Lab 3 - Education and Life Satisfaction Time Series

Christine Lucille Kuryla

2024-12-12

Introduction

Does having a college degree increase life excitement and satisfaction? Naively, there seems to be a relationship. Our question in this lab is whether having a bachelor degree is related to life satisfaction over time in the way they are related at the individual level. In other words, as more of our population is getting bachelor's and college degrees, are their lifes getting more exciting (higher life satisfaction)?

The variables we will be using come from the GSS dataset.

Variable: life

"In general, do you find life exciting, pretty routine, or dull?"

1 Exciting
2 Routine 3 Dull

We will recode to the more intuitive 3 Exciting
2 Routine 1 Dull

We will also dichotomize it so that exciting_life = 1 if people respond "exciting" and exciting_life = 0 otherwise.

Variable: degree / baplus

0 Less than high school
1 High school 2 Associate/junior college
3 Bachelor's
4 Graduate

We will dichotomize it so that baplus = 1 if people have a bachelors or graduate degree and baplus = 0 otherwise.

Naive Relationship

```
gss_for_ts <- read_csv("data/gss_raw_subset_for_ts.csv") %>%
       mutate(happiness = ifelse(happy == 1, 1, 0),
              excellent_health = ifelse(health == 1, 1, 0),
              good_health = ifelse(health == 2, 1, 0),
              fair_health = ifelse(health == 3, 1, 0),
              poor_health = ifelse(health == 4, 1, 0),
              exciting_life = ifelse(life == 1, 1, 0),
              nonextreme_views = ifelse(polviews %in% c(3, 4, 5), 1, 0),
              extreme_views = ifelse(polviews %in% c(1, 2, 6, 7), 1, 0),
              moderate_views = ifelse(polviews == 4, 1, 0),
              cohort = floor(year - age),
```

```
              over50 = ifelse(age >= 50, 1, 0),
              boomer = ifelse(cohort >= 1946 & cohort <= 1964, 1, 0) ,
              millenial = ifelse(cohort >= 1981 & cohort <= 1996, 1, 0) ,
              bornin40s = ifelse(cohort >= 1940 & cohort <= 1949, 1, 0),
              sat_w_finances = ifelse(satfin == 1, 1, 0),
              income = realinc) %>%
    mutate(life = 4 - life) %>% # more intuitive
    mutate(baplus = ifelse(degree >= 3, 1, 0)) %>% # Dichotomise bachelors/graduate vs no bachelors/gra
    mutate(happy = 4 - happy) %>%
    mutate(health = 5 - health)
```

```
## Rows: 72390 Columns: 14
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (14): year, trust, sex, age, partyid, wrkstat, happy, degree, realinc, h...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
# gss_for_ts <- read_csv("data/gss_raw_subset_for_ts.csv") %>%
#   mutate(life = 4 - life) %>% # more intuitive
#   mutate(baplus = ifelse(degree >= 3, 1, 0)) %>% # Dichotomise bachelors/graduate vs no bachelors/gra
#   mutate(exciting_life = ifelse(life == 1, 1, 0)) %>%
#   mutate(happy = 4 - happy) %>%
#   mutate(health = 5 - health)

colnames(gss_for_ts)
```

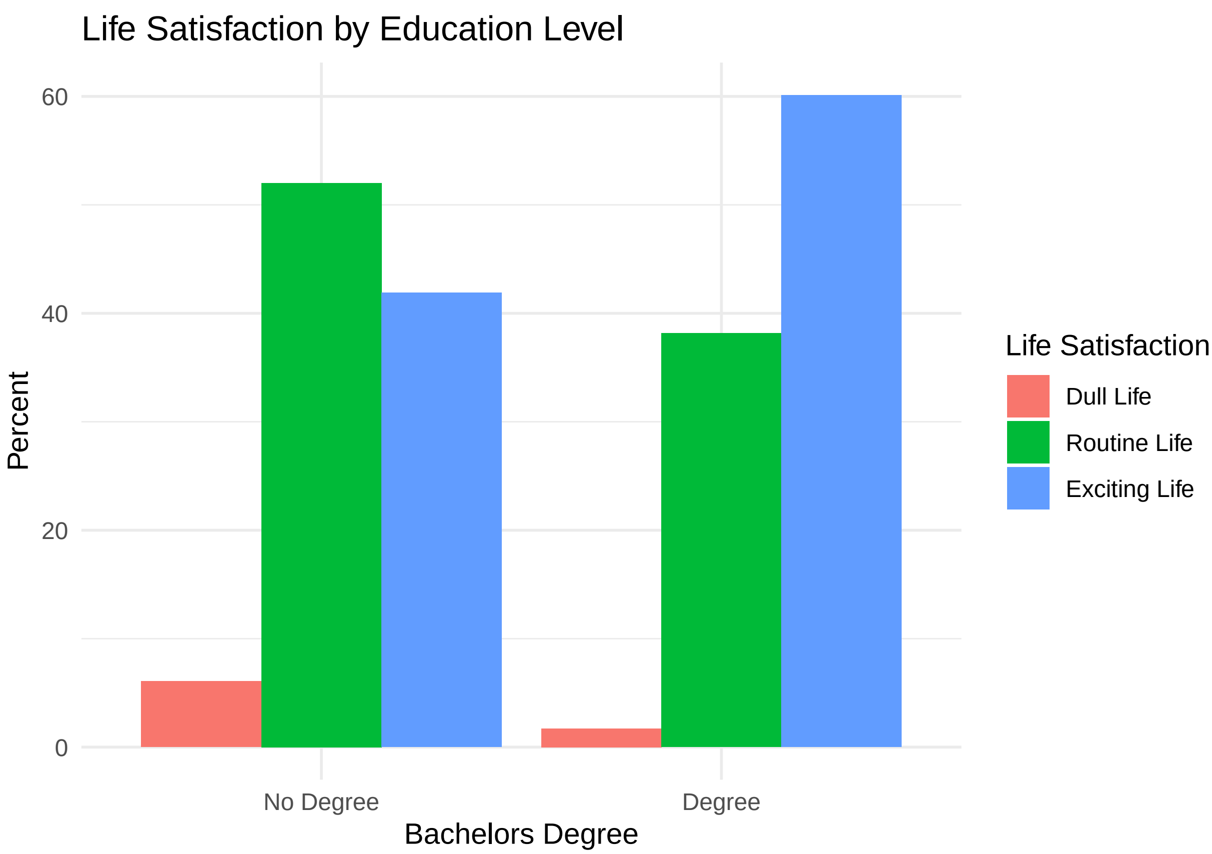
```
##  [1] "year"             "trust"            "sex"              "age"
##  [5] "partyid"          "wrkstat"          "happy"            "degree"
##  [9] "realinc"          "health"           "life"             "satfin"
## [13] "polviews"         "educ"             "happiness"        "excellent_health"
## [17] "good_health"      "fair_health"      "poor_health"      "exciting_life"
## [21] "nonextreme_views" "extreme_views"    "moderate_views"   "cohort"
## [25] "over50"           "boomer"           "millenial"        "bornin40s"
## [29] "sat_w_finances"   "income"           "baplus"
```

```
# Correlation between the variables?
gss_for_ts %>%
  dplyr::select(degree, life) %>%
  na.omit() %>%
  cor()
```

```
##           degree      life
## degree 1.0000000 0.1976825
## life   0.1976825 1.0000000
```

```
# Yes, and it is positive.

# Dichotomize bachelors vs no bachelors and visualize it
gss_for_ts %>%
  select(baplus, life, year) %>%
  na.omit() %>%
  mutate(life = factor(life,
                       levels = 1:3,
                       labels = c("Dull Life", "Routine Life", "Exciting Life"))) %>%
```

```r
  mutate(baplus = factor(baplus,
                   levels = c(0, 1),
                   labels = c("No Degree", "Degree"))) %>%
  dplyr::count(life, baplus) %>%
  group_by(baplus) %>%
  mutate(percent = n / sum(n) * 100) %>%
    ggplot(aes(x = factor(baplus), y = percent, fill = factor(life))) +
    geom_bar(stat = "identity", position = "dodge") +
    labs(
        title = "Life Satisfaction by Education Level",
        x = "Bachelors Degree",
        y = "Percent",
        fill = "Life Satisfaction"
    ) +
    theme_minimal()
```



Time Series Analysis

Q1: Create time series and interpolate

1. Create a multivariate time series; perform any interpolations.

```r
# get means by year
by.year <- aggregate(subset(gss_for_ts, sel = -year), list(year = gss_for_ts$year), mean, na.rm = T)

# interpolate for some missing years
```

```r
# First, add the extra years
unique(gss_for_ts$year) # years in dataset
```

```
##  [1] 1972 1973 1974 1975 1976 1977 1978 1980 1982 1983 1984 1985 1986 1987 1988
## [16] 1989 1990 1991 1993 1994 1996 1998 2000 2002 2004 2006 2008 2010 2012 2014
## [31] 2016 2018 2021 2022
```

```r
extra_years <- setdiff(seq(1972, 2018), unique(gss_for_ts$year)) # years missing for a continus TS; ski
dim(by.year)[1] # number of years in original data (34)
```

```
## [1] 34
```

```r
length(extra_years) # number of years to add (15)
```

```
## [1] 15
```

```r
dim(by.year)[1] + length(extra_years) # sum (49)
```

```
## [1] 49
```

```r
by.year[35:49, "year"] <- as.vector(extra_years) # add the extra years
by.year <- dplyr::arrange(by.year, year) # arrange by year

# Now make a time series object by.year.ts and interpolate using na.approx
by.year.ts <- ts(by.year)
by.year.ts <- na.approx(by.year.ts)

# calculate pct
by.year.ts <- as.data.frame(by.year.ts)
by.year.ts <- mutate(by.year.ts,
                     happy_pct = happiness*100,
                     excellent_health_pct = excellent_health*100,
                     exciting_life_pct = exciting_life*100,
                     boomer_pct = boomer*100,
                     bornin40s_pct = bornin40s*100,
                     over50_pct = over50*100,
                     sat_w_finances_pct = sat_w_finances*100,
                #     millenial_pct = millenial*100,
                     ba_pct = baplus*100)

# get rid of 1972,1973, after 2018 and convert back to time series object
 by.year.ts <- ts(subset(by.year.ts, year >= 1974 & year <= 2018))

head(by.year.ts)
```

```
## Time Series:
## Start = 1
## End = 6
## Frequency = 1
##   year    trust      sex      age  partyid  wrkstat    happy    degree  realinc
## 1 1974 1.608610 1.534367 44.59134 2.598220 3.585580 2.247973 0.9986514 32124.53
## 2 1975 1.648173 1.550336 44.30774 2.501010 3.575168 2.197980 0.9523170 29403.92
## 3 1976 1.593311 1.553702 45.28667 2.430769 3.625083 2.215477 0.9886135 28273.75
## 4 1977 1.623226 1.547059 44.66316 2.405797 3.228105 2.229208 0.9986877 32640.56
## 5 1978 1.653141 1.580287 44.00984 2.590701 3.355744 2.247858 1.0392413 30178.04
## 6 1979 1.617810 1.571819 44.49224 2.601494 3.346469 2.226870 1.0636780 30755.62
##     health     life   satfin polviews     educ happiness excellent_health
```

