that this model only explains 0.17% of the variability in support of assisted suicide. This is pretty low so let's consider another functional form, quadratic this time:

```
# quadratic linear regression
lm2 = lm(legalsuicide ~ year + yearsq, sub)
summary(lm2)
```

```
##
## Call:
## lm(formula = legalsuicide ~ year + yearsq, data = sub)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.7058 -0.6502  0.2983  0.3287  0.4029
##
## Coefficients:
##                 Estimate   Std. Error t value            Pr(>|t|)
## (Intercept) -973.5176473  109.3539104  -8.902 <0.0000000000000002 ***
## year           0.9751481    0.1096756   8.891 <0.0000000000000002 ***
## yearsq        -0.0002440    0.0000275  -8.874 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4673 on 27997 degrees of freedom
## Multiple R-squared:  0.00455,    Adjusted R-squared:  0.004479
## F-statistic: 63.98 on 2 and 27997 DF,  p-value: < 0.00000000000000022
```

Here again we obtain statistically significant results and our adjusted R^2 has improved to 0.4%.

Let's now investigate the jump in support that occured in 1982. To do so, we create a dummy variable categorizing our dataset into two time periods, before 1982 and after. We will then create a new model including this variable.

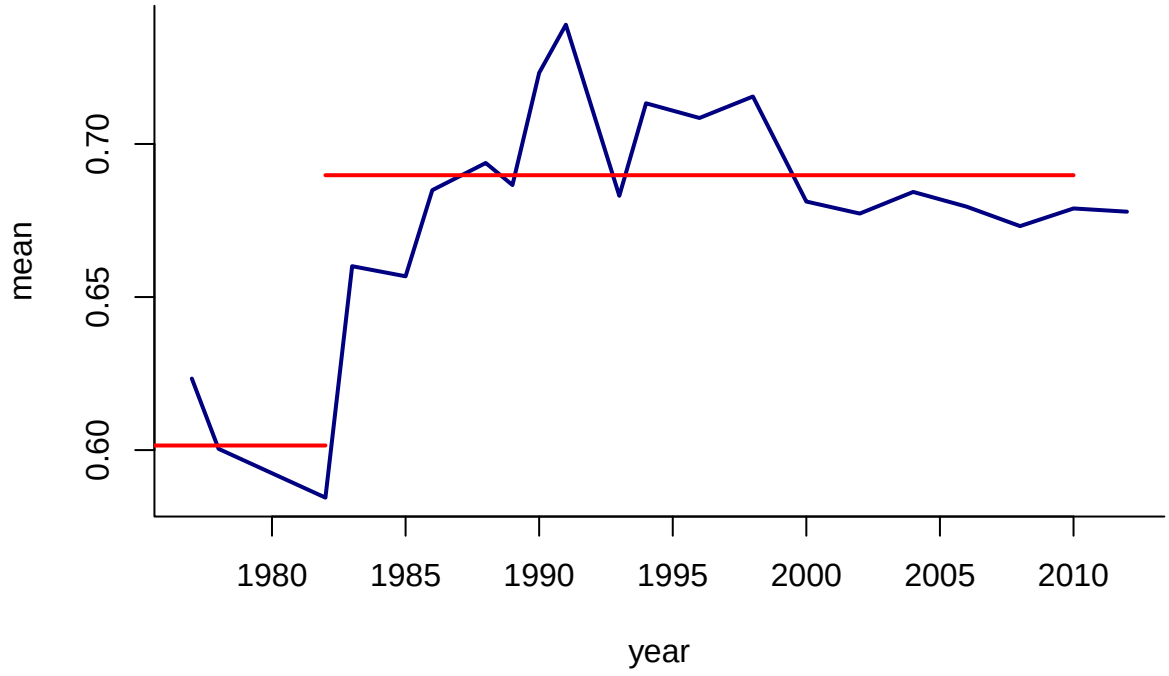
```
sub$late=ifelse((sub$year>1982), 1,0)

lm.suicide.period = lm(legalsuicide ~ late, data = sub)
summary(lm.suicide.period)
```

```
##
## Call:
## lm(formula = legalsuicide ~ late, data = sub)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.6899 -0.6899  0.3101  0.3101  0.3985
##
## Coefficients:
##             Estimate Std. Error t value            Pr(>|t|)
## (Intercept) 0.601544   0.006841   87.93 <0.0000000000000002 ***
## late        0.088390   0.007493   11.80 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4672 on 27998 degrees of freedom
```

```
## Multiple R-squared:  0.004945,   Adjusted R-squared:  0.004909
## F-statistic: 139.1 on 1 and 27998 DF,  p-value: < 0.00000000000000022
```

```
# plot the trend
plot(by.year, type = "l", lwd = 2, col = "navyblue", bty = "l")
lines(c(1972,1982), c(0.6015, 0.6015), lty=1,lwd = 2, col = "red")
lines(c(1982,2010), c(0.6898, 0.6898), lty=1,lwd = 2, col = "red")
```



