Name: Your name here
Due: 2024/11/27

Day 22 - Lab: ARIMA models

Introduction

In this assignment, we will get more practice with time series inference using GLS and ARIMA models. This lab is graded on completion - just make sure you turn something in by 11/27.

For this lab, we will focus on two data sets: one to practice forecasting via `auto.arima`, and the other to practice GLS. The data we will use for forecasting concerns the global mean methane abundance for marine surface sites, courtesy of NOAA[1].

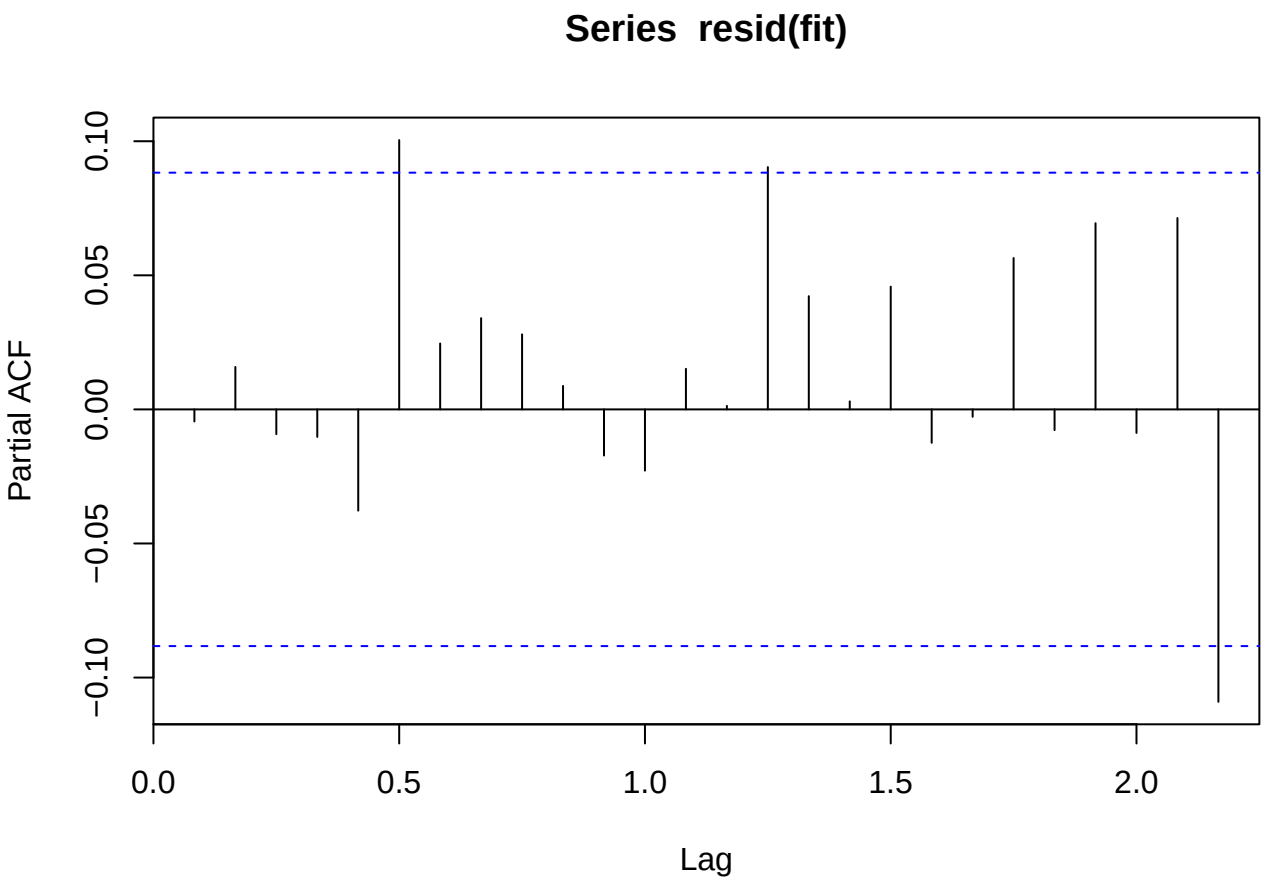
---

[1] Link to the NOAA site. You do not need to download anything from this link.

