Lab 3

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

```
vars <- c("year", "conpress", "sex", "age", "degree", "wrkstat")
GSS <- data.table::fread("/Users/hengyuai/Desktop/TS/Lab3/trends-gss.csv",
  sep = ",",
  select = vars,
  data.table = FALSE)


sub <- GSS[, vars]
sub = as.data.frame(sub)

sub <- mutate(sub,
              trustpress = ifelse(conpress < 3, 1, 0),
              baplus = ifelse(degree >= 3, 1, 0),
              degreelt50 = ifelse(baplus == 1 & age < 50, 1, 0),
              fulltime = ifelse(wrkstat == 1, 1, 0))
```

```
## My QUESTION is: Are people's confidence in the press and their working status related over time in t

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

# interpolate for some missing years
# add the extra years
by.year[30:40, "year"] <- c(1979, 1981, 1992, 1995, seq(1997, 2009, 2))
by.year <- arrange(by.year, year)

# 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 percent tvholic and percent under 50 with BA
by.year.ts <- as.data.frame(by.year.ts)
by.year.ts <- mutate(by.year.ts,
                     fulltime_pct = fulltime*100,
                     degreelt50_pct = degreelt50*100)

# only keep up to 1992 and convert back to time series object
by.year.ts <- ts(subset(by.year.ts, year <= 1992))

# correlations
cor.vars <- c("trustpress", "fulltime_pct", "degreelt50_pct", "age", "year")
cor.dat <- by.year.ts[, cor.vars]
cor(cor.dat, use = "complete")
```

```
##                trustpress fulltime_pct degreelt50_pct        age
## trustpress      1.0000000   -0.4804851     -0.9281833 -0.6562911
## fulltime_pct   -0.4804851    1.0000000      0.5796194  0.3590326
```

```
## degreelt50_pct -0.9281833    0.5796194      1.0000000  0.6255061
## age            -0.6562911    0.3590326      0.6255061  1.0000000
## year           -0.9250063    0.6809984      0.9177727  0.7119450
##                      year
## trustpress     -0.9250063
## fulltime_pct    0.6809984
## degreelt50_pct  0.9177727
## age             0.7119450
## year            1.0000000
```

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

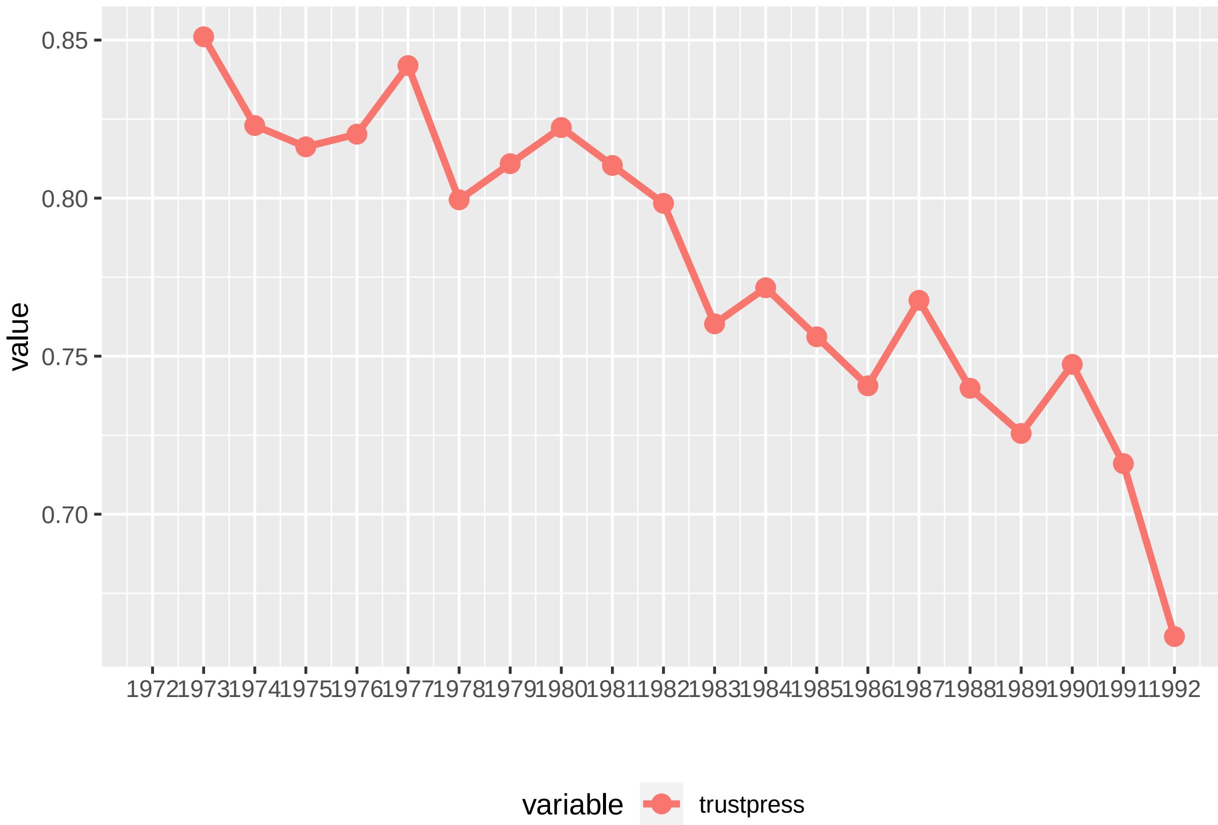
```r
# Time series plots with ggplot
# install.packages("reshape2")

# Make a character vector naming the variables we might want to plot
keep.vars <- c("year", "trustpress", "fulltime_pct", "degreelt50_pct", "age")

# Use meltMyTS to transform the data to a 3-column dataset containing a column for time, a column for v

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
  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)
}

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

