Name: Your name here
Due: 2024/11/04

Homework 6

Be sure to submit **both** the .pdf and .qmd file to Canvas by Monday, November 4th at 11:59 pm. The purpose of this assignment is to use simulation to better understand why we need to account for serial correlation in a time series.

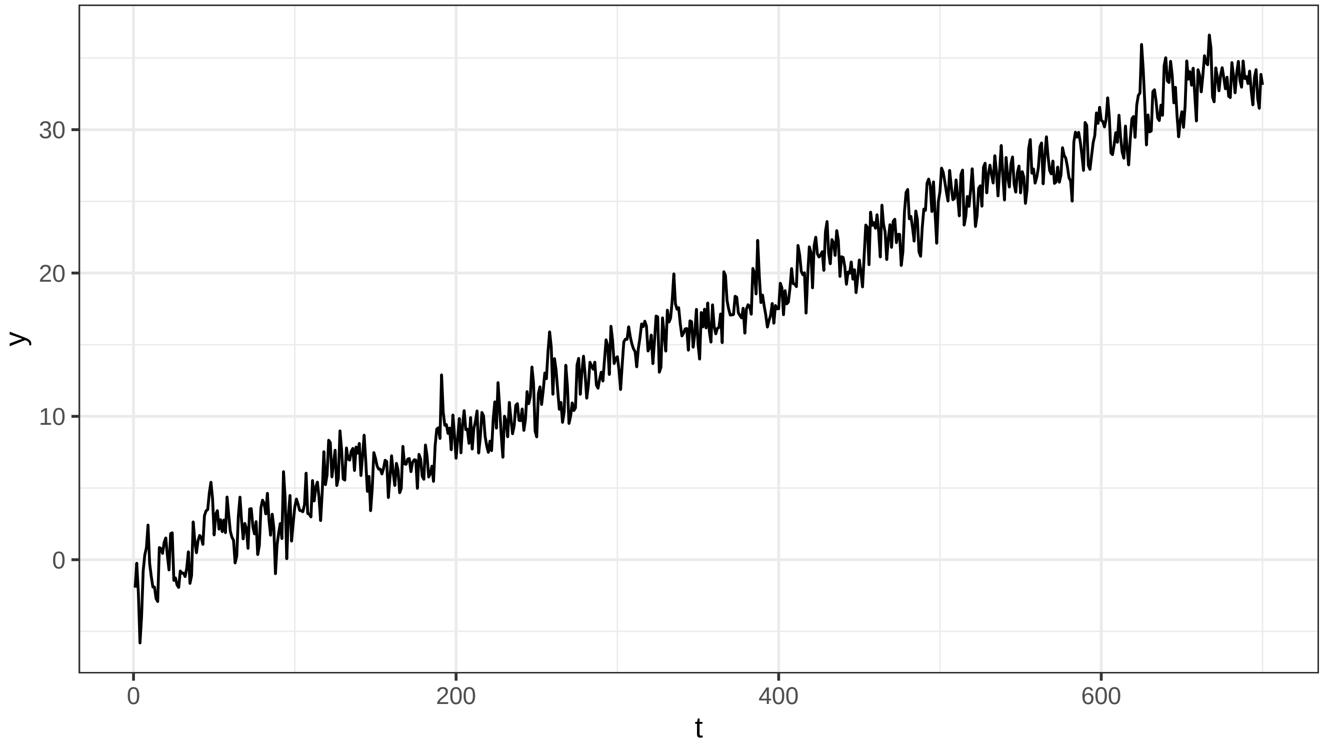
For this assignment, you may use the `generate_ts_reg` function contained in the `helpers.R` script to generate time series with trend, seasonality, and autocorrelated errors from an AR(1) process.

```r
source("helpers.R")

# example: 50 weeks of data
ex_ts <- generate_ts_reg(
  1027204,
  n = 100*7,
  freq = 7,
  betas = c(-1, .05, rnorm(6, sd = .5)) # beta0, beta1, and 6 harmonic cycles
)
str(ex_ts)
```

```
List of 4
 $ df    : tibble [700 x 8] (S3: tbl_df/tbl/data.frame)
  ..$ t    : int [1:700] 1 2 3 4 5 6 7 8 9 10 ...
  ..$ sin1t: num [1:700] 0.782 0.975 0.434 -0.434 -0.975 ...
  ..$ sin2t: num [1:700] 0.975 -0.434 -0.782 0.782 0.434 ...
  ..$ sin3t: num [1:700] 0.434 -0.782 0.975 -0.975 0.782 ...
  ..$ cos1t: num [1:700] 0.623 -0.223 -0.901 -0.901 -0.223 ...
  ..$ cos2t: num [1:700] -0.223 -0.901 0.623 0.623 -0.901 ...
  ..$ cos3t: num [1:700] -0.901 0.623 -0.223 -0.223 0.623 ...
  ..$ y    : num [1:700] -1.951 -0.246 -2.275 -5.809 -3.973 ...
 $ X     : num [1:700, 1:8] 1 1 1 1 1 1 1 1 1 1 ...
  ..- attr(*, "dimnames")=List of 2
  .. ..$ : chr [1:700] "1" "2" "3" "4" ...
  .. ..$ : chr [1:8] "(Intercept)" "t" "sin1t" "sin2t" ...
  ..- attr(*, "assign")= int [1:8] 0 1 2 3 4 5 6 7
 $ mean  : num [1:700] -1.784 0.174 -0.191 -1.363 -1.335 ...
 $ params:List of 3
  ..$ betas: num [1:8] -1 0.05 0.318 -0.847 -0.194 ...
  ..$ sigma: num 1
  ..$ alpha: num 0.662
```

```r
ex_ts$df %>%
  ggplot() +
  geom_line(aes(x = t, y = y)) +
  theme_bw()
```



```r
lm(y ~ ., ex_ts$df)
```

```

Call:
lm(formula = y ~ ., data = ex_ts$df)

Coefficients:
(Intercept)            t        sin1t        sin2t        sin3t        cos1t
   -1.39613      0.05156      0.30171     -0.87662     -0.15775     -0.18963
      cos2t        cos3t
   -0.21228      0.12073
```

```r
ex_ts$params$betas
```

```
[1] -1.0000000  0.0500000  0.3178850 -0.8473730 -0.1943643 -0.1861052 -0.1557460
[8]  0.1005507
```