Question 1

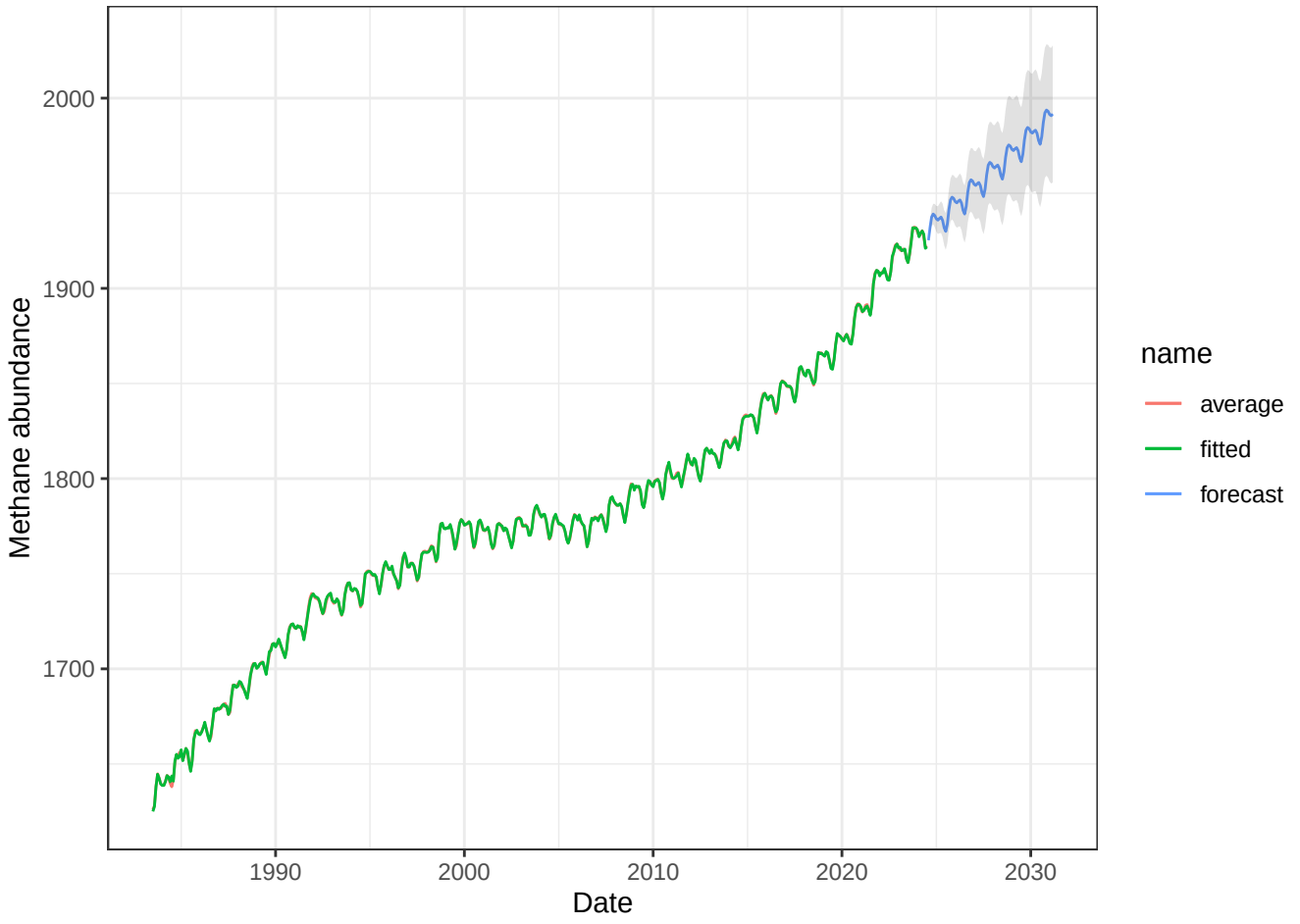
Suppose we are interested in determining the last month during which we are 95% confident that the mean atmospheric methane abundance is at most 2000. Fit a model to answer this question, demonstrate that there is no residual autocorrelation from this fit, and produce a figure that showcases your fitted model.

```
# recreate ts in evaluated chunk
ch4_ts <- ts(ch4$average, start = c(1983, 7), freq = 12)
fit <- readRDS("autofit.rds")

# residual autocorrelation?
pacf(resid(fit)) # not bad
```




```
# forecast
auto_forecast <- forecast(fit, h = 20*4, level = 95)

# plot
```

```
obs_df <- ch4 %>%
  mutate(time = c(time(ch4_ts)), fitted = fitted(fit)) %>%
  dplyr::select(time, average, fitted) %>%
  pivot_longer(-time)
forecast_df <- tibble(
  time = c(time(auto_forecast$mean)),
  value = c(auto_forecast$mean),
  lwr = c(auto_forecast$lower),
  upr = c(auto_forecast$upper),
  name = "forecast"
)

ggplot() +
  geom_line(data = obs_df, aes(x = time, y = value, col = name)) +
  geom_line(data = forecast_df, aes(x = time, y = value, col = name)) +
  geom_ribbon(data = forecast_df, aes(x = time, ymin = lwr, ymax = upr), alpha = .15) +
  theme_bw() +
  labs(x = "Date", y = "Methane abundance")
```



```r
forecast_df %>% filter(upr < 2000) %>% tail()
```

```
# A tibble: 6 x 5
   time value   lwr   upr name
  <dbl> <dbl> <dbl> <dbl> <chr>
1 2029. 1974. 1949. 1999. forecast
2 2029  1973. 1947. 2000. forecast
3 2029. 1972. 1946. 1999. forecast
4 2029. 1969. 1940. 1997. forecast
5 2030. 1967. 1938. 1995. forecast
6 2030. 1971. 1942. 2000. forecast
```