Finally, I also considered usinga comprehensive set of dummy variable, one for each year.

```
ols.legalsuicide = lm(legalsuicide ~ as.factor(year), data = sub)
summary(ols.legalsuicide)
```

```
##
## Call:
## lm(formula = legalsuicide ~ as.factor(year), data = sub)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.7389 -0.6568  0.2915  0.3211  0.4155
##
## Coefficients:
##                      Estimate Std. Error t value             Pr(>|t|)
## (Intercept)           0.62336    0.01227  50.789 < 0.0000000000000002 ***
## as.factor(year)1978 -0.02295    0.01731  -1.326             0.184848
## as.factor(year)1982 -0.03888    0.01658  -2.345             0.019056 *
## as.factor(year)1983  0.03672    0.01715   2.141             0.032303 *
## as.factor(year)1985  0.03342    0.01725   1.937             0.052742 .
```

```
## as.factor(year)1986  0.06156    0.01748   3.521             0.000430 ***
## as.factor(year)1988  0.07043    0.01960   3.594             0.000326 ***
## as.factor(year)1989  0.06322    0.01947   3.247             0.001167 **
## as.factor(year)1990  0.09988    0.01997   5.001        0.00000057462 ***
## as.factor(year)1991  0.11559    0.01936   5.972        0.00000000237 ***
## as.factor(year)1993  0.05971    0.01911   3.125             0.001781 **
## as.factor(year)1994  0.08993    0.01635   5.499        0.00000003850 ***
## as.factor(year)1996  0.08513    0.01634   5.210        0.00000019039 ***
## as.factor(year)1998  0.09215    0.01658   5.558        0.00000002758 ***
## as.factor(year)2000  0.05784    0.01655   3.494             0.000477 ***
## as.factor(year)2002  0.05391    0.01996   2.701             0.006910 **
## as.factor(year)2004  0.06097    0.02004   3.042             0.002352 **
## as.factor(year)2006  0.05616    0.01625   3.455             0.000551 ***
## as.factor(year)2008  0.04981    0.02247   2.216             0.026667 *
## as.factor(year)2010  0.05557    0.01755   3.167             0.001541 **
## as.factor(year)2012  0.05453    0.01804   3.023             0.002503 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4669 on 27979 degrees of freedom
## Multiple R-squared:  0.006809,   Adjusted R-squared:  0.006099
## F-statistic: 9.591 on 20 and 27979 DF,  p-value: < 0.00000000000000022
```

```
# testing the significance of the 1982's hike in support
linearHypothesis(ols.legalsuicide, "as.factor(year)1982 = as.factor(year)1983")
```

```
## Linear hypothesis test
##
## Hypothesis:
## as.factor(year)1982 - as.factor(year)1983 = 0
##
## Model 1: restricted model
## Model 2: legalsuicide ~ as.factor(year)
##
##   Res.Df    RSS Df Sum of Sq      F      Pr(>F)
## 1  27980 6103.2
## 2  27979 6098.6  1    4.6489 21.328 0.000003887 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We obtain a model that is highly statistically significant (p-value near zero) and that has the highest adjusted R^2 of all the models (=0.6%). This model can be interpreted as follows: At the beginning of the year 1977, the support for assisted suicide was 62% and it decreases until 1982. We also see the hike in support after 1982 from a - 0.038 coefficient in 1982 to a +0.036 coefficient. SUpport reached its highest point in 1991 were support was 11% than that in 1977. As seen previously, support has been levelling in the few years trailing 2012 with a support on average 6% higher than in 1977.