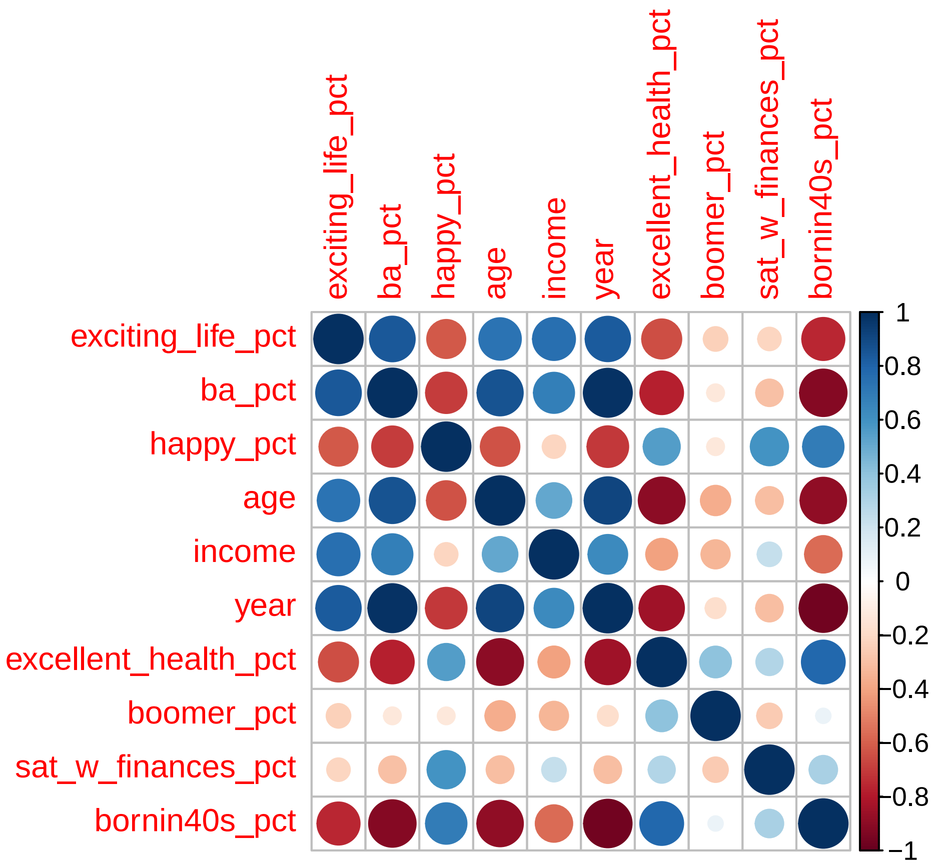
```
## 1 2.993243 2.387656 1.920162 3.979433 11.80081 0.3790541        0.3283784
## 2 2.980524 2.399252 1.956728 3.960630 11.68258 0.3286195        0.3237072
## 3 2.976636 2.410847 1.926273 4.020700 11.70127 0.3408939        0.3130841
## 4 2.975115 2.376337 1.879684 4.043359 11.68816 0.3483955        0.3176162
## 5 2.980226 2.385410 1.899935 4.096167 11.91743 0.3434410        0.3177014
## 6 2.985338 2.394483 1.941759 4.115263 11.96418 0.3413511        0.3177866
##   good_health fair_health poor_health exciting_life nonextreme_views
## 1   0.3979730   0.2121622  0.06148649     0.4348128        0.6684636
## 2   0.3975823   0.2142377  0.06447280     0.4411352        0.6624161
## 3   0.4198932   0.1975968  0.06942590     0.4474576        0.6444296
## 4   0.4086444   0.2049771  0.06876228     0.4438503        0.6725490
## 5   0.4124933   0.2021357  0.06766962     0.4491150        0.6873368
## 6   0.4163421   0.1992943  0.06657697     0.4543796        0.7006166
##   extreme_views moderate_views   cohort    over50    boomer millenial bornin40s
## 1     0.2816712      0.4000000 1929.409 0.3903924 0.2368065         0 0.2212449
## 2     0.2751678      0.4001432 1930.692 0.3824916 0.2727273         0 0.2121212
## 3     0.2901935      0.3990007 1930.713 0.4065640 0.2833222         0 0.2277294
## 4     0.2771242      0.3881624 1932.337 0.3919895 0.2843073         0 0.2048588
## 5     0.2493473      0.3825784 1933.990 0.3718033 0.3501639         0 0.2249180
## 6     0.2544420      0.3949281 1934.508 0.3822690 0.3615110         0 0.2173254
##   sat_w_finances   income    baplus happy_pct excellent_health_pct
## 1      0.3119080 32124.53 0.1429535  37.90541             32.83784
## 2      0.3096687 29403.92 0.1276024  32.86195             32.37072
## 3      0.3069705 28273.75 0.1426658  34.08939             31.30841
## 4      0.3418803 32640.56 0.1397638  34.83955             31.76162
## 5      0.3387835 30178.04 0.1393067  34.34410             31.77014
## 6      0.3120046 30755.62 0.1478638  34.13511             31.77866
##   exciting_life_pct boomer_pct bornin40s_pct over50_pct sat_w_finances_pct
## 1          43.48128   23.68065      22.12449   39.03924           31.19080
## 2          44.11352   27.27273      21.21212   38.24916           30.96687
## 3          44.74576   28.33222      22.77294   40.65640           30.69705
## 4          44.38503   28.43073      20.48588   39.19895           34.18803
## 5          44.91150   35.01639      22.49180   37.18033           33.87835
## 6          45.43796   36.15110      21.73254   38.22690           31.20046
##     ba_pct
## 1 14.29535
## 2 12.76024
## 3 14.26658
## 4 13.97638
## 5 13.93067
## 6 14.78638
```

```r
colnames(by.year.ts)
```

```
##  [1] "year"                 "trust"                "sex"
##  [4] "age"                  "partyid"              "wrkstat"
##  [7] "happy"                "degree"               "realinc"
## [10] "health"               "life"                 "satfin"
## [13] "polviews"             "educ"                 "happiness"
## [16] "excellent_health"     "good_health"          "fair_health"
## [19] "poor_health"          "exciting_life"        "nonextreme_views"
## [22] "extreme_views"        "moderate_views"       "cohort"
## [25] "over50"               "boomer"               "millenial"
## [28] "bornin40s"            "sat_w_finances"       "income"
## [31] "baplus"               "happy_pct"            "excellent_health_pct"
```

```
## [34] "exciting_life_pct"    "boomer_pct"           "bornin40s_pct"
## [37] "over50_pct"           "sat_w_finances_pct"   "ba_pct"
```

```r
# correlations
cor.vars <- c("exciting_life_pct", "ba_pct", "happy_pct", "age", "income", "year", "excellent_health_pc
# cor.vars <- colnames(by.year.ts)
cor.dat <- by.year.ts[, cor.vars]


corrplot::corrplot(cor(cor.dat))
```



Exiting life and percent with a higher degree are positively correlated with year, which suggests that both increase as time passes. The percent of the population with a bachelors or higher is positively correlated with people reporting a more exciting life, which suggestions that as the proportion of the population with a bachelors or higher increases, life satisfaction/ excitement also tends to increase.

Q2: Graph relationships

2. Graph the relationships between X and Y. Explain how you think Y should relate to your key Xs.

```r
library(reshape2)


meltMyTS <- function(mv.ts.object, time.var, keep.vars){
  # mv.ts.object = a multivariate ts object
  # keep.vars = character vector with names of variables to keep
  # time.var = character string naming the time variable
```

```r
  require(reshape2)

  if(missing(keep.vars)) {
    melt.dat <- data.frame(mv.ts.object)
  }
  else {
    if (!(time.var %in% keep.vars)){
      keep.vars <- c(keep.vars, time.var)
    }
    melt.dat <- data.frame(mv.ts.object)[, keep.vars]
  }
  melt.dat <- melt(melt.dat, id.vars = time.var)
  colnames(melt.dat)[which(colnames(melt.dat) == time.var)] <- "time"
  return(melt.dat)
}

# Make a character vector naming the variables we might want to plot
keep.vars <- c("year", "happy_pct", "age", "ba_pct", "income", "excellent_health_pct", "exciting_life_p
keep.vars <- setdiff(colnames(by.year.ts), "year")

# Use meltMyTS to transform the data to a 3-column dataset containing a column
# for time, a column for variable names, and a column of values corresponding to
# the variable names


plot.dat <- meltMyTS(mv.ts.object = by.year.ts, time.var = "year", keep.vars = keep.vars)
head(plot.dat)
```

```
##   time variable    value
## 1 1974    trust 1.608610
## 2 1975    trust 1.648173
## 3 1976    trust 1.593311
## 4 1977    trust 1.623226
## 5 1978    trust 1.653141
## 6 1979    trust 1.617810
```

```r
# Use ggMyTS to plot any of the variables or multiple variables together


ggMyTS <- function(df, varlist, line = TRUE, point = TRUE, pointsize = 3, linewidth = 1.25, ...){
  require(ggplot2)
  # varlist = character vector with names of variables to use
  if(missing(varlist)){
    gg <- ggplot(df, aes(time, value, colour = variable))
  }
  else{
    include <- with(df, variable %in% varlist)
    gg <- ggplot(df[include,], aes(time, value, colour = variable))
  }
  if(line == FALSE & point == FALSE) {
    stop("At least one of 'line' or 'point' must be TRUE")
  }
  else{
    if(line == TRUE) gg <- gg + geom_line(size = linewidth, aes(color = variable), ...)
```

```
    if(point == TRUE) gg <- gg + geom_point(size = pointsize, aes(color = variable), ...)
  }

  gg + xlab("") + theme(legend.position = "bottom") + scale_x_continuous(breaks = min(df$time):max(df$ti
}
```