The plot above includes the observed series, fitted series, and forecasted series. Based on this plot, and the print table, the last month that we expect, with 95% confidence, the methane abundance to be no greater than 2000 is July of 2029. Enjoy the next five good years. :) It is also worth noting that the PACF plot suggests little to no evidence of serial autocorrelation in the errors. A few lags extend above the blue bands, but not by much!

Question 2

Let us return to GLS to get a bit more practice answering research questions. To motivate this exercise, we will use a cool data set containing small rodent abundance and rainfall data in Chile. From a rather interesting paper[2]:

> Rodent outbreaks or irruptions in semiarid Chile are associated with rainfall pulses driven by the El Niño Southern Oscillation (ENSO). During the last decade, north-central Chile has experienced an almost uninterrupted sequence of dry years, the so-called megadrought, which had led to a new ecological situation in this region. We employ a diagnostic approach to analyze the abundance of data regarding two rodent species, Phyllotis darwini and Abrothrix olivacea, using a 33-yr time series spanning from 1987 to 2019. Our population dynamic models provide evidence of competitive interactions within and among both species of rodents. This result is novel since rainfall variability influences the degree of interspecific competition and is asymmetric. The diagnostic approach used here offers a way to develop simple population models that are useful for understanding the causes of population fluctuations and for predicting population changes under a climate change scenario.