```
# testing the significance of the 1982's hike in support
linearHypothesis(ols.legalsuicide, "as.factor(year)1982 = as.factor(year)1983")
```

```
## Linear hypothesis test
##
## Hypothesis:
## as.factor(year)1982 - as.factor(year)1983 = 0
##
## Model 1: restricted model
```

```
## Model 2: legalsuicide ~ as.factor(year)
##
##   Res.Df    RSS Df Sum of Sq      F      Pr(>F)
## 1  27980 6103.2
## 2  27979 6098.6  1    4.6489 21.328 0.000003887 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

A formal F test for whether concern over crime in 1982 (B= - 0.038) is different from 1983 (B=+0.036) indicates a real change, as the p-value on that F-statistic is statistically significantly different from zero.

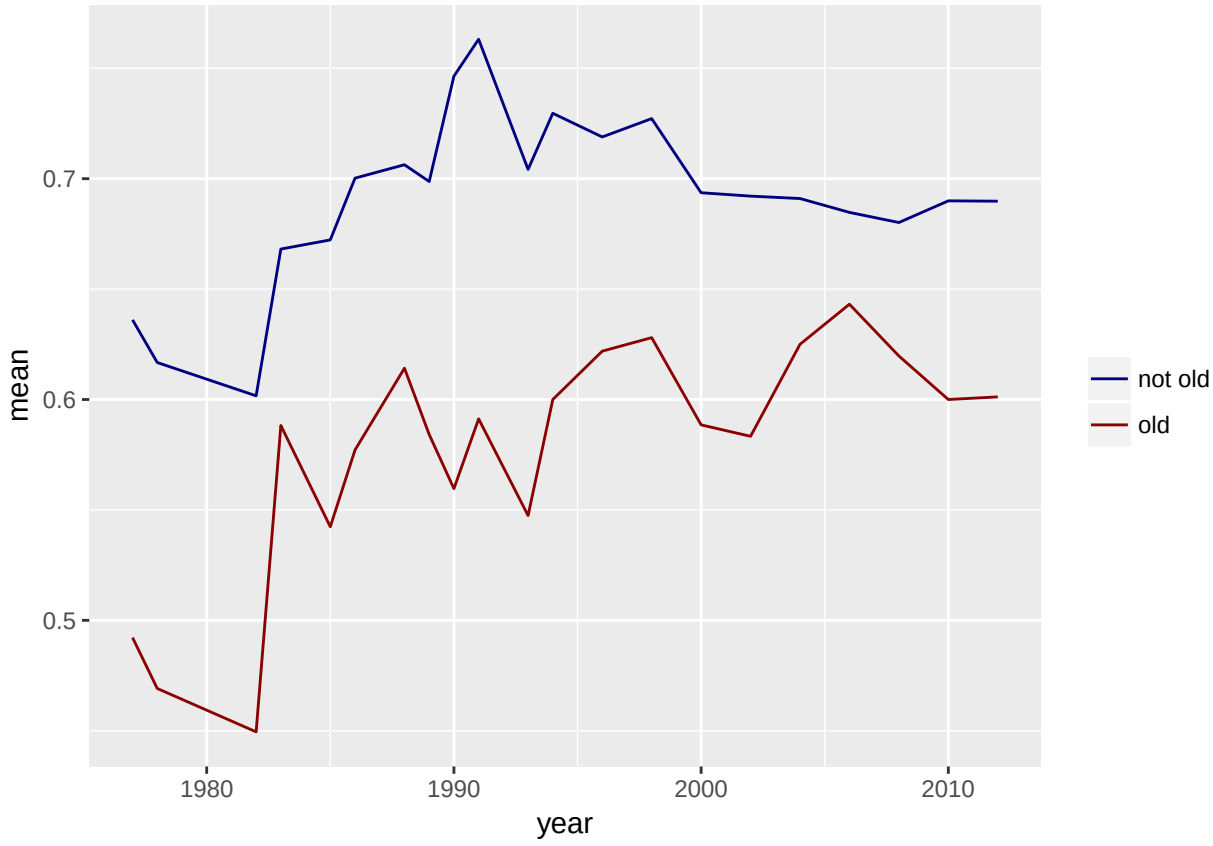
Subgroup analysis

Lastly, I want to study the difference in support between young and old respondents, with old defined as being 70 year old or more. Below the mean response is plotted for both groups against time.

```
#Graph trend over time by old
# get mean of potlegal by year and old
by.year.old <- ddply(sub, c("year", "old"), summarise,
                     mean = mean(legalsuicide, na.rm = T))

# define some labels and colors to use
color_and_labels <- scale_color_manual(values = c("navyblue", "darkred"),
                                       labels = c("not old","old"), name = "")
# set up the plot (declare x, y, grouping and coloring by married)
g_by.year.old <- ggplot(by.year.old, aes(x=year, y=mean, group=old, color = factor(old)))

# view the trend
g_by.year.old <- g_by.year.old + geom_line()
g_by.year.old + color_and_labels
```



There is a clear gap between respondents, with young people more favorable to assisted suicide than older people.

