Name:
Due: 2024/12/01

Exam 2

Be sure to submit **both** the .pdf and .qmd file to Canvas by Sunday, December 1st at 11:59 pm. The purpose of this exam is to synthesize some of the concepts we have discussed this semester, particularly with respect to forecasting and inference.

> **!** Important
>
> You may use any resource to complete this except **except** for other people (classmates, other faculty members, your parents, etc.) Please do not discuss this exam with anyone but me, and be sure to **cite all references and materials used to answer each question.** Please type your name below to acknowledge that you have not discussed this exam with anyone else.

Your name:

Question 1A [27 pt]

Understanding the dynamics between predator species, prey species, and how each group leverages their environment is a fundamental objective of ecology. To explore these dynamics, researchers collected monthly measurements of the activity of a predatory fish, the Eastern Mosquitofish (*Gambusia holbrooki*) and its littoral zone prey species, the least Killifish (*Heteraandria formosa*) in three regions of Northern Florida (`location`) using throw traps.[1] At each survey event, a throw net was cast three times and the average log density of the Mosquitofish (`me4loggambo`) and Killifish (`me2loghetads`) was recorded. Researchers also collected the average percent vegetative cover within the survey location (`cover1`), which is a measure of how the Killifish utilize surrounding vegetation to escape the Mosquitofish. Note that while no trends in fish activity over time are expected, researchers do anticipate monthly seasonal effects resulting from changes in temperature throughout the year.