```
##    time      variable     value
## 1  1972    trustpress        NA
## 2  1973    trustpress 0.8510494
## 3  1974    trustpress 0.8229665
## 4  1975    trustpress 0.8162275
## 5  1976    trustpress 0.8202324
```

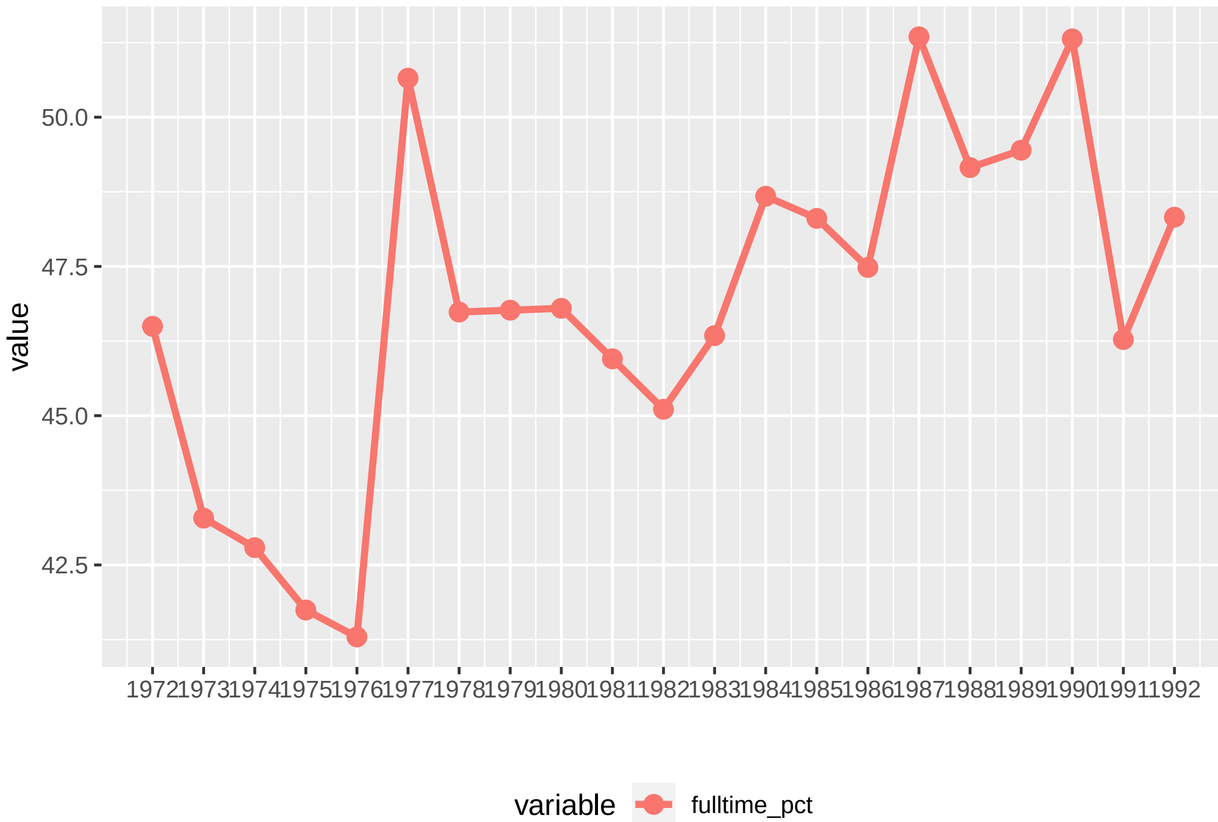
```
## 6  1977     trustpress  0.8419290
## 7  1978     trustpress  0.7994670
## 8  1979     trustpress  0.8108992
## 9  1980     trustpress  0.8223315
## 10 1981     trustpress  0.8103347
## 11 1982     trustpress  0.7983380
## 12 1983     trustpress  0.7602302
## 13 1984     trustpress  0.7716371
## 14 1985     trustpress  0.7561245
## 15 1986     trustpress  0.7406120
## 16 1987     trustpress  0.7676653
## 17 1988     trustpress  0.7398543
## 18 1989     trustpress  0.7255489
## 19 1990     trustpress  0.7474048
## 20 1991     trustpress  0.7160121
## 21 1992     trustpress  0.6613006
## 22 1972   fulltime_pct 46.4972102
## 23 1973   fulltime_pct 43.2845745
## 24 1974   fulltime_pct 42.7897574
## 25 1975   fulltime_pct 41.7449664
## 26 1976   fulltime_pct 41.2941961
## 27 1977   fulltime_pct 50.6535948
## 28 1978   fulltime_pct 46.7362924
## 29 1979   