```
lm3 = lm(legalsuicide ~ year + yearsq , sub, subset = old==1)
summary(lm3)
```

```
##
## Call:
## lm(formula = legalsuicide ~ year + yearsq, data = sub, subset = old ==
##     1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.6187 -0.5802  0.3827  0.4045  0.5309
##
## Coefficients:
##                  Estimate    Std. Error t value Pr(>|t|)
## (Intercept) -861.91845732  338.46552940  -2.547   0.0109 *
## year           0.86107762    0.33938790   2.537   0.0112 *
## yearsq        -0.00021491    0.00008508  -2.526   0.0116 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4922 on 3311 degrees of freedom
## Multiple R-squared:  0.007848,   Adjusted R-squared:  0.007248
## F-statistic: 13.09 on 2 and 3311 DF,  p-value: 0.000002165
```

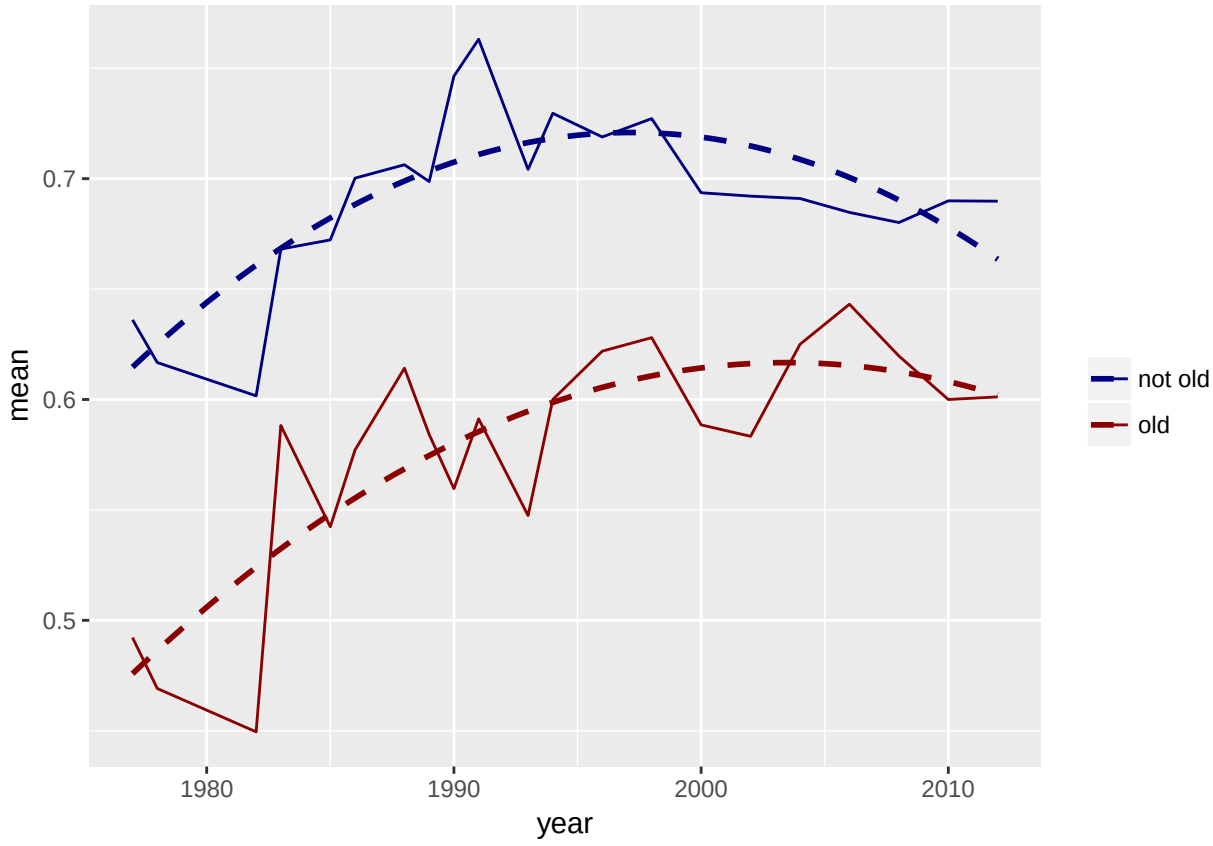
```
## the older group has lower intercept but overall identical pattern

lm4 = lm(legalsuicide ~ year + yearsq , sub, subset = old==0)
summary(lm4)
```

```
##
## Call:
## lm(formula = legalsuicide ~ year + yearsq, data = sub, subset = old ==
##     0)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.7192 -0.6576  0.2858  0.3147  0.3883
##
## Coefficients:
##                    Estimate     Std. Error t value            Pr(>|t|)
## (Intercept) -1016.12891833   115.09737590  -8.828 <0.0000000000000002 ***
## year            1.01809810     0.11543921   8.819 <0.0000000000000002 ***
## yearsq         -0.00025484     0.00002894  -8.804 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4622 on 24683 degrees of freedom
## Multiple R-squared:  0.004614,   Adjusted R-squared:  0.004533
## F-statistic:  57.2 on 2 and 24683 DF,  p-value: < 0.00000000000000022
```