Least Killifish

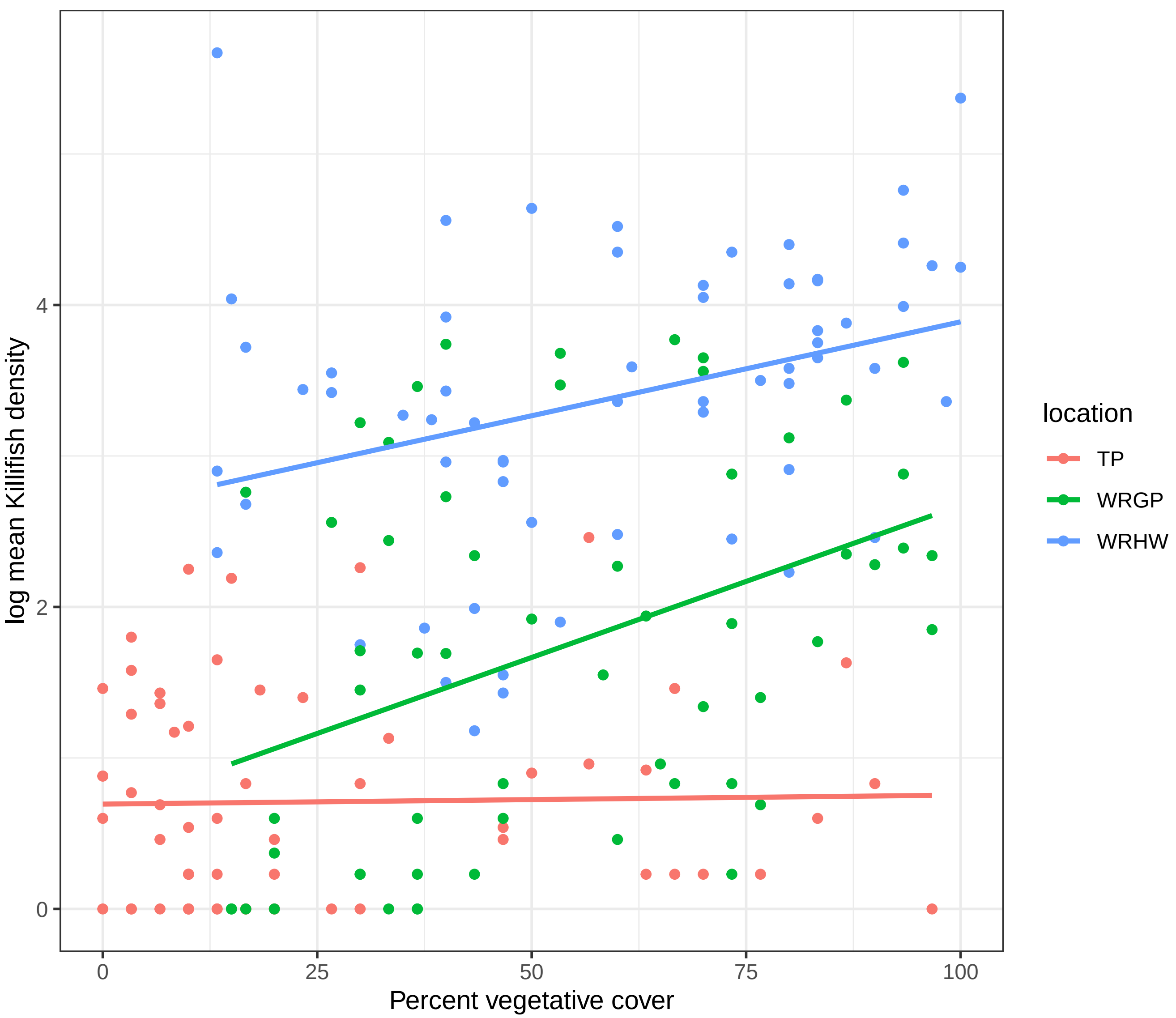
Eastern Mosquitofish

Researchers are interested in determining whether the relationship between the Killifish density and the percent vegetative cover depends upon the location, after accounting for the density of the Mosquitofish and any seasonal effects due to month. Such a result would imply that the Killifish utilize the vegetative cover differently across the three regions, even after accounting for seasonal differences and the impact of the predator species. Within the .qmd file, there is a code chunk that cleans the data for you; you should use the `fish_clean.rds` file for this analysis. For this first question, your goal is to use **ordinary least squares** to address the research question. Be sure to address the model assumptions, but you may ignore any violations of the independence assumption when answering the research question. Your response should be formatted as paragraph, which references an exploratory graphic, the model assumptions, and the results of the statistical test. Use the following rubric to guide your response:

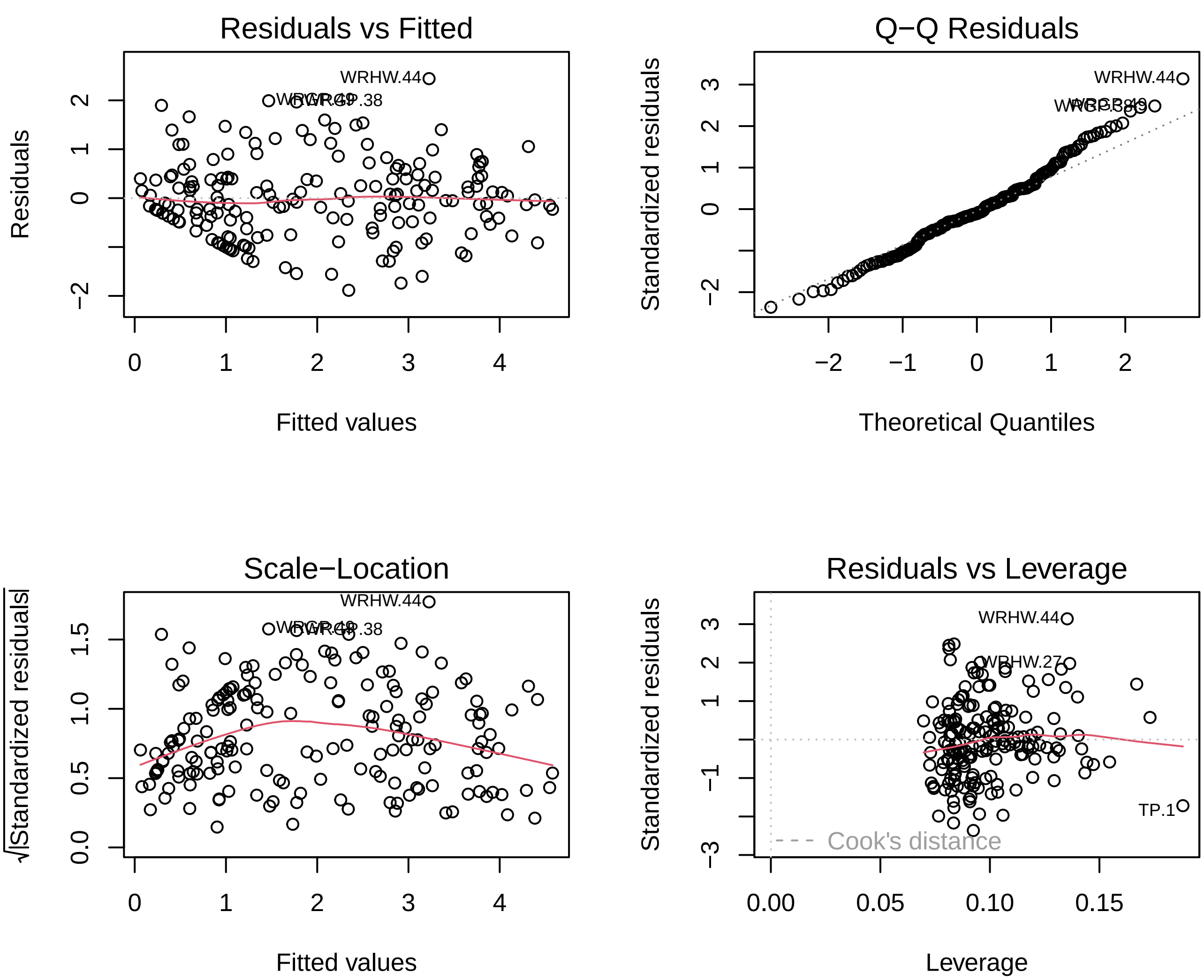
- [3 pt] Exploratory visual that addresses the research question is created and discussed
- [6 pt] Appropriate statistical model(s) used
- [6 pt] Model assumptions appropriately addressed
- [4 pt] Appropriate statistical test used to address research question
- [4 pt] A conclusion, supported by statistical evidence, is given and the evidence is referenced (statistic and p-value)
- [4 pt] Formatting (written response, output is reasonably clean, no callout box errors, complete sentences, spelling, figures of reasonable size, etc.)