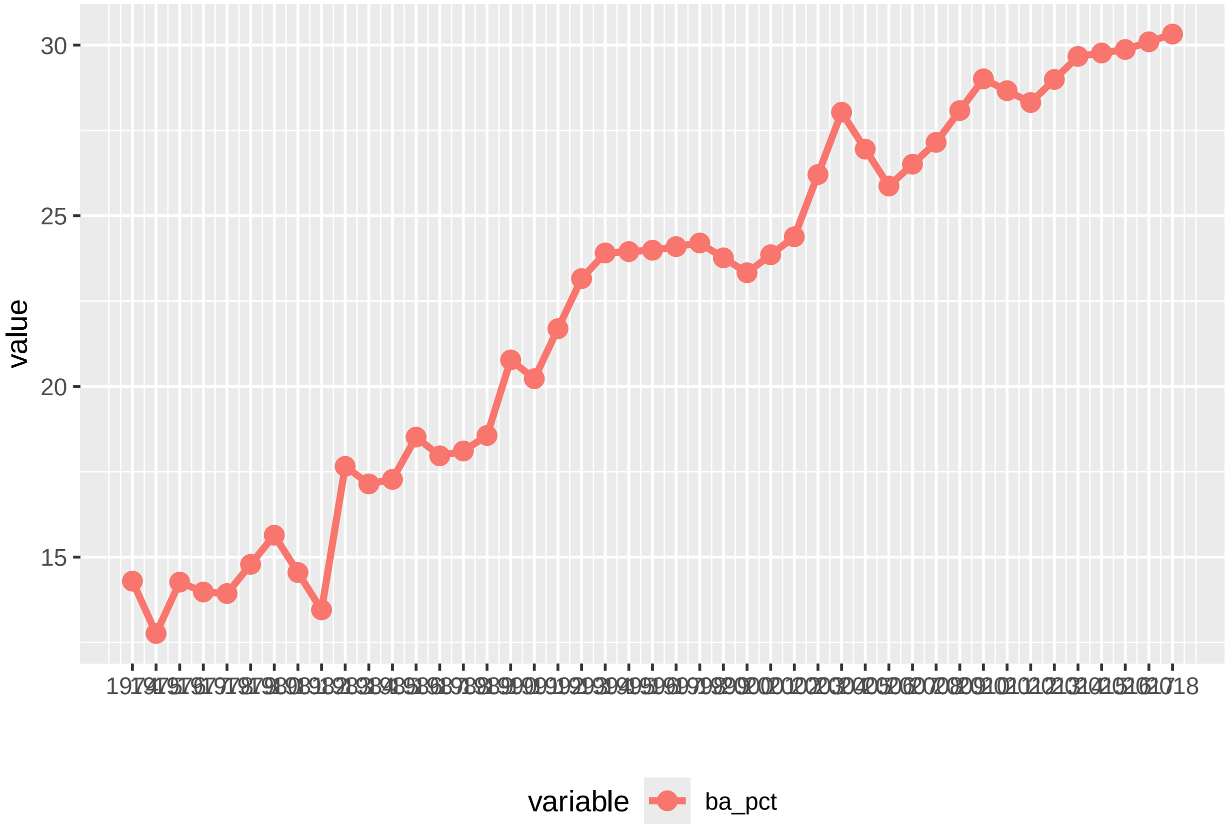
Key Xs

I expect that the percentage of the population with bachelor's or higher will increase through the years because in modern society, people are getting more higher degrees. I expect percent of people who rate their life exciting to be increasing as well, because the world has become interesting, and, because it is possible that more people having higher degrees enables them to make time for an exciting life. I expect financial satisfaction to decrease or fluctuate, because comparing to external wealth has become worse as time has passed. I would normally expect happiness and health to increase, but from my previous work, I know they decrease.

Let's first look at the percentage of the population with bachelor's or higher:

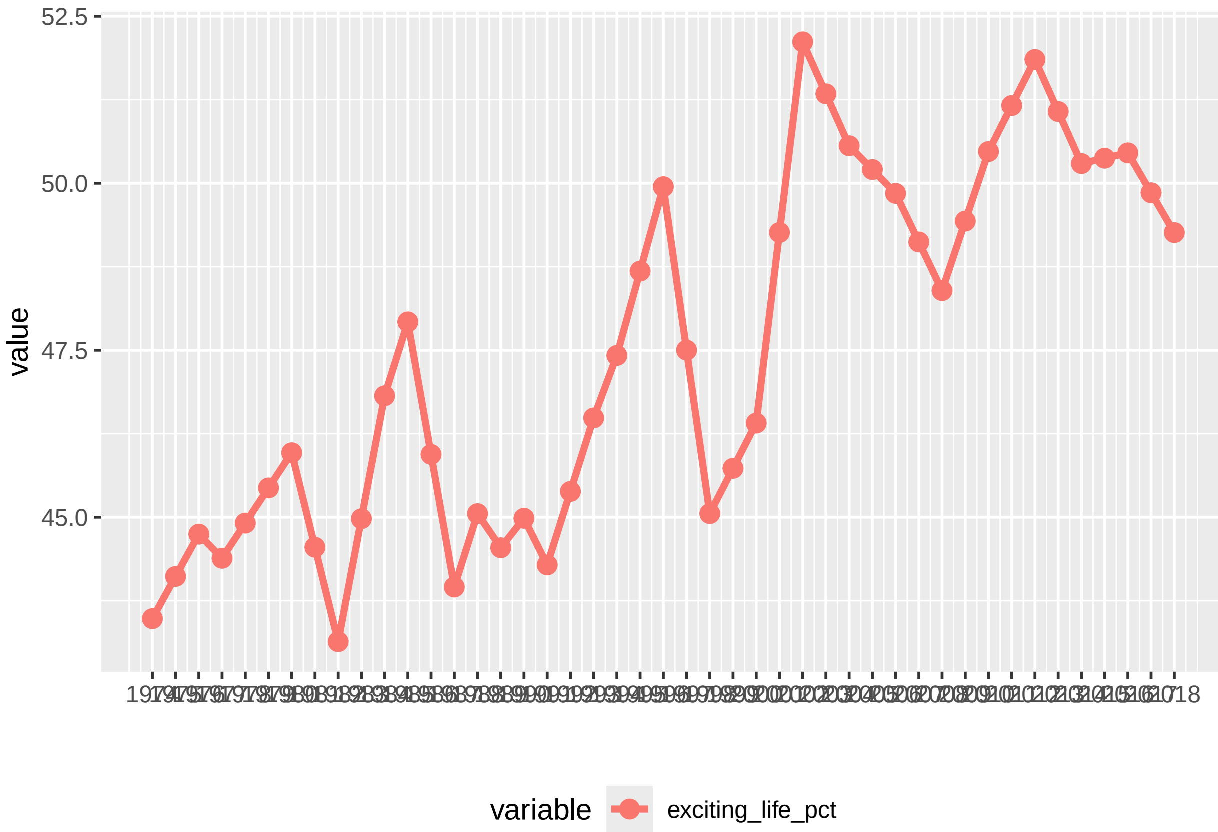
```
ggMyTS(df = plot.dat, varlist = c("ba_pct"))
```

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



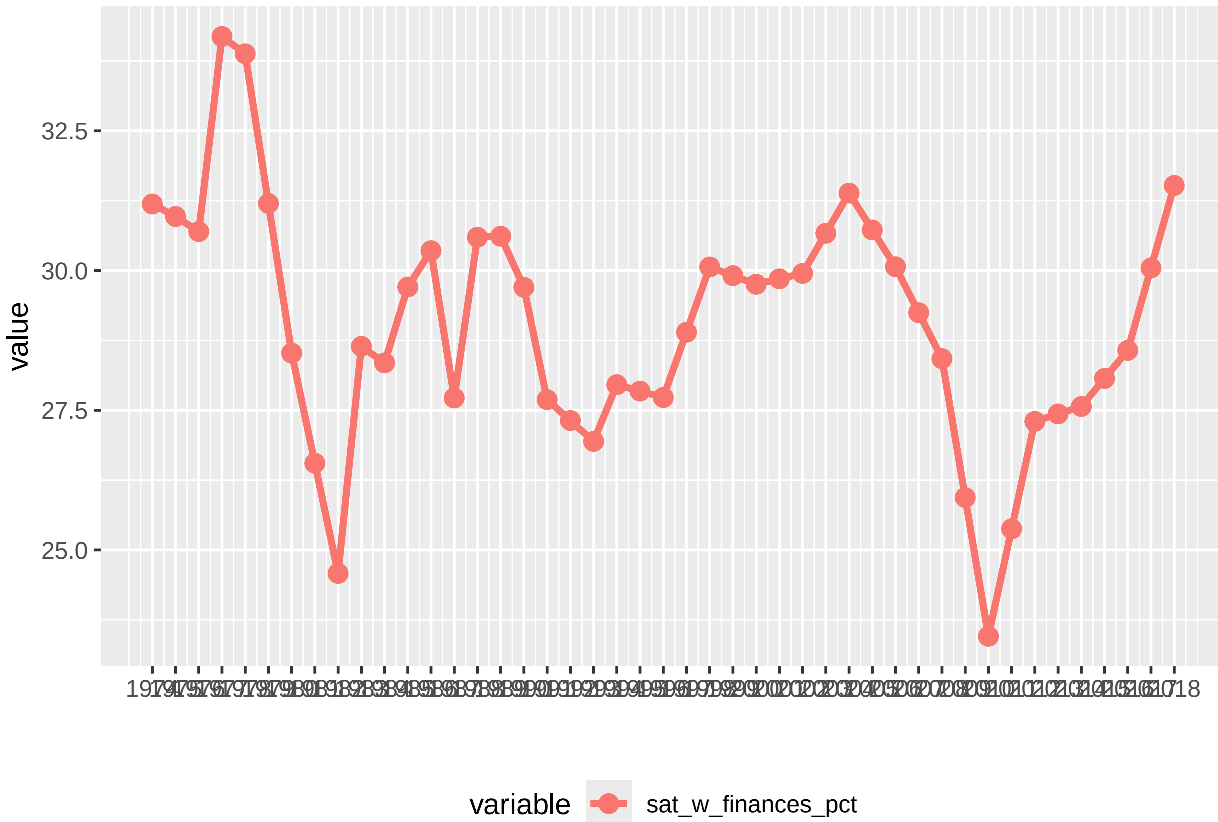
Life satisfaction

```r
ggMyTS(df = plot.dat, varlist = c("exciting_life_pct"))
```