The intercept for young people is larger than for older people, so young people are indeed more supportive of assisted suicide than the other group. We also notice that the gap has been tightening over the years since the *year* coefficient for young people is smaller than for older people (0.86 vs 1.01). This means that although older people were more conservative regarding assisted suicide in 1977, they have grown more supportive faster tha younger people. This pattern will be more obvious if we fit quadratic trend lines over the data.

```
# use a quadratic fit
g_quad <- g_by.year.old + stat_smooth(method = "lm", formula = y ~ poly(x,2), se = F, lty = 2)
g_quad + color_and_labels
```



Indeed, we clearly see a tightening effect on the gap between older and younger respondents which means that opinion is becoming more homegeneous. To test this systematically we are running a regression interracting the *old* term with the post-1982 period.

```
lm.legalsuicide.period.int = lm(legalsuicide ~ late*as.factor(old), data = sub)
summary(lm.legalsuicide.period.int)
```

```
##
## Call:
## lm(formula = legalsuicide ~ late * as.factor(old), data = sub)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.7027 -0.6172  0.2973  0.2973  0.5328
##
## Coefficients:
##                       Estimate Std. Error t value             Pr(>|t|)
## (Intercept)           0.617246   0.007208  85.636 < 0.0000000000000002 ***
## late                  0.085499   0.007907  10.813 < 0.0000000000000002 ***
## as.factor(old)1      -0.150032   0.022281  -6.734      0.0000000000169 ***
## late:as.factor(old)1  0.044244   0.024161   1.831               0.0671 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4657 on 27996 degrees of freedom
## Multiple R-squared:  0.01107,    Adjusted R-squared:  0.01097
## F-statistic: 104.5 on 3 and 27996 DF,  p-value: < 0.00000000000000022
```