[1] A link to the description of the data if you are curious. You do not need to download anything from this link.

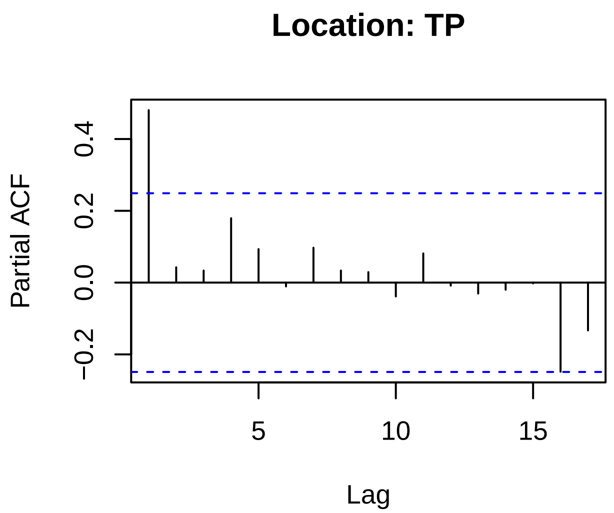
```
# read data
fish_clean <- readRDS("fish_clean.rds")

# exploratory plots
fish_clean %>%
  ggplot(aes(y = me2loghetads, x = cover1, col = location)) +
  geom_point() +
  theme_bw() +
  geom_smooth(method = "lm", se = F) +
  labs(
    y = "log mean Killifish density",
    x = "Percent vegetative cover"
  )
```