Here is a bit of information about each column from the metadata:

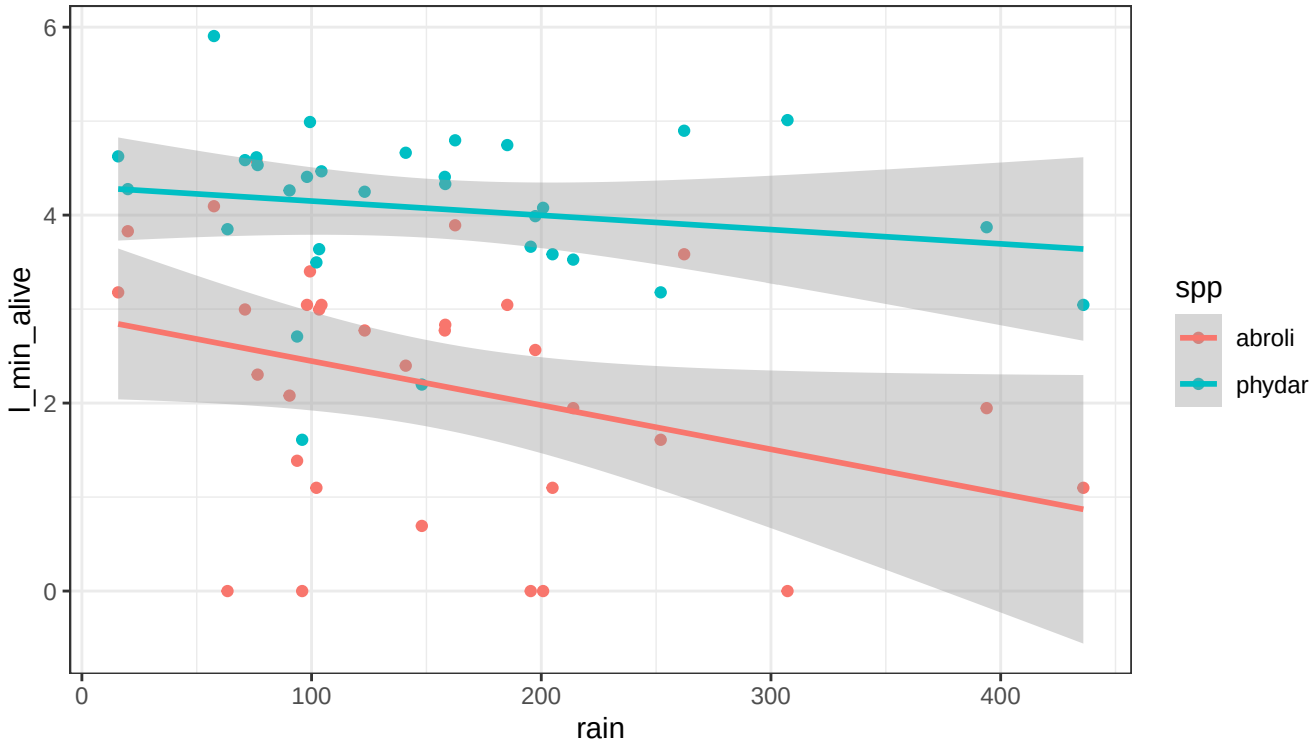
- `PD` = Phyllotis darwini minimun number alive
- `PDD` = Phyllotis darwini density (minimun number alive/trapping area)
- `XPD` = natural logarithm of P. darwini density
- `AO` = Abrothrix olivacea minimun number alive
- `AOD` = Abrothrix olivacea density (minimun number alive/trapping area)
- `XAO` = natural logarithm of Abrothrix olivacea density
- `RPD` = logarithmic (per capita) reproductive rate of Phyllotis darwini
- `RAO` = logarithmic (per capita) reproductive rate of Abrothrix olivacea
- `P` = Annual accumulated rainfall (mm)
- `P1` = one year lagged annual accumulated rainfall (mm)

[2] Link to the repository. You do not need to download anything from this website.