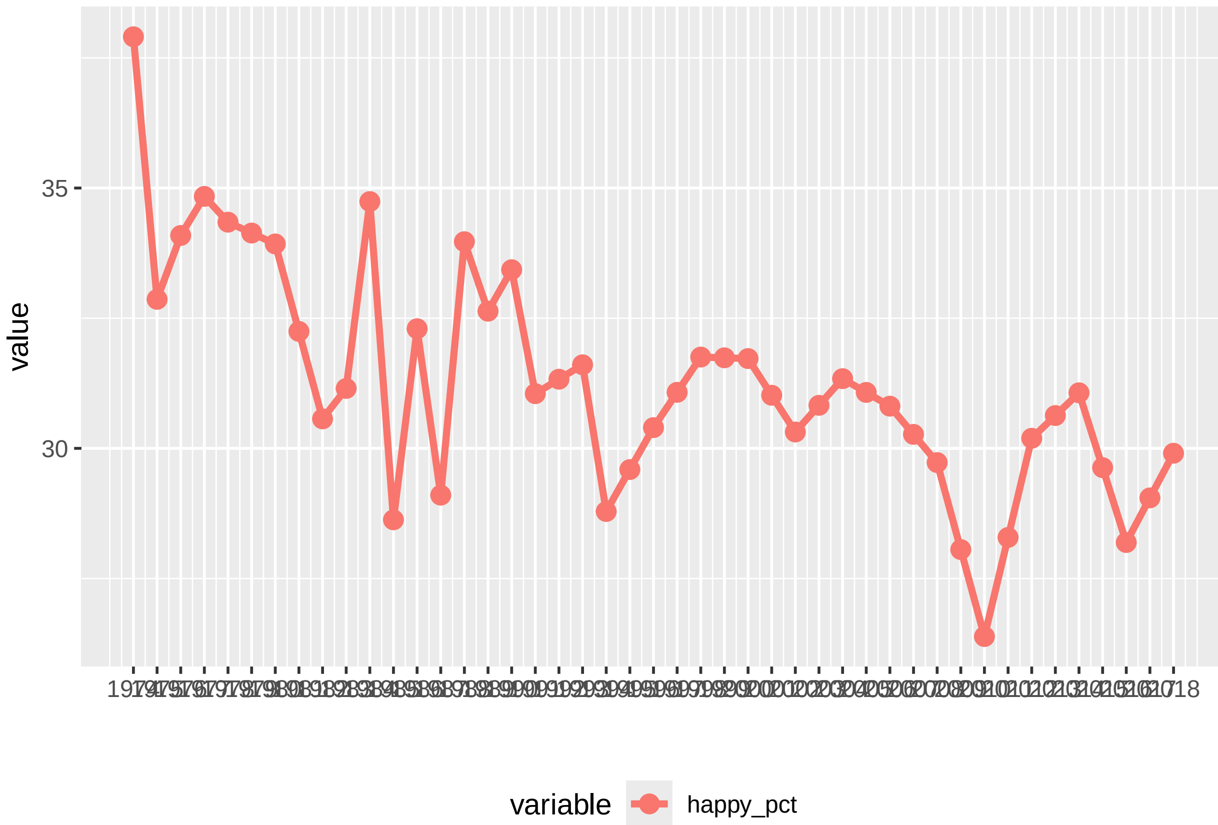
Financial satisfaction

```r
ggMyTS(df = plot.dat, varlist = c("sat_w_finances_pct"))
```



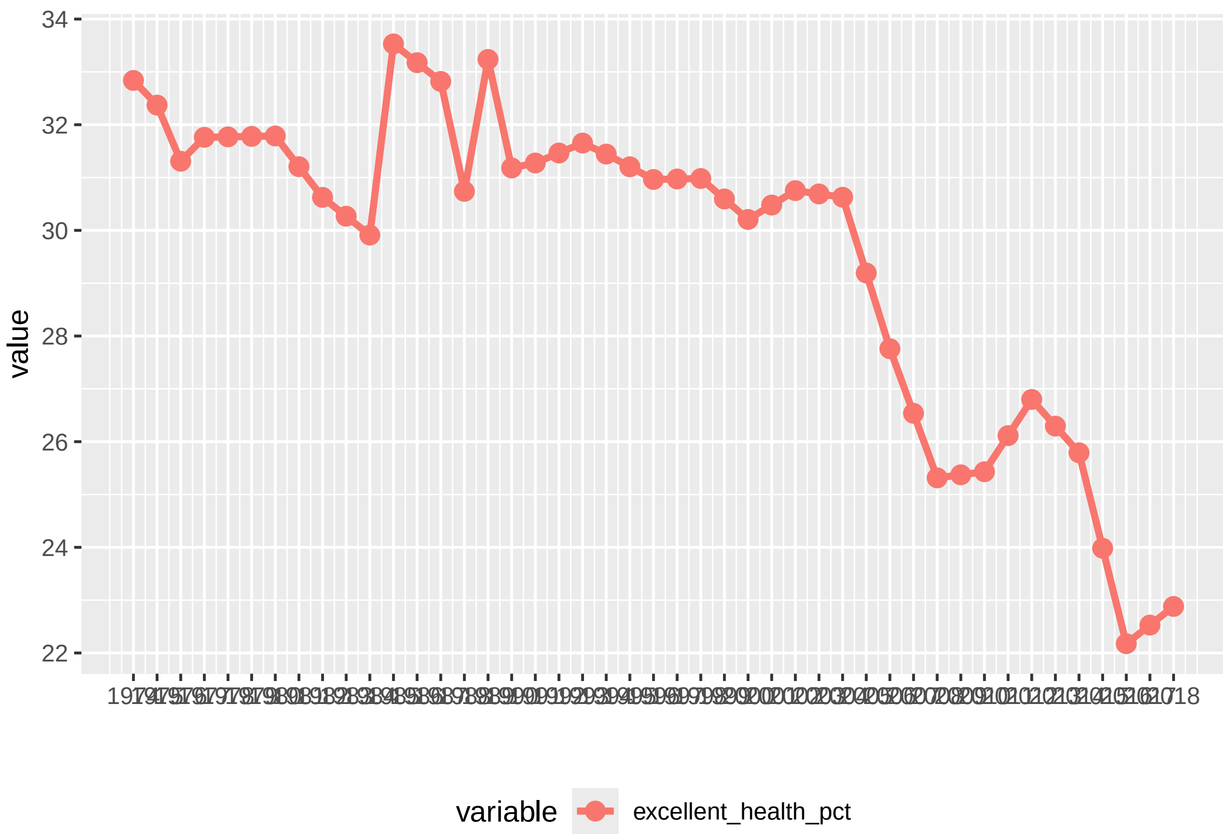
Happiness

```
(g_happy_pct <- ggMyTS(df = plot.dat, varlist = c("happy_pct")))
```



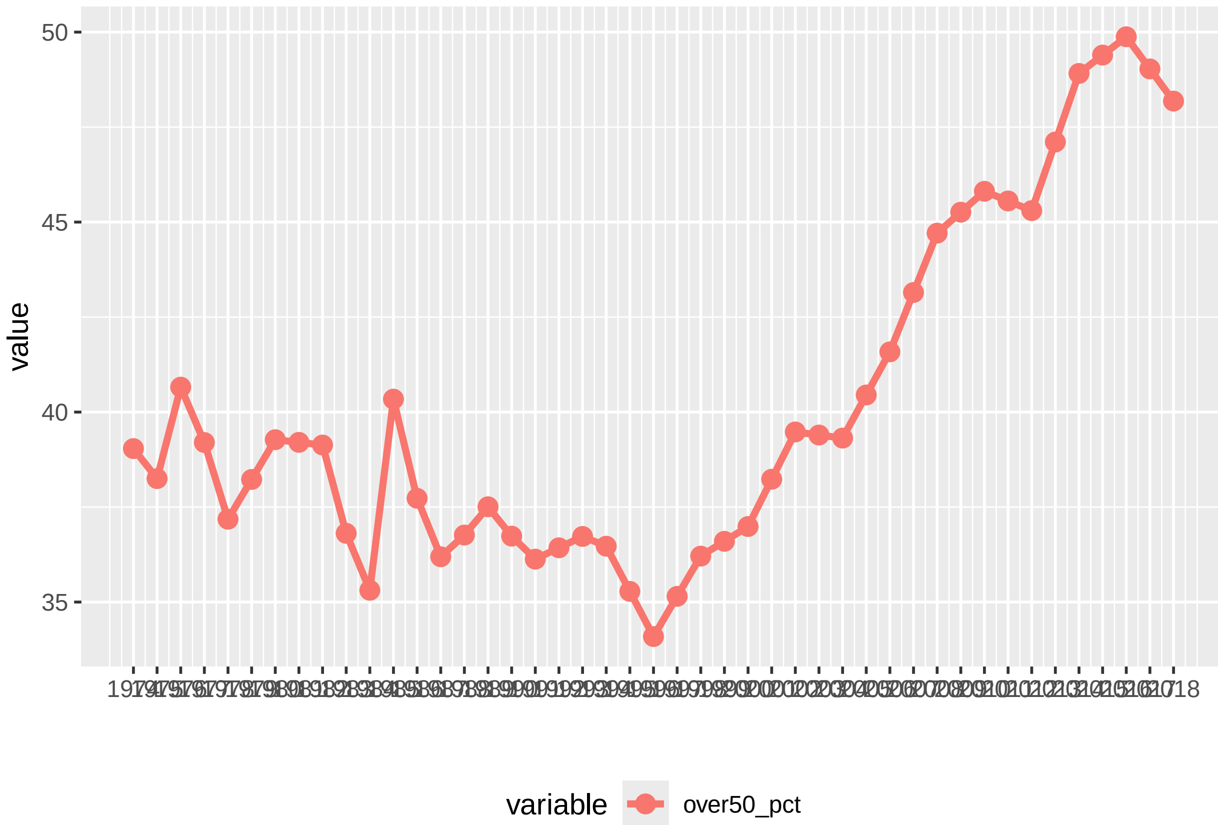
Self-Rated Health

```
ggMyTS(df = plot.dat, varlist = c("excellent_health_pct"))
```



Age Composition

```
ggMyTS(df = plot.dat, varlist = c("over50_pct"))
```



As expected, the percentage of the population with bachelor's or higher increased through the years, as did percentage of people who rate their life as exciting. Financial satisfaction fluctuates. As stated before, health and happiness are decreasing. However, there are some cohort and age effects happening, and I will explore them further later.