fulltime_pct 46.7673288
## 30 1980   fulltime_pct 46.7983651
## 31 1981   fulltime_pct 45.9529460
## 32 1982   fulltime_pct 45.1075269
## 33 1983   fulltime_pct 46.3414634
## 34 1984   fulltime_pct 48.6761711
## 35 1985   fulltime_pct 48.3050847
## 36 1986   fulltime_pct 47.4829932
## 37 1987   fulltime_pct 51.3468939
## 38 1988   fulltime_pct 49.1559757
## 39 1989   fulltime_pct 49.4469746
## 40 1990   fulltime_pct 51.3119534
## 41 1991   fulltime_pct 46.2755438
## 42 1992   fulltime_pct 48.3245714
## 43 1972 degreelt50_pct  7.6971214
## 44 1973 degreelt50_pct  9.7855228
## 45 1974 degreelt50_pct 10.2564103
## 46 1975 degreelt50_pct  9.8723976
## 47 1976 degreelt50_pct 10.1808439
## 48 1977 degreelt50_pct  9.7769029
## 49 1978 degreelt50_pct 10.1894187
## 50 1979 degreelt50_pct 10.5554602
## 51 1980 degreelt50_pct 10.9215017
## 52 1981 degreelt50_pct 10.2367897
## 53 1982 degreelt50_pct  9.5520777
## 54 1983 degreelt50_pct 12.2180451
## 55 1984 degreelt50_pct 13.1793478
## 56 1985 degreelt50_pct 12.3939987
## 57 1986 degreelt50_pct 13.9740968
## 58 1987 degreelt50_pct 14.0650855
## 59 1988 degreelt50_pct 13.5043889
```