Is there evidence that minimum number of rodents alive depends upon the rainfall after accounting for differences between species? **Hint**: use the log transformed minimum number alive.

First, we create a plot to visualize the research question. The plot below shows the log of the minimum rodents alive by the rainfall for each species. There may be some evidence of a negative linear relationship, but it is honestly hard to tell since the relationships are so uncertain.

```
rodents_clean %>%
  ggplot(aes(x = rain, y = l_min_alive, col = spp)) +
  geom_point() +
  theme_bw() +
  geom_smooth(method = "lm", se = T)
```



Next, we fit a good ole fashion OLS model to see if that is sufficient to answer the question - note the form of the linear predictor! The research question asks if there is a relationship between the minimum number alive and the rainfall after accounting for the species; therefore, `l_min_alive ~ rain + spp` is the appropriate predictor.

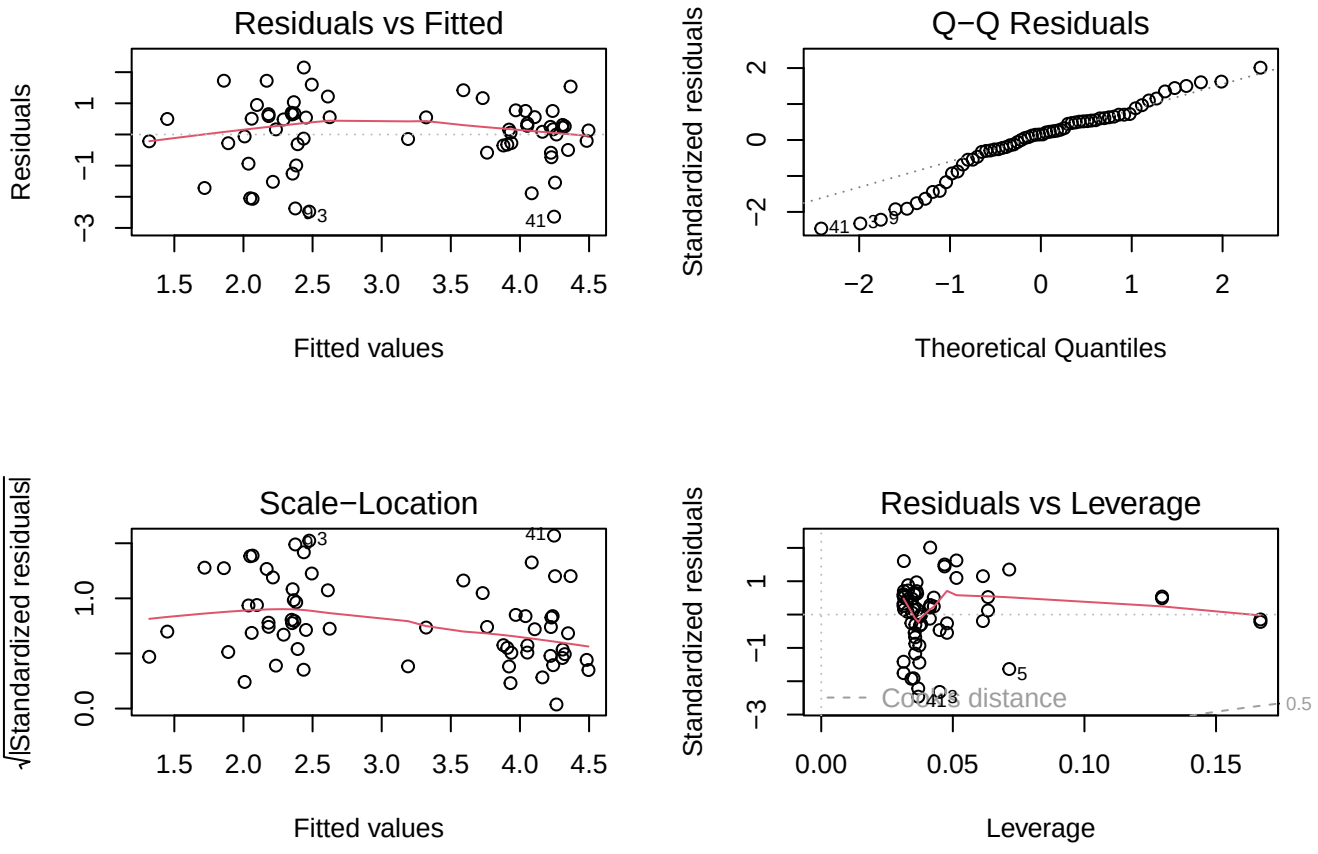
We then look at the diagnostic plots for the OLS model. In general, they look quite poor. There is some evidence of curvature in the residuals vs fitted suggesting a violation of linearity. The spread about the 0 line seems fairly constant, so there is not very much evidence against the constant variance assumption. The QQ plot looks terrible, with the points following off the hypothesized 1-1