Q3: Simple time series regression

3. Run a simple time series regression, with one X and no trend. Interpret it.

```
# simplest regression
lm.exciting_life <- lm(exciting_life_pct ~ ba_pct, data = by.year.ts)
summary(lm.exciting_life)
```

```
##
## Call:
## lm(formula = exciting_life_pct ~ ba_pct, data = by.year.ts)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.1695 -0.7673 -0.0436  0.8981  3.8187
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 38.36390    0.89022   43.09  < 2e-16 ***
## ba_pct       0.40736    0.03859   10.56 1.62e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 1.436 on 43 degrees of freedom
## Multiple R-squared:  0.7216, Adjusted R-squared:  0.7151
## F-statistic: 111.5 on 1 and 43 DF,  p-value: 1.621e-13
```

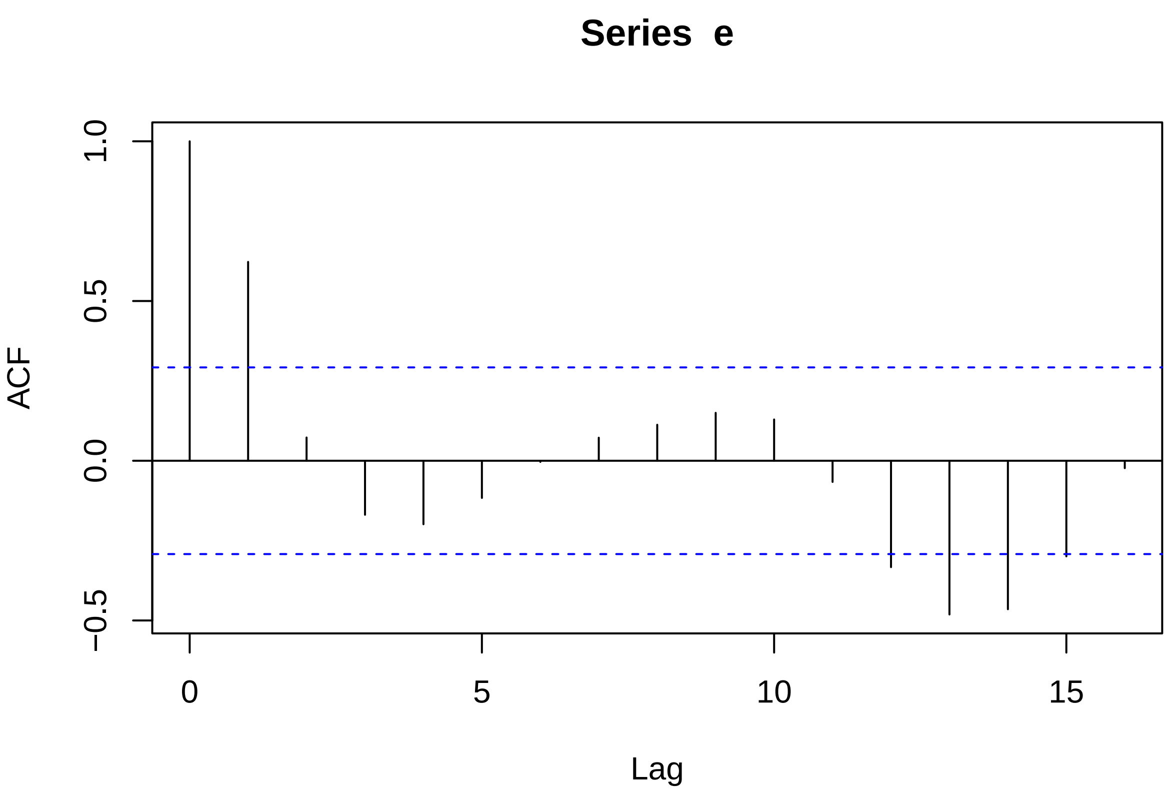
The positive, highly significant coefficient, and high R squared of 0.72, suggests a strong association between percentage of people with higher degrees and people finding their life exciting over time. This indicates that years with a higher proportion of educated individuals correspond to higher life excitement.

```
# test for heteroskedasticity
bptest(lm.exciting_life)
```

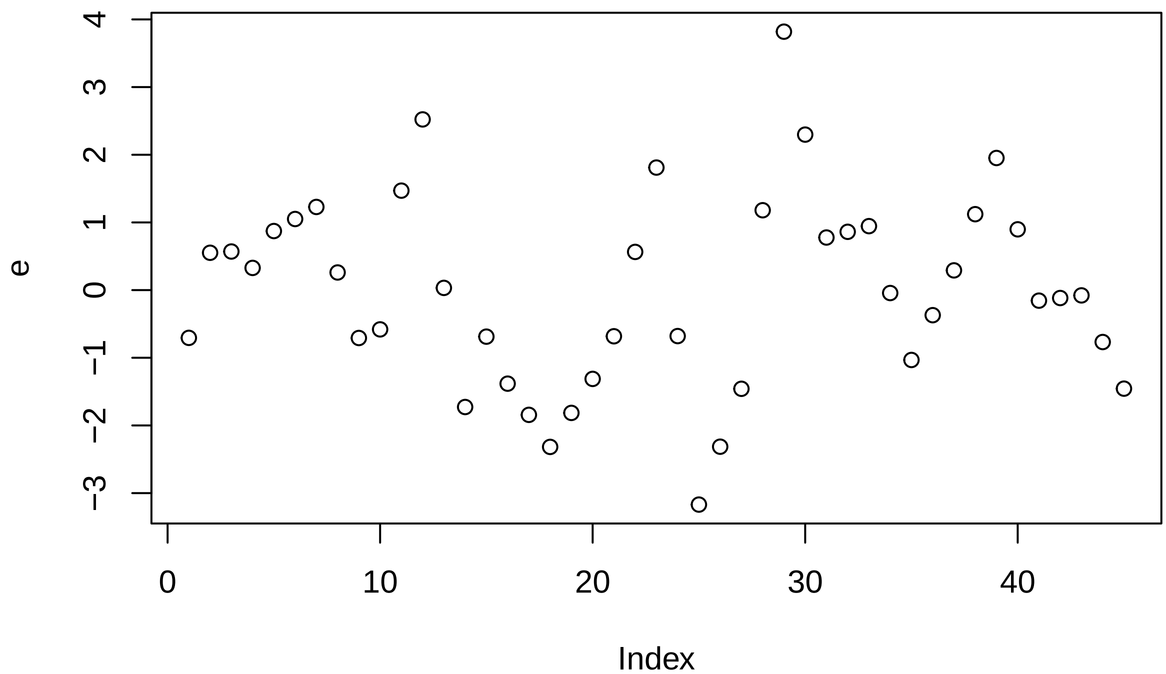
```
##
##  studentized Breusch-Pagan test
##
## data:  lm.exciting_life
## BP = 0.099067, df = 1, p-value = 0.753
```

The BP is not significant, so there is no heteroskedasticity.

```
# look for autocorrelation in errors
e <- lm.exciting_life$resid
acf(e)
```




```
plot(e) # plot residuals over time
```




```r
dwtest(lm.exciting_life) # Durbin-Watson test
```

```
##
##  Durbin-Watson test
##
## data:  lm.exciting_life
## DW = 0.72554, p-value = 2.17e-07
## alternative hypothesis: true autocorrelation is greater than 0
```

```r
bgtest(lm.exciting_life) # Breusch-Godfrey test
```

```
##
##  Breusch-Godfrey test for serial correlation of order up to 1
##
## data:  lm.exciting_life
## LM test = 17.913, df = 1, p-value = 2.312e-05
```

```r
durbinWatsonTest(lm.exciting_life, max.lag=3) # Durbin-Watson with more lags
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1      0.62246085      0.725537   0.000
##    2      0.07278515      1.814817   0.492
##    3     -0.16910079      2.294853   0.256
##  Alternative hypothesis: rho[lag] != 0
```

The tests strongly indicate positive serial correlation:

The DW statistic of ~0.7255 with a very small p-value (~2.17e-07) is much less than 2, indicating strong positive autocorrelation in the residuals. The Breusch-Godfrey test also confirms serial correlation (p-value = 2.312e-05). The ACF plot of residuals shows a large spike at lag 1 and 2.

Hence, the errors from the regression model are not independent over time. This violates OLS assumptions so we must explore more methods.

Q4: TS regression with one X, trent, including autocorr diagnostics

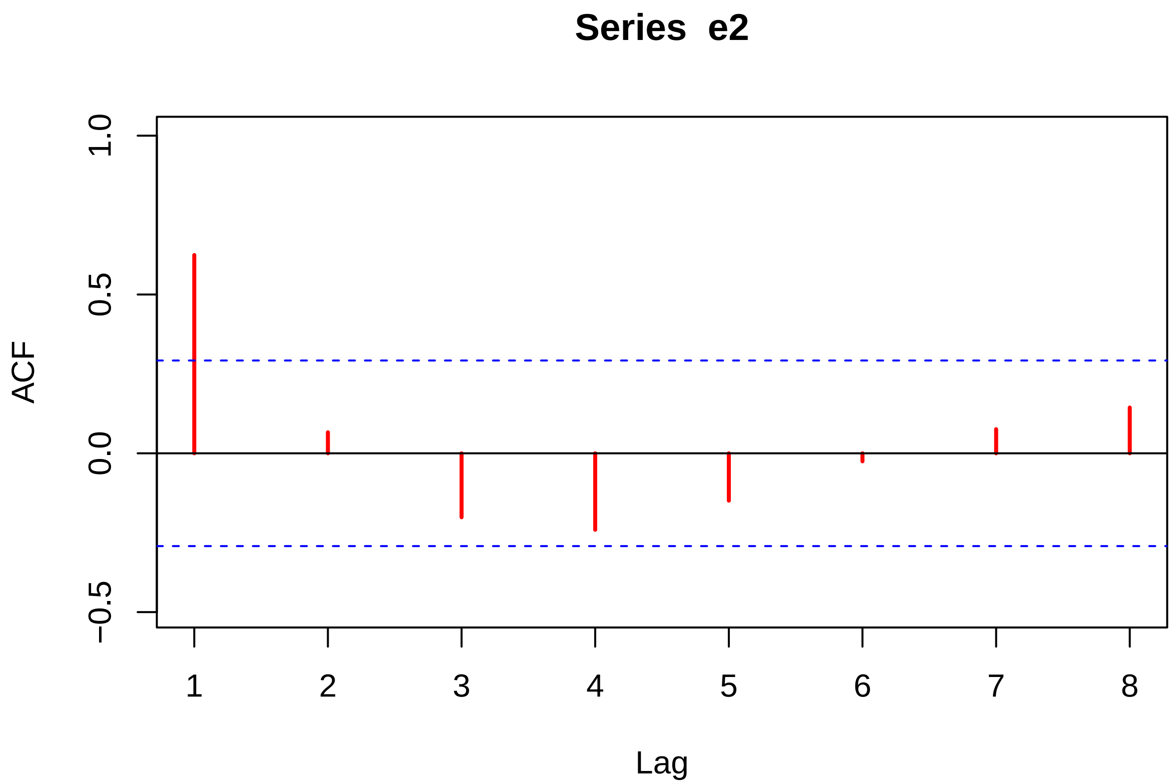
4. Run a time series regression with one X and trend. Interpret it. Perform autocorrelation diagnostics. Explain what you found.

```r
# include year trend
lm.exciting_life2 <- update(lm.exciting_life, ~ . + year)
summary(lm.exciting_life2)
```