```
## 60 1989 degreelt50_pct 13.9124755
## 61 1990 degreelt50_pct 13.5865595
## 62 1991 degreelt50_pct 15.1655629
## 63 1992 degreelt50_pct 15.7963979
## 64 1972            age 44.9508706
## 65 1973            age 44.1820000
## 66 1974            age 44.5913396
## 67 1975            age 44.3077441
## 68 1976            age 45.2866711
## 69 1977            age 44.6631648
## 70 1978            age 44.0098361
## 71 1979            age 44.4922381
## 72 1980            age 44.9746402
## 73 1981            age 44.9168594
## 74 1982            age 44.8590786
## 75 1983            age 44.2964824
## 76 1984            age 44.0047716
## 77 1985            age 45.7111984
## 78 1986            age 45.4306220
## 79 1987            age 44.9236303
## 80 1988            age 45.3744076
## 81 1989            age 45.4435747
## 82 1990            age 45.9569971
## 83 1991            age 45.6261559
## 84 1992            age 45.8374377
```

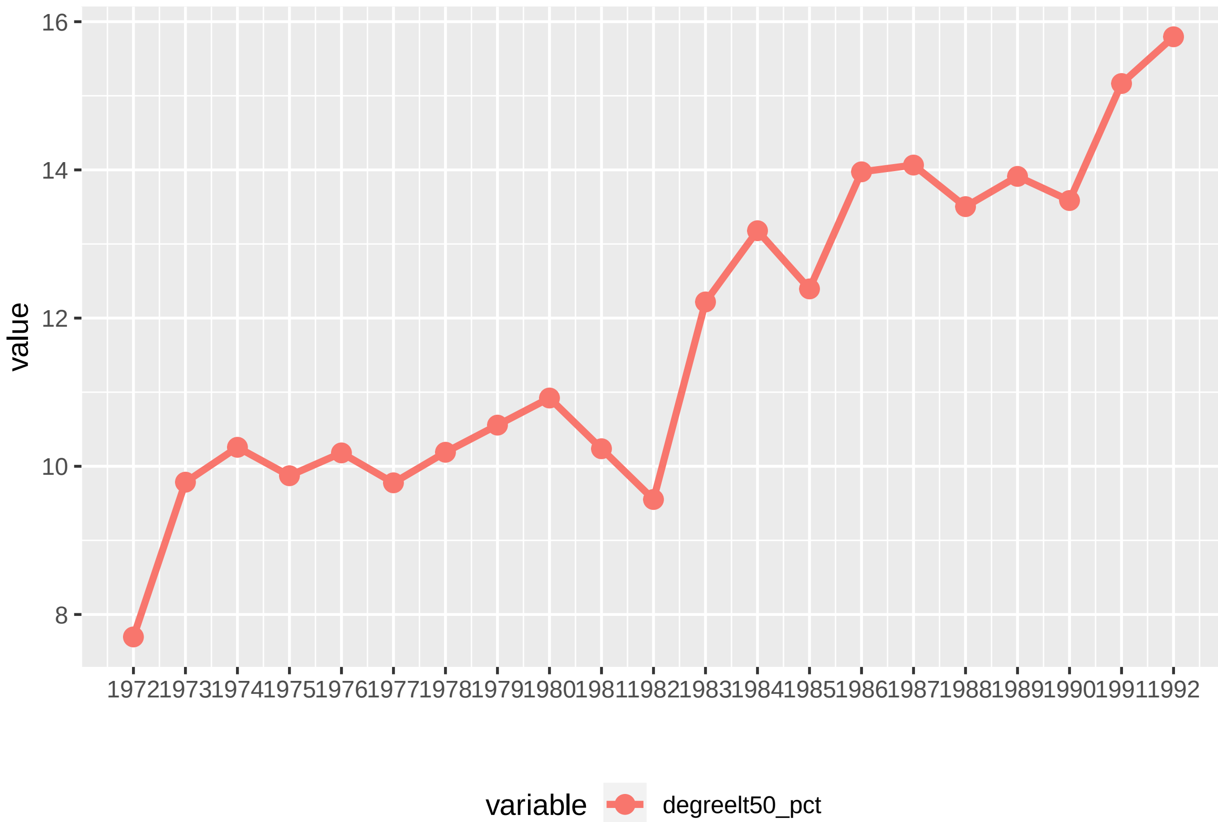
```
ggMyTS <- function(df, varlist, line = TRUE, point = TRUE, pointsize = 3, linewidth = 1.25, ...){
  require(ggplot2)
  # varlist = character vector with names of variables to use
  if(missing(varlist)){
    gg <- ggplot(df, aes(time, value, colour = variable))
  }
  else{
    include <- with(df, 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
}