line very early, suggesting an egregious violation of normality. Moreover, with only 64 observations, the CLT may not be enough to ensure approximately normally distributed sampling distributions for our regression coefficients. To be honest, the QQ plot is bad enough that we should consider using nonparametric methods. But we will ignore all the above for now. :)

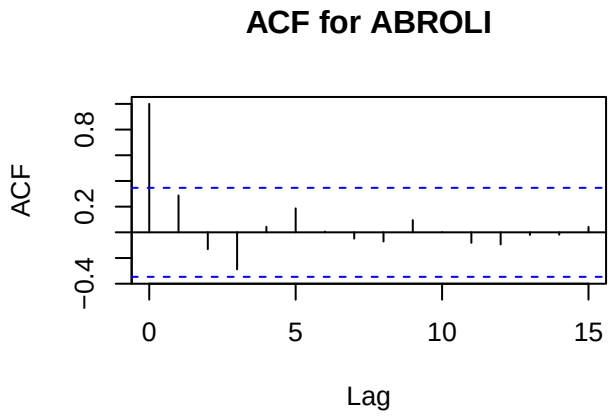
The PACF plot for PHYDAR suggests that there may be some evidence of an AR(1) process in the residuals. It is right at the blue bands, but the magnitude is fairly meaningful (about .35). So that is something we would like to address using GLS.

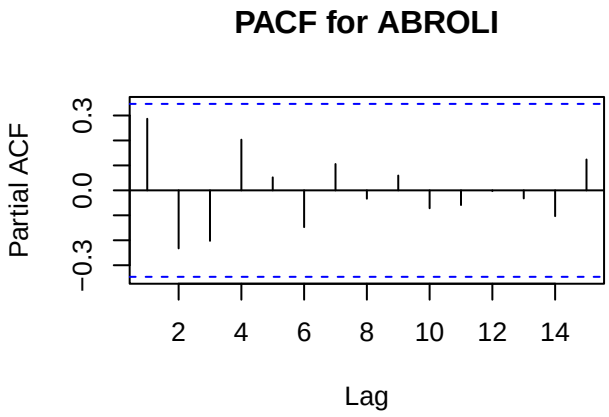
Finally, I will note that based on the OLS fit (which has problems, not the least of which being violations of independence), there is moderate evidence to suggest that there is a linear relationship between the log minimum number of rodents alive and the rainfall, after accounting for differences in species ($t = -2.184$ on 61 degrees of freedom yielding a p-value of 0.0328). But we should not base on conclusions on that p-value, as it is likely too small given the positive serial autocorrelation!