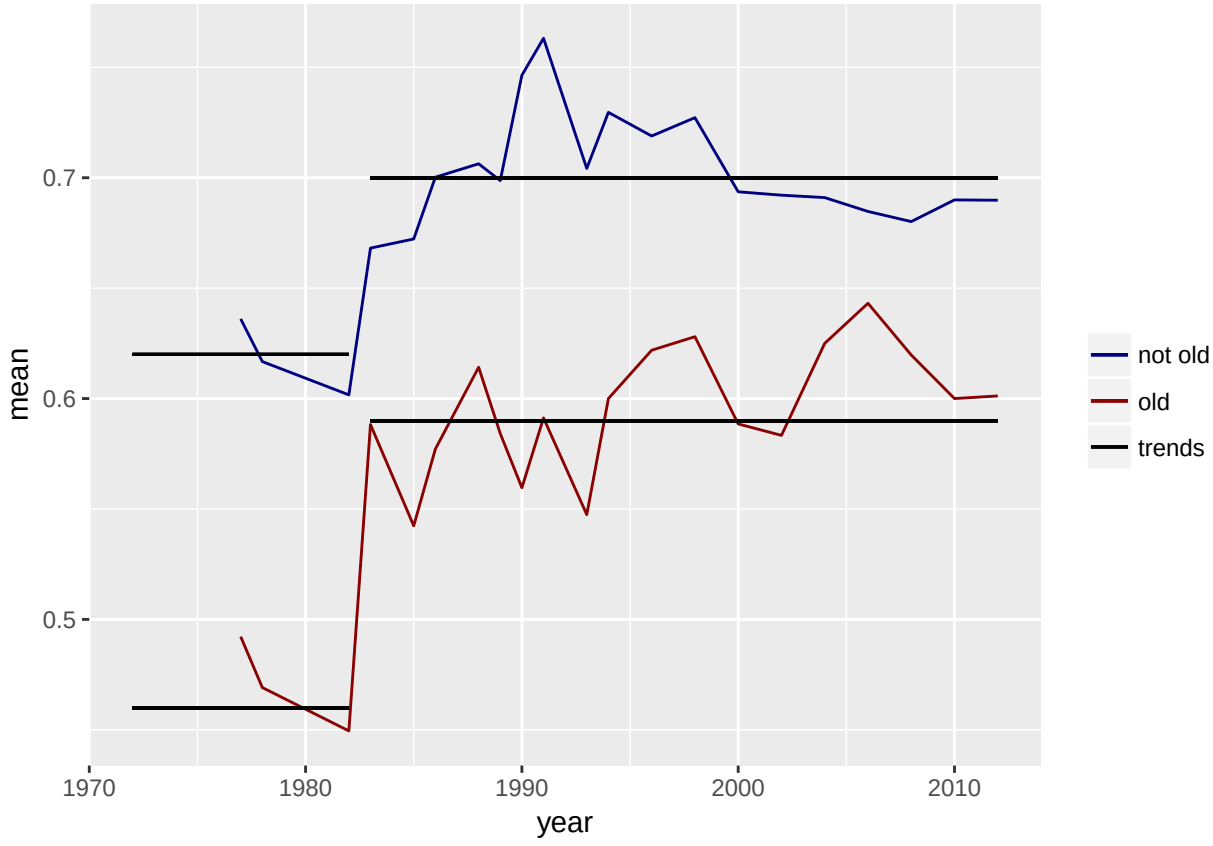
The results from the regression show that in the earlier period, old people were 15% points less supportive of assisted assisted-suicide than young people. Young people's support went up 8.5% in the later period. But old people's support went up even further by 13% (= 0.085 + 0.044) in the later period. The interaction term is somewhat statistically significant ($p < 0.07$) but we will leave it here. The adjusted R^2 of 1.1% in this model is a lot higher than in the previous models because age is in general a meaningful predictor of support for assisted suicide.

The fit of this model is graphed below:

```
#Graph trend over time by old
# get mean of potlegal by year and old
by.year.old <- ddply(sub, c("year", "old"), summarise,
                     mean = mean(legalsuicide, na.rm = T))

# define some labels and colors to use
color_and_labels <- scale_color_manual(values = c("navyblue", "darkred", "black"),
                                       labels = c("not old","old", "trends"), name = "")
# set up the plot (declare x, y, grouping and coloring by married)
g_by.year.old <- ggplot(by.year.old, aes(x=year, y=mean, group=old, color = factor(old)))

# view the trend
g_by.year.old <- g_by.year.old + geom_line()
g_by.year.old + color_and_labels +
  geom_segment(aes(x = 1972, y = 0.46, xend = 1982, yend = 0.46, colour = "trends")) +
  geom_segment(aes(x = 1983, y = 0.59, xend = 2012, yend = 0.59, colour = "trends")) +
  geom_segment(aes(x = 1972, y = 0.62, xend = 1982, yend = 0.62, colour = "trends")) +
  geom_segment(aes(x = 1983, y = 0.70, xend = 2012, yend = 0.70, colour = "trends"))
```



Interestingly we observe that support for assisted suicide among old people in 2012 has almost reached the level of support shown by young people 35 years before in 1977.