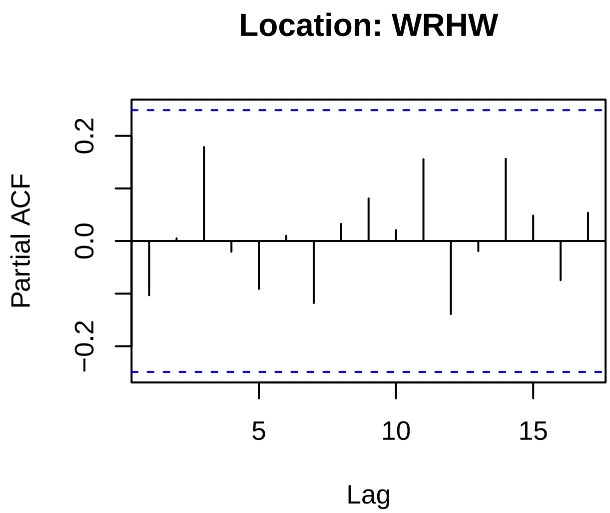


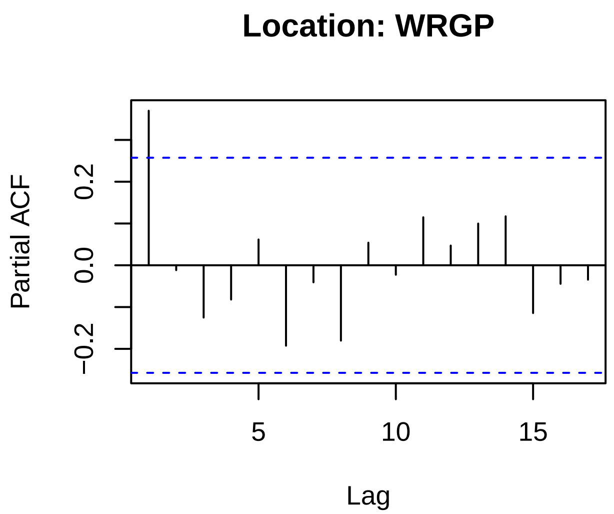
```
# statistical models
alt <-  lm(me2loghetads ~ me4loggambo + location*cover1 + month, fish_clean)
null <- lm(me2loghetads ~ me4loggambo + location + cover1 + month, fish_clean)

# plots for assumptions
par(mfrow = c(2,2))
plot(alt)
```




```
pacf(resid(alt)[1:62], main = "Location: TP") # location: TP
pacf(resid(alt)[63:124], main = "Location: WRHW") # location: WRHW
pacf(resid(alt)[125:182], main = "Location: WRGP") # location: WRGP
par(mfrow = c(1,1))
```








```
# tests
anova(null, alt)
```

```
Analysis of Variance Table

Model 1: me2loghetads ~ me4loggambo + location + cover1 + month
Model 2: me2loghetads ~ me4loggambo + location * cover1 + month
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1    166 119.46
2    164 115.02  2     4.433 3.1602 0.04501 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

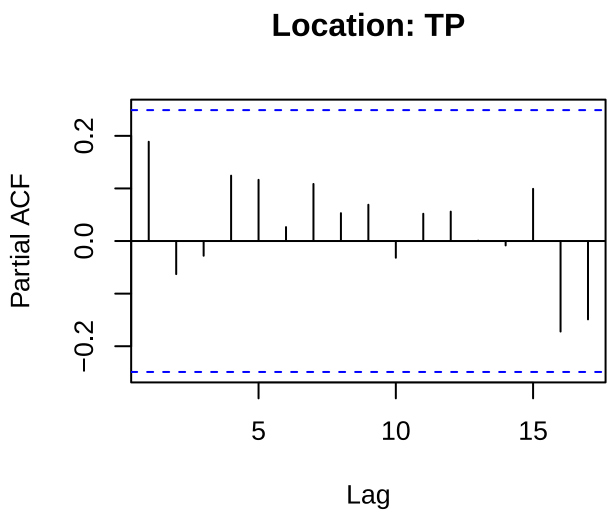
Question 1B [24 pt]

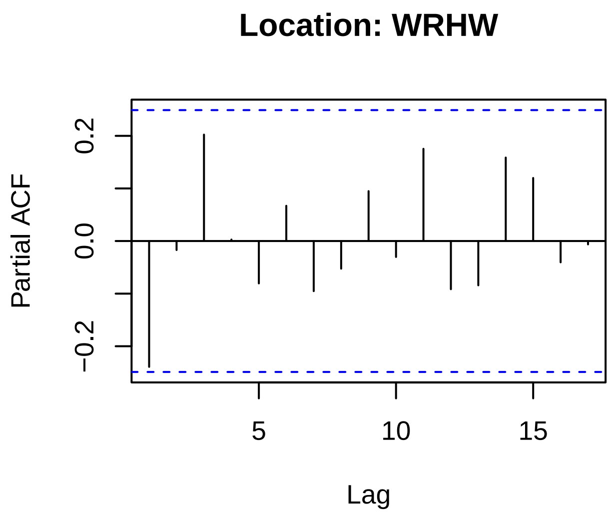
This is a continuation of question 1A. For this question, your goal is to use **generalized least squares** to address the research question. Be sure to address the model assumptions, and show that the model you fit resolves any violations of the independence assumption. Your response should be formatted as paragraph, which references the model assumptions, proof that the model you fit addresses any serial correlation in the error structure, and the results of the statistical test. Use the following rubric to guide your response:

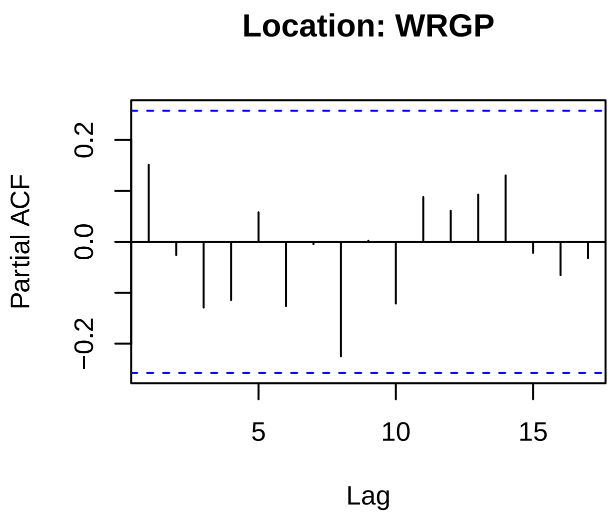
- [6 pt] Appropriate statistical model(s) used
- [6 pt] Model assumptions appropriately addressed. You need only reassess the independence assumption.
- [4 pt] Appropriate statistical test used to address research question
- [4 pt] A conclusion, supported by statistical evidence, is given and the evidence is referenced (statistic and p-value)
- [4 pt] Formatting (written response, output is reasonably clean, no callout box errors, complete sentences, spelling, figures of reasonable size, etc.)