(g_trustpress <- ggMyTS(df = plot.dat, varlist = c("trustpress")))
```




```
(g_tvholicpct <- ggMyTS(df = plot.dat, varlist = c("fulltime_pct")))
```




```
(g_degreelt50_pct <- ggMyTS(df = plot.dat, varlist = c("degreelt50_pct")))
```



Explain how you think Y should relate to your key Xs: From the graphs above, we can find that the percentage of people who have fulltime jobs and the percentage of people who with at least a BA increased between 1972 and 1992 overall. However, people's average confidence in press declined. Therefore, I think that people's average confidence in press was negatively related to the percentage of people with fulltime jobs and a BA under 50.

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

```
# simplest regression
lm.trust <- lm(trustpress ~ fulltime_pct, data = by.year.ts)
summary(lm.trust)
```

```
##
## Call:
## lm(formula = trustpress ~ fulltime_pct, data = by.year.ts)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.10638 -0.02173  0.00428  0.02386  0.09262
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.149076   0.159507   7.204 1.05e-06 ***
## fulltime_pct -0.007892   0.003395  -2.324    0.032 *
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.04365 on 18 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.2309, Adjusted R-squared:  0.1881
## F-statistic: 5.403 on 1 and 18 DF,  p-value: 0.032
```

The percent people working full-time was negatively related to average confidence in press. The coefficient is statistically significant at 0.05 and we can reject the null of no effect.

```r
# test for heteroskedasticity
bptest(lm.trust)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  lm.trust
## BP = 1.3552, df = 1, p-value = 0.2444
```

There is no heteroskadasticity from the above regression.

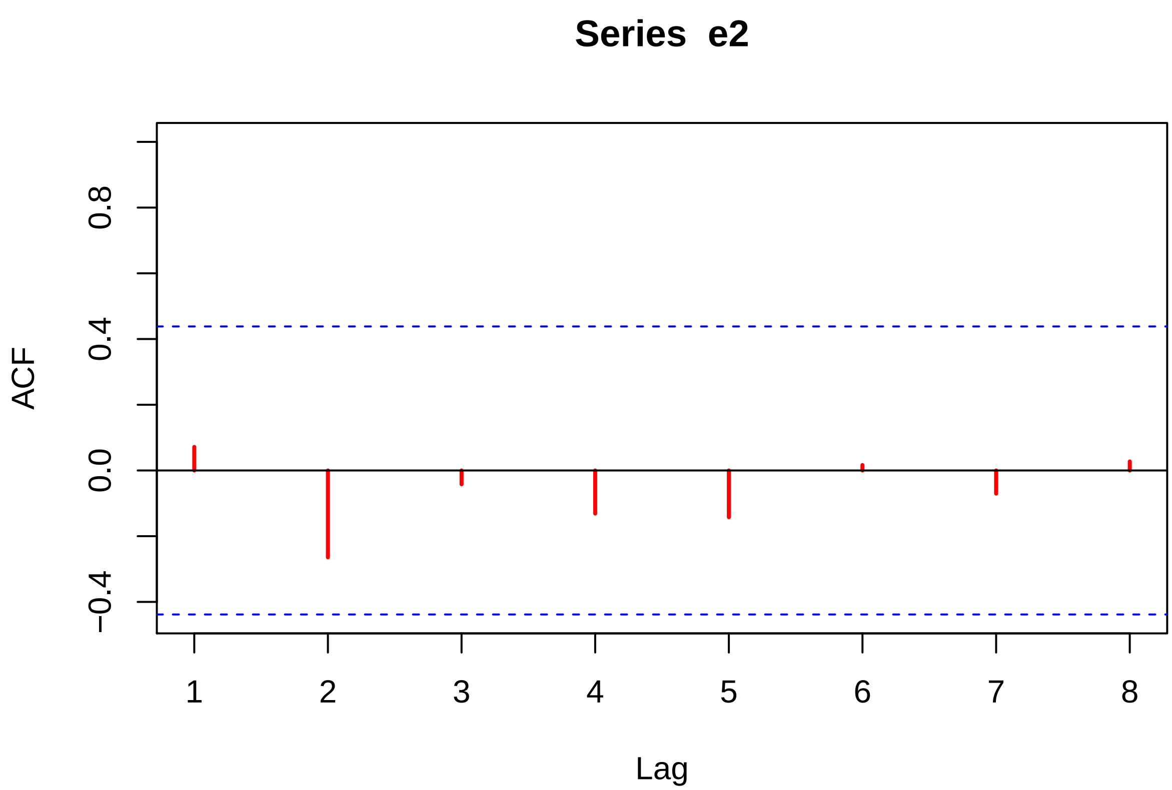
4. Run a time series regression with one X and trend. Interpret it. Perform autocorrelation diagnostics. Explain what you found.