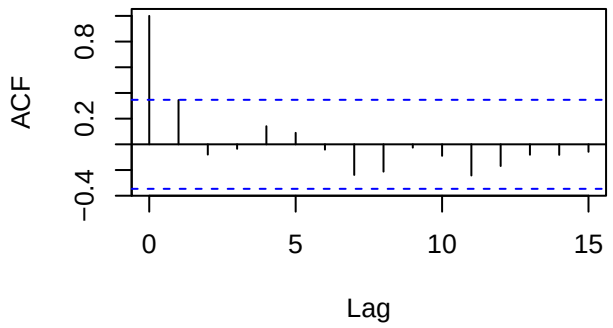
```
fit <- lm(l_min_alive ~ rain + spp, rodents_clean)
par(mfrow = c(2,2))
plot(fit)
```

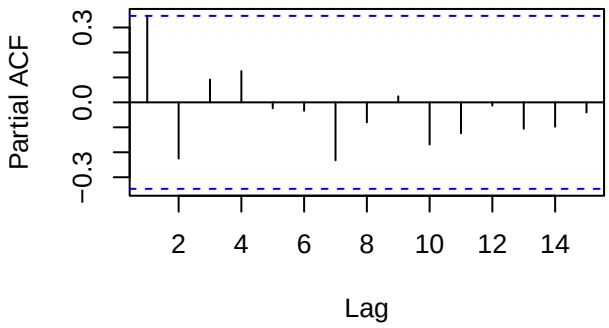


```
acf(resid(fit)[1:32], main = "ACF for ABROLI")
pacf(resid(fit)[1:32], main = "PACF for ABROLI")
acf(resid(fit)[33:64], main = "ACF for PHYDAR")
pacf(resid(fit)[33:64], main = "PACF for PHYDAR")
```










```
summary(fit)
```

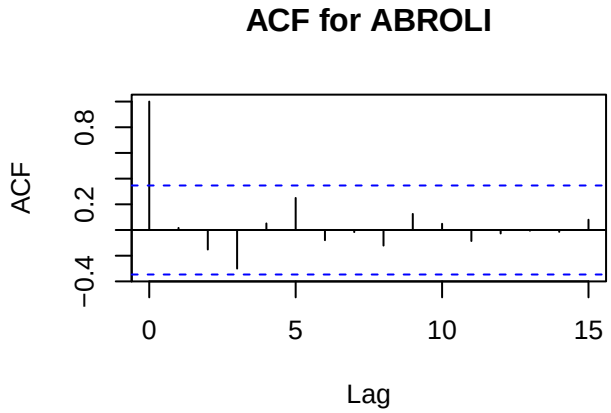
```
Call:
lm(formula = l_min_alive ~ rain + spp, data = rodents_clean)

Residuals:
    Min      1Q  Median      3Q     Max
-2.6376 -0.3901  0.1646  0.6458  2.1482

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.672793   0.291175   9.179 4.25e-13 ***
```

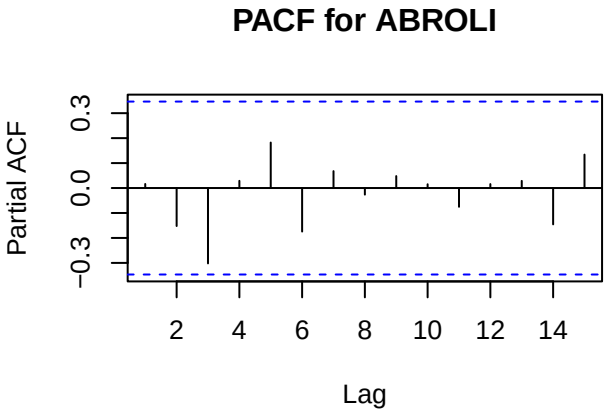
```
rain        -0.003106   0.001422  -2.184   0.0328 *
sppphydar    1.872167   0.272868   6.861 4.01e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.091 on 61 degrees of freedom
Multiple R-squared:  0.4594,    Adjusted R-squared:  0.4417
F-statistic: 25.92 on 2 and 61 DF,  p-value: 7.11e-09
```