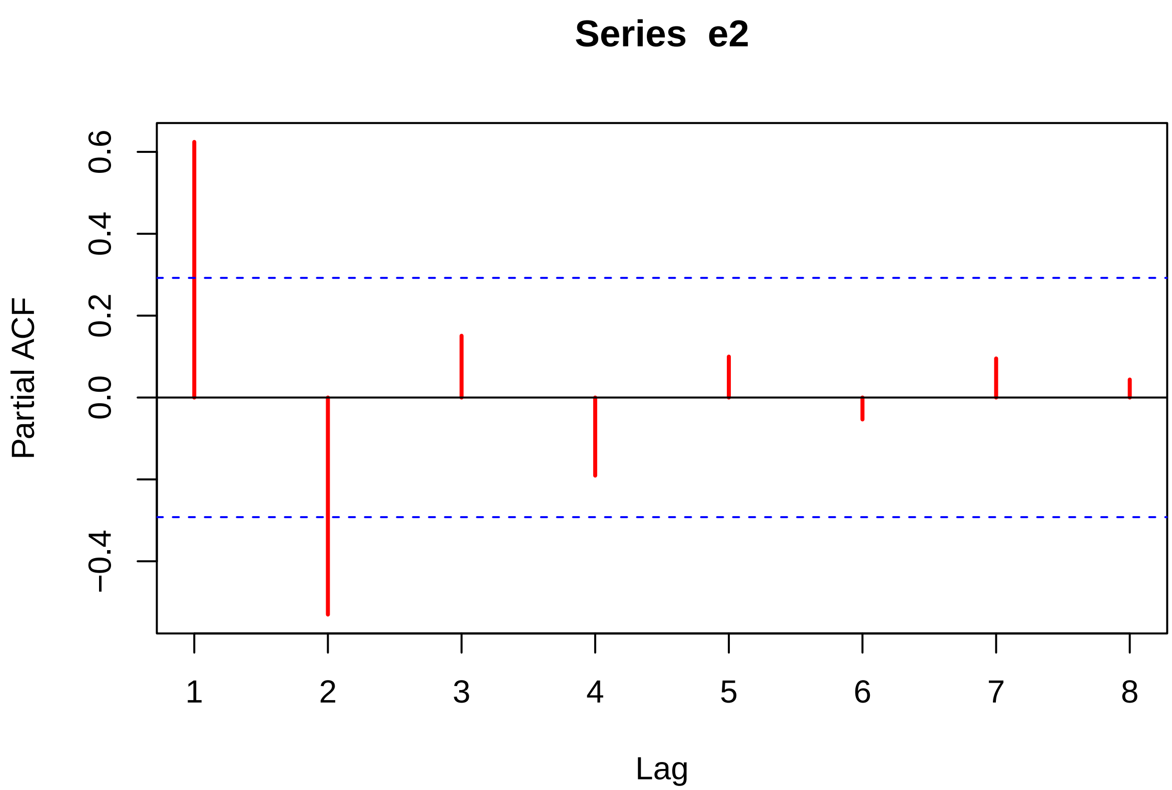
```
##
## Call:
## lm(formula = exciting_life_pct ~ ba_pct + year, data = by.year.ts)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.0967 -0.9286 -0.0960  0.9107  3.7703
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -27.42900  165.56367  -0.166    0.869
## ba_pct        0.32969    0.19932   1.654    0.106
## year          0.03383    0.08514   0.397    0.693
##
## Residual standard error: 1.45 on 42 degrees of freedom
## Multiple R-squared:  0.7227, Adjusted R-squared:  0.7094
## F-statistic: 54.72 on 2 and 42 DF,  p-value: 2.011e-12
```

After controlling for a linear time trend, the relationship between ba_pct and exciting_life_pct is no longer statistically significant. This may indicate that the original relationship is partly spurious due to time trends. In addition, however, the time trend itself is not significant.

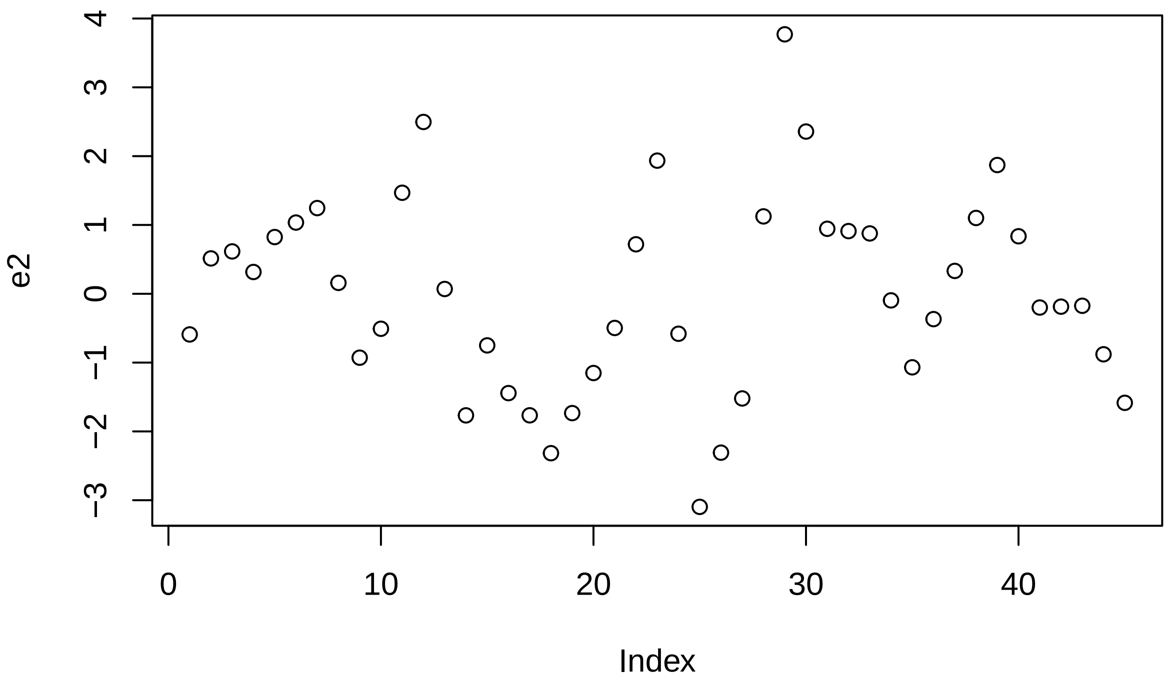
```r
# look for autocorrelation
e2 <- lm.exciting_life2$resid
acf(e2, xlim = c(1,8), col = "red", lwd = 2)
```




```r
pacf(e2, xlim = c(1,8), col = "red", lwd = 2)
```




```r
plot(e2)
```




```r
coeftest(lm.exciting_life2, vcov = NeweyWest(lm.exciting_life2, lag = 1))
```

```
##
## t test of coefficients:
##
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -27.429001 176.810074 -0.1551   0.8775
## ba_pct        0.329686   0.202725  1.6263   0.1114
## year          0.033834   0.090666  0.3732   0.7109
```

```r
dwtest(lm.exciting_life2)
```

```
##
##  Durbin-Watson test
##
## data:  lm.exciting_life2
## DW = 0.71906, p-value = 9.822e-08
## alternative hypothesis: true autocorrelation is greater than 0
```

```r
bgtest(lm.exciting_life2)
```

```
##
##  Breusch-Godfrey test for serial correlation of order up to 1
##
## data:  lm.exciting_life2
## LM test = 18.1, df = 1, p-value = 2.095e-05
```

```r
durbinWatsonTest(lm.exciting_life2, max.lag=3)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1      0.62427829     0.7190619   0.000
##    2      0.06622718     1.8234203   0.526
##    3     -0.20146932     2.3541809   0.202
##  Alternative hypothesis: rho[lag] != 0
```

The ACF plot shows strong positive autocorrelation at lag 1, which means the residuals are not independent over time (the residual at time 1 is correlated with the residual at t - 1). The PACF plot supports this. The residuals look random, but the other tests point to correlation.

The DW statistic for the new model is about 0.719, with a very small p-value. This is almost the same outcome as the original model without the trend and still indicates strong positive serial correlation in the residuals. The BG test small p value also indicates serial correlation in the residuals.