```r
lm.trust2 <- update(lm.trust, ~ . + year)
summary(lm.trust2)
```

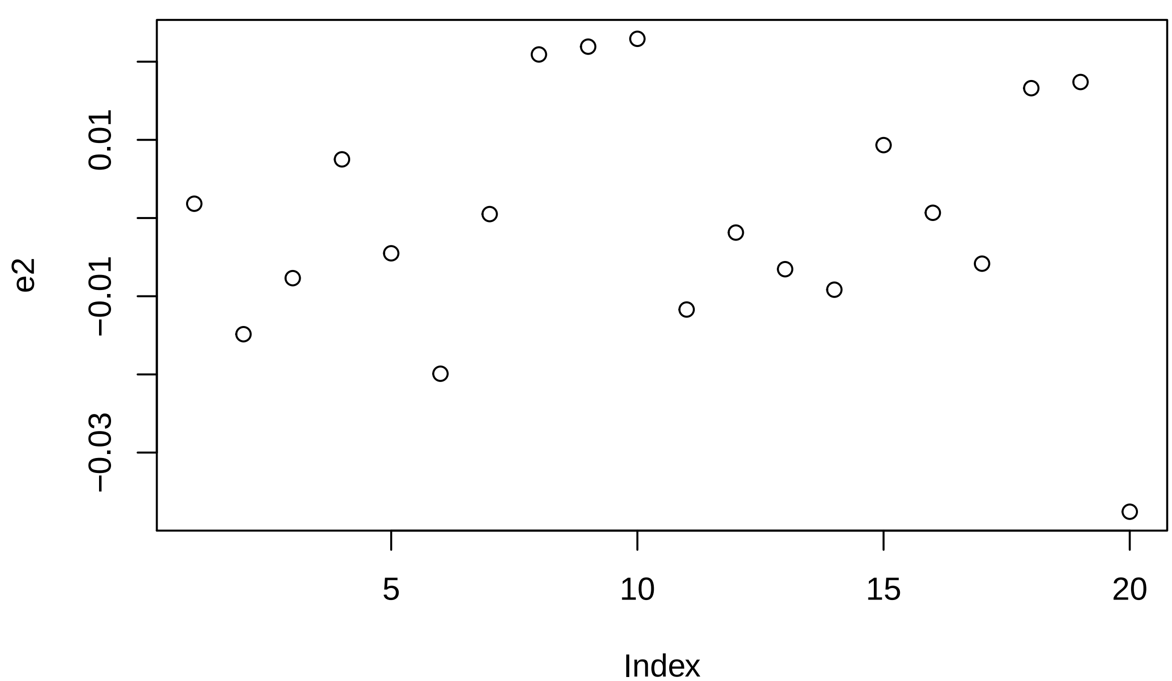
```
##
## Call:
## lm(formula = trustpress ~ fulltime_pct + year, data = by.year.ts)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.037563 -0.008056 -0.000676  0.011149  0.022925
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  18.6597301  1.6684148  11.184 2.93e-09 ***
## fulltime_pct  0.0045776  0.0017436   2.625   0.0177 *
## year         -0.0091275  0.0008691 -10.502 7.51e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01641 on 17 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.8973, Adjusted R-squared:  0.8852
## F-statistic: 74.25 on 2 and 17 DF,  p-value: 3.971e-09
```

Net of the year trend, each percent more of full-time employed people increases ave. confidence in press by 0.0045776. This coefficient is significant at 0.05 level and we can reject the null of no effects.