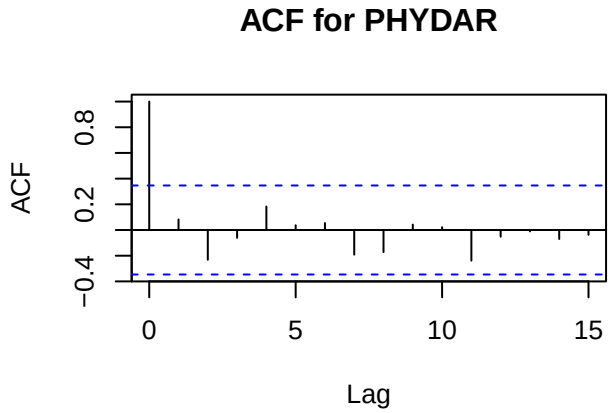
Next we fit the GLS model - note the form of the correlation structure! Looking at the PACF plots, the AR(1) correlation structure seemed to solve our problems. There is little-to-no evidence of serial correlation in the residuals. It is also worth noting that the estimated AR(1) parameter is .41, which is fairly large.

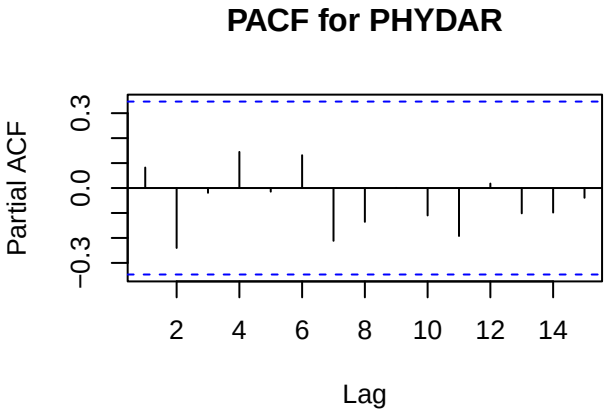
Ignoring some of the other assumption violations for now, let us take a look at the test results. After accounting for the serial correlation in the residuals and the differences between species, we no longer have evidence of a relationship between log minimum rodents alive and the rainfall ($t = -1.4699$ on 61 degrees of freedom yielding a p-value of 0.1467)! Hopefully this demonstration will serve as a stoic reminder of why it is important to account for serial correlation in the residuals! Ignoring serial correlation can absolutely lead you to the wrong conclusion.

```
fit_gls <- gls(
  l_min_alive ~ rain + spp, data = rodents_clean,
  correlation = corARMA(form = ~ 1 | spp, p = 1)
)
par(mfrow = c(2,2))
acf(resid(fit_gls, "normalized")[1:32], main = "ACF for ABROLI")
pacf(resid(fit_gls, "normalized")[1:32], main = "PACF for ABROLI")
acf(resid(fit_gls, "normalized")[33:64], main = "ACF for PHYDAR")
pacf(resid(fit_gls, "normalized")[33:64], main = "PACF for PHYDAR")
```










```
summary(fit_gls)
```

```
Generalized least squares fit by REML
  Model: l_min_alive ~ rain + spp
  Data: rodents_clean
       AIC      BIC    logLik
  205.0819 215.6363 -97.54097

Correlation Structure: AR(1)
 Formula: ~1 | spp
 Parameter estimate(s):
      Phi
0.4094209

Coefficients:
                 Value Std.Error   t-value p-value
(Intercept)  2.5350804 0.3508534  7.225469  0.0000
```

```
rain        -0.0017856 0.0012147 -1.469962  0.1467
sppphydar    1.8413253 0.4254986  4.327453  0.0001

 Correlation:
          (Intr) rain
rain      -0.514
sppphydar -0.606  0.000

Standardized residuals:
       Min         Q1        Med         Q3        Max
-2.3066023 -0.4919621  0.1389360  0.5790389  1.9421430

Residual standard error: 1.125349
Degrees of freedom: 64 total; 61 residual
```