Q5: TS regression with many Xs and trend, including VIF

5. Consider running a time series regression with many Xs and trend. Interpret that. Check VIF.

```r
# add some more predictors
lm.exciting_life3 <- update(lm.exciting_life2, ~ . + age + happy_pct + sat_w_finances_pct)
summary(lm.exciting_life3)
```

```
##
## Call:
## lm(formula = exciting_life_pct ~ ba_pct + year + age + happy_pct +
##     sat_w_finances_pct, data = by.year.ts)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2226 -0.8213  0.0290  0.8175  3.5080
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -62.12276  222.35492  -0.279    0.781
## ba_pct               0.29485    0.22139   1.332    0.191
## year                 0.05586    0.11936   0.468    0.642
## age                 -0.16202    0.39865  -0.406    0.687
## happy_pct           -0.12781    0.17530  -0.729    0.470
## sat_w_finances_pct   0.10304    0.13392   0.769    0.446
##
## Residual standard error: 1.486 on 39 degrees of freedom
## Multiple R-squared:  0.7296, Adjusted R-squared:  0.695
## F-statistic: 21.05 on 5 and 39 DF,  p-value: 3.813e-10
```

```r
vif(lm.exciting_life3) # variance inflation factor
```

```
##             ba_pct               year                age          happy_pct
##          30.744641          48.975997           7.482813           2.938035
## sat_w_finances_pct
##           1.641323
```

```r
durbinWatsonTest(lm.exciting_life3, max.lag=2)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1      0.60992424      0.743718   0.000
##    2      0.06234572      1.826034   0.416
##  Alternative hypothesis: rho[lag] != 0
```

No predictors are significant in the multivariate regression with trend. The R-squared remains around 0.73, similar to prior models, suggesting that adding these variables does not substantially improve the model's explanatory power. Similarly with the RSE, it is similar, so predictive accuracy has not improved.

ba_pct and year have very high VIFs (» 10), suggesting high multicolinearity. The VIF for age is also high, indicating multicolinearity.

Let's dry first differences as these models show they have issues.

Q6: First differenced TS regression

6. Run a first differenced time series regression. Interpret that.

```r
firstD <- function(var, group, df){
  bad <- (missing(group) & !missing(df))
  if (bad) stop("if df is specified then group must also be specified")

  fD <- function(j){ c(NA, diff(j)) }

  var.is.alone <- missing(group) & missing(df)

  if (var.is.alone) {
    return(fD(var))
  }
  if (missing(df)){
    V <- var
    G <- group
  }
  else{
    V <- df[, deparse(substitute(var))]
    G <- df[, deparse(substitute(group))]
  }

  G <- list(G)
  D.var <- by(V, G, fD)
  unlist(D.var)
}

firstD <- function(var, group, df){
  bad <- (missing(group) & !missing(df))
  if (bad) stop("if df is specified then group must also be specified")

  fD <- function(j){ c(NA, diff(j)) }

  var.is.alone <- missing(group) & missing(df)

  if (var.is.alone) {
    return(fD(var))
  }
  if (missing(df)){
    V <- var
    G <- group
  }
  else{
    V <- df[, deparse(substitute(var))]
    G <- df[, deparse(substitute(group))]
  }

  G <- list(G)
  D.var <- by(V, G, fD)
  unlist(D.var)
```

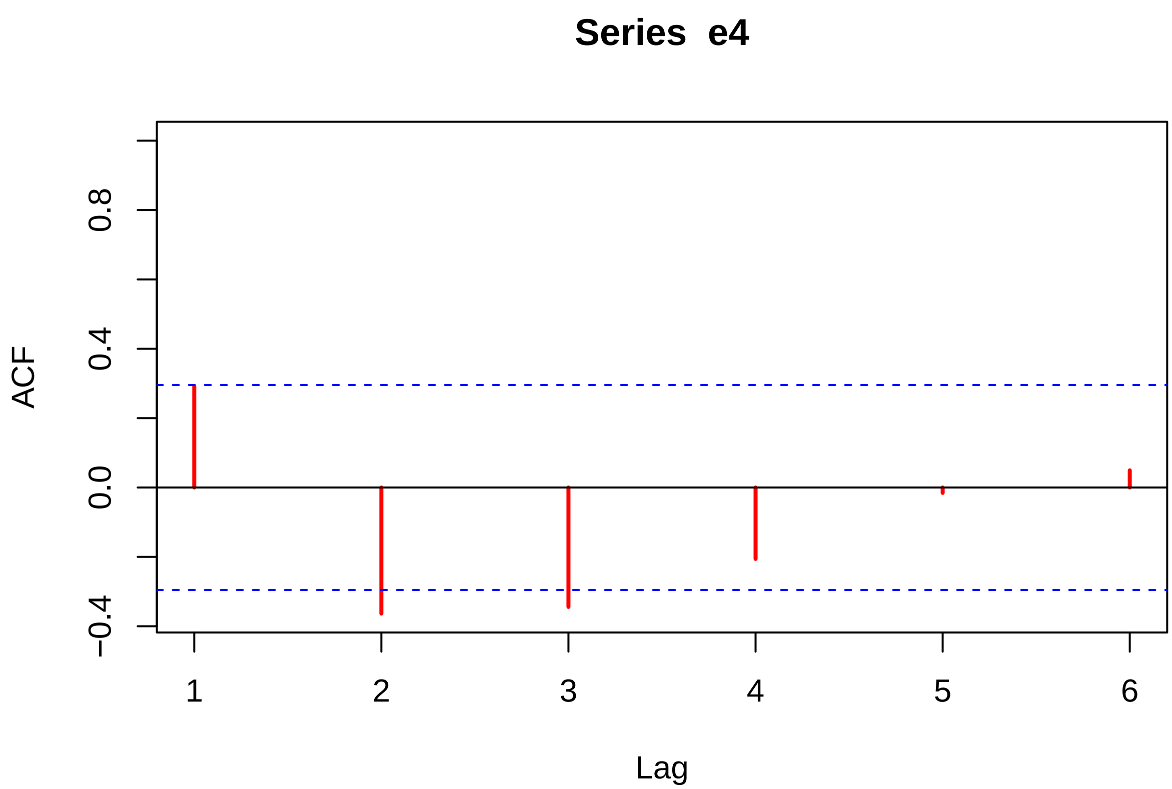
```r
}

## Use the first differences
by.yearFD <- summarise(data.frame(by.year.ts),
                       exciting_life_pct = firstD(exciting_life_pct), # using firstD functon from QMSS p
                       age = firstD(age),
                       ba_pct = firstD(ba_pct),
                       happy_pct = firstD(happy_pct),
                       income = firstD(income),
                       year = year)
```

```
## Warning: Returning more (or less) than 1 row per `summarise()` group was deprecated in
## dplyr 1.1.0.
## i Please use `reframe()` instead.
## i When switching from `summarise()` to `reframe()`, remember that `reframe()`
##   always returns an ungrouped data frame and adjust accordingly.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

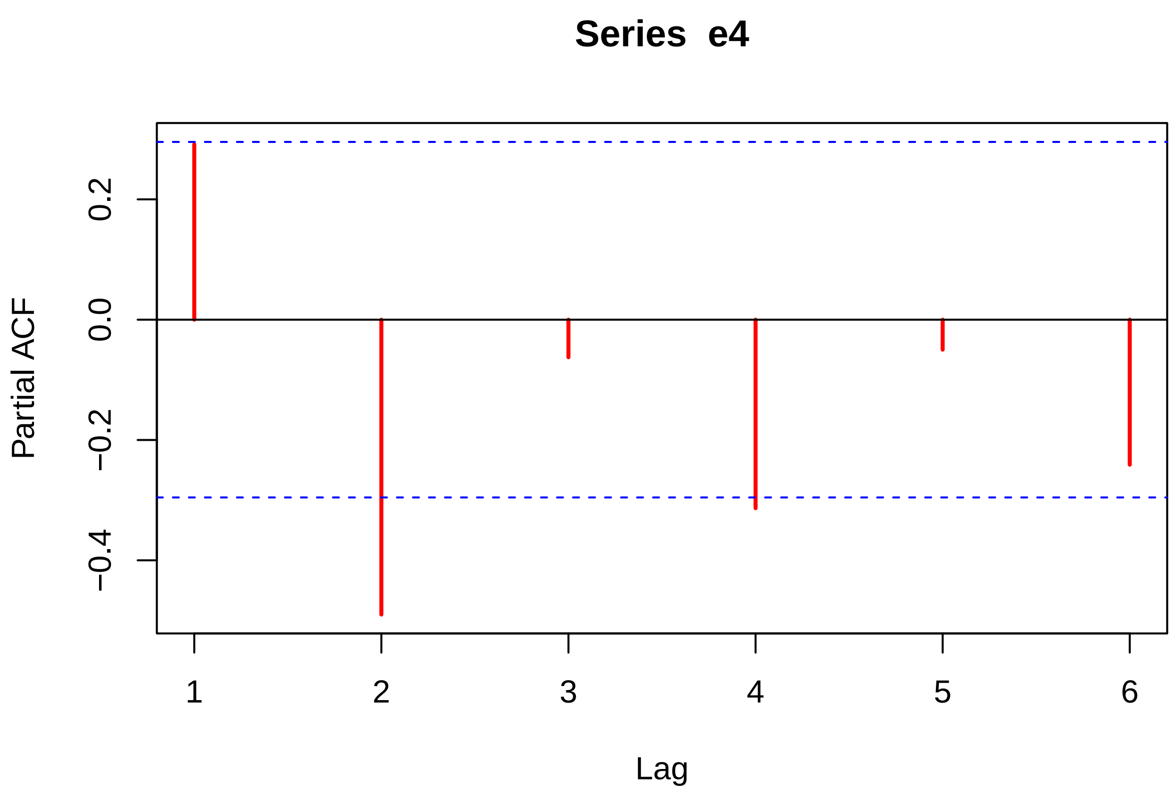
```r
lm.exciting_life4 <- update(lm.exciting_life2, data = by.yearFD)
summary(lm.exciting_life4)
```

```
##
## Call:
## lm(formula = exciting_life_pct ~ ba_pct + year, data = by.yearFD)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.5091 -0.8515  0.1463  0.8360  2.7187
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13.89344   29.35720   0.473    0.639
## ba_pct       0.25440    0.18138   1.403    0.168
## year        -0.00694    0.01470  -0.472    0.639
##
## Residual standard error: 1.238 on 41 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.0509, Adjusted R-squared:  0.004605
## F-statistic: 1.099 on 2 and 41 DF,  p-value: 0.3427
```

```r
e4 <- lm.exciting_life4$resid
acf(e4, xlim = c(1,6), col = "red", lwd = 2)
```