```r
# look for autocorrelation
e2 <- lm.trust2$resid
acf(e2, xlim = c(1,8), col = "red", lwd = 2)
```




```r
plot(e2)
```



```r
dwtest(lm.trust2)
```

```
##
##  Durbin-Watson test
##
## data:  lm.trust2
## DW = 1.5479, p-value = 0.08736
## alternative hypothesis: true autocorrelation is greater than 0
```

```r
bgtest(lm.trust2)
```

```
##
##  Breusch-Godfrey test for serial correlation of order up to 1
##
## data:  lm.trust2
## LM test = 0.17355, df = 1, p-value = 0.677
```

```r
durbinWatsonTest(lm.trust2, max.lag=3)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1      0.07160435      1.547916   0.192
##    2     -0.26429238      2.105385   0.842
##    3     -0.04201482      1.587673   0.492
##  Alternative hypothesis: rho[lag] != 0
```

From the ACF graph and residual trend graph, we cannot see any AR(1) left. In the dwtest and bgtest result, a prob of chi2 > 0.05 indicates no serial correlation.

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

```r
lm.trust3 <- update(lm.trust2, ~ . + degreelt50_pct)
summary(lm.trust3)
```

```
##
## Call:
## lm(formula = trustpress ~ fulltime_pct + year + degreelt50_pct,
##     data = by.year.ts)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.030039 -0.006378 -0.001098  0.009174  0.020660
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    11.958147   3.053569   3.916  0.00123 **
## fulltime_pct    0.003976   0.001543   2.576  0.02029 *
## year           -0.005671   0.001579  -3.592  0.00244 **
## degreelt50_pct -0.010286   0.004117  -2.498  0.02376 *
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01435 on 16 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.9261, Adjusted R-squared:  0.9123
## F-statistic: 66.84 on 3 and 16 DF,  p-value: 2.869e-09
```

Net of the year trend, each percent more of full-time employed people increases ave. confidence in press by 0.003976. This coefficient is significant at 0.05 level and we can reject the null of no effects. Net of the year trend, each percent more of under 50 BA degree people decreases ave. confidence in press by 0.010286. This coefficient is significant at 0.05 level and we can reject the null of no effects.

```
vif(lm.trust3) # variance inflation factor
```

```
##   fulltime_pct           year degreelt50_pct
##       1.911386       8.048782       6.499728
```

Given such high correlations among variables, we want to look out for multicollinearity, which we might have with year and % of people under 50 with a BA+ degree.

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

```
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)
}

by.yearFD <- summarise(data.frame(by.year.ts),
                       trustpress = firstD(trustpress), # using firstD functon from QMSS package
                       age = firstD(age),
```