Question 1 [11 pt] The goal of this assignment is to conduct a *simulation study* that demonstrates why we need to account for serial autocorrelation when fitting regression models. The purpose of this first question is to get our feet wet with the idea of a simulation study.

1. [2 pt] Use the `generate_ts_reg` function to generate a time series that represents 20 years of monthly data. Set the `beta` vector equal to `c(-1, 0.05, 1, -1, rnorm(10, sd = .5))` and the autocorrelation parameter to `.8`. Plot the resulting time series.

2. [2 pt] Fit a linear regression model that includes the time index and all 12 harmonic seasonal cycles. Create a PACF plot of the residuals and comment on how the ACF plot relates to the way you generated the series.

   > The PACF plot suggests that the residuals are an AR(1) series with $\alpha = .8$, which makes sense because that is exactly what we simulated.

3. [2 pt] Create confidence intervals for the regression coefficients using `confint` and determine which intervals captured the generating parameters (which you can find contained within the output of the `generate_ts_reg` function).

4. [4 pt] Fit a GLS model with an AR(1) correlation structure, create confidence intervals for the regression coefficients, and again determine which intervals captured the generating values.

5. [1 pt] If we were to repeatedly simulate time series like in question 1-1 through 1-4 and calculate 95% confidence intervals for the regression coefficients in the model, approximately what percent of the constructed intervals should capture the generating values, if we are appropriately modeling the uncertainty?

Question 2 [10 pt] We are now ready to conduct our simulation study. To do so, repeat the following process 100 times:

- simulate new data with a distinct seed (using the same `betas` and `alpha` as before)
- fit an OLS model and determine for which parameters the confidence interval captures the generating values, store these values
- fit a GLS model and determine for which parameters the confidence interval captures the generating values, store these values

Then, calculate the proportion of times (out of 100 simulations) that each model captured the generating values for each of the regression coefficients. Create a visual that displays the resulting proportions and comment on what the results suggest about the importance of accounting for serial autocorrelation when estimating regression coefficients.

```
# conduct your simulation here and save the results using
# save(sim, file = "sim.rdata")
```

```
# then load it here: load("sim.rdata")
```