```r
pacf(e4, xlim = c(1,6), col = "red", lwd = 2)
```




```r
# library(forecast)
 auto.arima(e4, trace=TRUE)
```

```
##
##  ARIMA(2,0,2) with non-zero mean : Inf
##  ARIMA(0,0,0) with non-zero mean : 144.8753
##  ARIMA(1,0,0) with non-zero mean : 143.298
##  ARIMA(0,0,1) with non-zero mean : 135.6212
##  ARIMA(0,0,0) with zero mean     : 142.6778
##  ARIMA(1,0,1) with non-zero mean : 137.3828
##  ARIMA(0,0,2) with non-zero mean : 135.2193
##  ARIMA(1,0,2) with non-zero mean : Inf
##  ARIMA(0,0,3) with non-zero mean : Inf
##  ARIMA(1,0,3) with non-zero mean : Inf
##  ARIMA(0,0,2) with zero mean     : 132.7978
##  ARIMA(0,0,1) with zero mean     : 133.3141
##  ARIMA(1,0,2) with zero mean     : Inf
##  ARIMA(0,0,3) with zero mean     : 128.6516
##  ARIMA(1,0,3) with zero mean     : Inf
##  ARIMA(0,0,4) with zero mean     : 130.4606
##  ARIMA(1,0,4) with zero mean     : 133.0495
##
##  Best model: ARIMA(0,0,3) with zero mean
```

```
## Series: e4
## ARIMA(0,0,3) with zero mean
##
## Coefficients:
##          ma1      ma2      ma3
##       0.2957  -0.6477  -0.3556
## s.e.  0.1459   0.1506   0.1385
##
## sigma^2 = 0.9224:  log likelihood = -59.81
## AIC=127.63   AICc=128.65   BIC=134.76
```

The relationship between the percent of people with bachelors or higher and life excitement does not appear to be significant. After differencing, it appears that changes in exciting_life_pct are not strongly or significantly related to changes in ba_pct or to the passage of time (year), so the model does not provide evidence of a meaningful relationship in the differenced series. The dramatic drop in R-squared from 0.72 to 0.05 in the first-differenced model suggests that much of the relationship between education and life excitement was driven by common trends rather than an actual relationship. This either means there is no relationship or that we need to try another model.

Q7: Check for unit roots

7. Check your variables for unit roots. Do some tests. Interpret them.

```r
## 7. Check your variables for unit roots.  Do some tests.  Interpret them.

adfTest(by.year.ts[,"exciting_life_pct"], lags = 0, type="ct")
```

```
##
## Title:
##  Augmented Dickey-Fuller Test
##
## Test Results:
##   PARAMETER:
##     Lag Order: 0
##   STATISTIC:
```

```
##     Dickey-Fuller: -2.7971
##   P VALUE:
##     0.2576
##
## Description:
##  Sat Dec 14 18:04:03 2024 by user:
```

```r
adfTest(by.year.ts[,"exciting_life_pct"], lags = 1, type="ct")
```

```
## Warning in adfTest(by.year.ts[, "exciting_life_pct"], lags = 1, type = "ct"):
## p-value smaller than printed p-value

##
## Title:
##  Augmented Dickey-Fuller Test
##
## Test Results:
##   PARAMETER:
##     Lag Order: 1
##   STATISTIC:
##     Dickey-Fuller: -4.9471
##   P VALUE:
##     0.01
##
## Description:
##  Sat Dec 14 18:04:03 2024 by user:
```

```r
adfTest(by.year.ts[,"exciting_life_pct"], lags = 2, type="ct")
```

```
##
## Title:
##  Augmented Dickey-Fuller Test
##
## Test Results:
##   PARAMETER:
##     Lag Order: 2
##   STATISTIC:
##     Dickey-Fuller: -3.6634
##   P VALUE:
##     0.03883
##
## Description:
##  Sat Dec 14 18:04:03 2024 by user:
```

```r
adfTest(by.year.ts[,"exciting_life_pct"], lags = 3, type="ct")
```

```
##
## Title:
##  Augmented Dickey-Fuller Test
##
## Test Results:
##   PARAMETER:
##     Lag Order: 3
##   STATISTIC:
##     Dickey-Fuller: -3.354
##   P VALUE:
##     0.07578
```

```
##
## Description:
##  Sat Dec 14 18:04:03 2024 by user:
```

```r
adfTest(by.year.ts[,"exciting_life_pct"], lags = 4, type="ct")
```

```
##
## Title:
##  Augmented Dickey-Fuller Test
##
## Test Results:
##   PARAMETER:
##     Lag Order: 4
##   STATISTIC:
##     Dickey-Fuller: -2.8407
##   P VALUE:
##     0.2403
##
## Description:
##  Sat Dec 14 18:04:03 2024 by user:
```

ADF tests with 1, 2, and 3 lags showed significance, while ADF tests with 0 and 4 lags did not. This means that effects are present for lags of 1, 2, 3. This suggests that the relationship between education and life excitement may be more complex than a simple linear trend.

```r
# Phillips-Perron test
PP.test(by.year.ts[,"exciting_life_pct"],lshort=TRUE)
```

```
##
##  Phillips-Perron Unit Root Test
##
## data:  by.year.ts[, "exciting_life_pct"]
## Dickey-Fuller = -3.0058, Truncation lag parameter = 3, p-value = 0.1748
```

The Phillips-Perron test fails to reject the null hypothesis of a unit root in the life excitement series. This result aligns with some of the ADF test specifications and suggests that the series may be non-stationary. This finding indicates that shocks to life excitement may have permanent effects, rather than reverting to the mean.

```r
# BTW, Solution 1: use Newey & West autocorrelation consistent covariance matrix
# estimator

library(sandwich)
coeftest(lm.exciting_life3, vcov = NeweyWest(lm.exciting_life2, lag = 2))
```

```
##
## t test of coefficients:
##
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -62.122762 155.856501 -0.3986   0.6924
## ba_pct        0.294855   0.177906  1.6574   0.1055
## year          0.055858   0.079874  0.6993   0.4885
```

Q8: Automatic ARIMA on residuals

8. Perform an Automatic ARIMA on the residuals from one of your earlier models. Tell me what it says.

```r
library(forecast)
auto.arima(e2, trace=TRUE)
```

```
##
##  ARIMA(2,0,2) with non-zero mean : 127.8571
##  ARIMA(0,0,0) with non-zero mean : 162.338
##  ARIMA(1,0,0) with non-zero mean : 141.9318
##  ARIMA(0,0,1) with non-zero mean : 129.4635
##  ARIMA(0,0,0) with zero mean     : 160.1453
##  ARIMA(1,0,2) with non-zero mean : 126.3288
##  ARIMA(0,0,2) with non-zero mean : 123.7933
##  ARIMA(0,0,3) with non-zero mean : 126.3286
##  ARIMA(1,0,1) with non-zero mean : 125.6544
##  ARIMA(1,0,3) with non-zero mean : 129.0007
##  ARIMA(0,0,2) with zero mean     : 121.4035
##  ARIMA(0,0,1) with zero mean     : 127.1733
##  ARIMA(1,0,2) with zero mean     : 123.8156
##  ARIMA(0,0,3) with zero mean     : 123.8155
##  ARIMA(1,0,1) with zero mean     : 123.2696
##  ARIMA(1,0,3) with zero mean     : 126.3539
##
##  Best model: ARIMA(0,0,2) with zero mean
```

```
## Series: e2
## ARIMA(0,0,2) with zero mean
##
## Coefficients:
##          ma1     ma2
##       1.2262  0.4262
## s.e.  0.1401  0.1282
##
## sigma^2 = 0.756:  log likelihood = -57.41
## AIC=120.82   AICc=121.4   BIC=126.24
```

Applying auto ARIMA to the residuals identified ARIMA(0,0,2) with zero mean as the optimal specification by mimimizing AIC. The model's two MA coefficients are statistically significant, which suggests that changes to to life excitement are still present for approximately 2 periods, but without any autoregressive components. The model's sigma^2 and log likelihood suggest reasonable fit. This is somewhat consistent with the ADF tests, though they suggested lags 1, 2, 3, and this just suggests lags 1, 2.

Q9: ARIMA

9. Run an ARIMA that follows from Step 8. Interpret that, too.

```r
## 9. Run an ARIMA that follows from Step 7.  Interpret that, too.

xvars.fat <- by.year.ts[,c("ba_pct", "year")]

arima.001 <- arima(by.year.ts[,"exciting_life_pct"], order = c(0,0,2), xreg = xvars.fat)
summary(arima.001)
```

```
##
## Call:
## arima(x = by.year.ts[, "exciting_life_pct"], order = c(0, 0, 2), xreg = xvars.fat)
##
## Coefficients:
```

```
##          ma1     ma2  intercept  ba_pct     year
##       1.3544  0.5292    93.2761  0.4810  -0.0284
## s.e.  0.1916  0.1724   102.7286  0.1139   0.0526
##
## sigma^2 estimated as 0.6921:  log likelihood = -56.67,  aic = 125.34
##
## Training set error measures:
##                      ME      RMSE       MAE         MPE     MAPE      MASE
## Training set 0.01274246 0.8319259 0.6257964 0.001845549 1.321035 0.6085466
##                     ACF1
## Training set -0.05052605
```

```r
Box.test(resid(arima.001), lag = 20, type = c("Ljung-Box"), fitdf = 0)
```

```
##
##  Box-Ljung test
##
## data:  resid(arima.001)
## X-squared = 13.978, df = 20, p-value = 0.8316
```

The coefficients for MA1, MA2, and ba_pct are positive with a relatively small standard error suggesting that fluctuations in exciting_life_pct are partly explained by changes in BA percentage (coefficient ba_pct) and are influenced by recent shocks at lags 1 and 2 (as captured by the MA terms). The year variable doesn't show a trend after accounting for education and the moving average components, as the coefficient is small and the standard error is very large, even larger than the magnitude of the coefficient. The training set errors suggest a decent fit.

The BL test suggests the residuals are essentially white noise (no detectable autocorrelation). This suggests that the MA(2) structure and the ba_pct have adequately captured the time-dependent patterns in the data, so this is a good model.

Overall, this analysis shows a possible relationship between the percent of the population with a bachelors or higher and life excitement. It is not shown clearly in all models, but in the ARIMA, it shows that increases in the educated share of the population correlate positively with people's reporting of having an exciting life, while time does not add much explanatory power. The model is statistically sound with no leftover autocorrelation, indicating an appropriate model specification for this particular data set. That being said, the first difference model and other forms showed no relationship, so care should be taken when interpreting this analysis.

Several factors could explain the patterns that are not a direct education-excitement relationship, such as omitted economic and social variables affecting both education and life satisfaction, generational differences in baseline education and life satisfaction, changes in how people interpret questions about life excitement over time, the possibility that excited/optimistic people pursue more education rather than vice versa, and there may have been changes in who has access to higher education over the study period.

All of the above implies that there may be a relationship, as some models suggest, but there may not. If there is a relationship, it is not overwhelmingly strong.