```r
                        fulltime_pct = firstD(fulltime_pct),
                        degreelt50_pct = firstD(degreelt50_pct),
                        year = year)

lm.trust4 <- update(lm.trust3, data = by.yearFD)
summary(lm.trust4)
```

```
##
## Call:
## lm(formula = trustpress ~ fulltime_pct + year + degreelt50_pct,
##     data = by.yearFD)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.045549 -0.011698 -0.001073  0.015786  0.024447
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     0.8948219  1.7097579   0.523   0.6084
## fulltime_pct    0.0039378  0.0015902   2.476   0.0257 *
## year           -0.0004557  0.0008623  -0.528   0.6049
## degreelt50_pct -0.0066733  0.0054022  -1.235   0.2357
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.02022 on 15 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:  0.3871, Adjusted R-squared:  0.2646
## F-statistic: 3.159 on 3 and 15 DF,  p-value: 0.05572
```

For each 1 percentage point change in people working-full time, average confidence in press increases by 0.0039378, net of all other differences in the Xs and at any point in time. This coefficient is significant at 0.05 level and we can reject the null of no effects. For each 1 percentage point change in people getting BA, average confidence in press decreases by 0.0066733, net of all other differences in the Xs and at any point in time. This coefficient is not significant.

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

```r
# install.packages("fUnitRoots")
library(fUnitRoots)
```

```
## Loading required package: timeDate

## Loading required package: timeSeries

##
## Attaching package: 'timeSeries'

## The following object is masked from 'package:zoo':
##
##     time<-
```

```
## Loading required package: fBasics

##
## Attaching package: 'fBasics'

## The following object is masked from 'package:car':
##
##     densityPlot
```

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

```
##
## Title:
##  Augmented Dickey-Fuller Test
##
## Test Results:
##   PARAMETER:
##     Lag Order: 0
##   STATISTIC:
##     Dickey-Fuller: -2.842
##   P VALUE:
##     0.2516
##
## Description:
##  Fri Nov 29 01:27:25 2019 by user:
```

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

```
##
## Title:
##  Augmented Dickey-Fuller Test
##
## Test Results:
##   PARAMETER:
##     Lag Order: 4
##   STATISTIC:
##     Dickey-Fuller: -1.6725
##   P VALUE:
##     0.6971
##
## Description:
##  Fri Nov 29 01:27:25 2019 by user:
```

Either with 0 lag or with 4 lags, p-value is too high to be able to reject the null of Unit Root, therefore, we might have a unit root here.

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

```r
library(forecast)
```

```
## Registered S3 method overwritten by 'xts':
##   method     from
##   as.zoo.xts zoo

## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo

## Registered S3 methods overwritten by 'forecast':
##   method             from
##   fitted.fracdiff    fracdiff
##   residuals.fracdiff fracdiff
```

```r
e <- lm.trust$resid
auto.arima(e, trace=TRUE)
```

```
##
##  ARIMA(2,1,2) with drift         : Inf
##  ARIMA(0,1,0) with drift         : -63.006
##  ARIMA(1,1,0) with drift         : -63.08589
##  ARIMA(0,1,1) with drift         : Inf
##  ARIMA(0,1,0)                    : -64.81855
##  ARIMA(1,1,1) with drift         : Inf
##
##  Best model: ARIMA(0,1,0)

## Series: e
## ARIMA(0,1,0)
##
## sigma^2 estimated as 0.001717:  log likelihood=33.53
## AIC=-65.05   AICc=-64.82   BIC=-64.11
```

auto.arima suggests that the errors from the static model is a random walk and we cannot reject unit root.

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

```r
xvars.fat <- by.year.ts[,c("fulltime_pct")]

arima.010 <- arima(by.year.ts[,"trustpress"], order = c(0,1,0), xreg = xvars.fat)
summary(arima.010)
```

```
##
## Call:
## arima(x = by.year.ts[, "trustpress"], order = c(0, 1, 0), xreg = xvars.fat)
##
## Coefficients:
```

```
##       xvars.fat
##          0.0039
## s.e.     0.0017
##
## sigma^2 estimated as 0.0004906:  log likelihood = 45.43,  aic = -86.86
##
## Training set error measures:
##                       ME       RMSE        MAE       MPE     MAPE
## Training set -0.01044549 0.02158966 0.01621065 -1.425352 2.157985
##                   MASE       ACF1
## Training set 0.7558092 -0.1133399
```

Each 1 percentage point difference in the percent of people with full-time job increases people's confidence in press by 0.0039 percentage points.