```r
# models
library(nlme)
alt_gls <- gls(
  me2loghetads ~ me4loggambo + location*cover1 + month,
  data = fish_clean,
  correlation = corARMA(form = ~ 1 | location, p = 1, q = 1),
  method = "ML"
)
null_gls <- gls(
  me2loghetads ~ me4loggambo + location + cover1 + month,
  data = fish_clean,
  correlation = corARMA(form = ~ 1 | location, p = 1, q = 1),
  method = "ML"
)

# assumptions
par(mfrow = c(2, 2))
pacf(resid(alt_gls, type = "normalized")[1:62], main = "Location: TP")
pacf(resid(alt_gls, type = "normalized")[63:124], main = "Location: WRHW")
pacf(resid(alt_gls, type = "normalized")[125:182], main = "Location: WRGP")

# tests
anova(null_gls, alt_gls)
```

```
         Model df      AIC      BIC    logLik   Test  L.Ratio p-value
null_gls     1 19 457.4149 518.2910 -209.7074
alt_gls      2 21 458.7884 526.0726 -208.3942 1 vs 2 2.626405   0.269
```


Location: TP
Partial ACF
Lag
−0.2
0.0
0.2
5
10
15


Location: WRHW
Partial ACF
Lag
−0.2
0.0
0.2
5
10
15


Location: WRGP
Partial ACF
Lag
−0.2
0.0
0.2
5
10
15

Question 2 [17 pt]

For this question, we return to the Mauna Loa CO_2 measurements. As a reminder, this data set describes the average monthly CO_2 readings from the Mauna Loa observatory[2], which have been collected continuously since March of 1958. Suppose that researchers are interested in determining when the CO_2 emissions from Mauna Loa are expect to exceed 450 parts per million.

Your goal is to fit a model to the Mauna Loa time series, use that model to forecast the CO_2 emissions, and determine the month during which you are 95% confident that the CO_2 emissions are at least 450 ppm. Your response should be formatted as paragraph, and include the following:

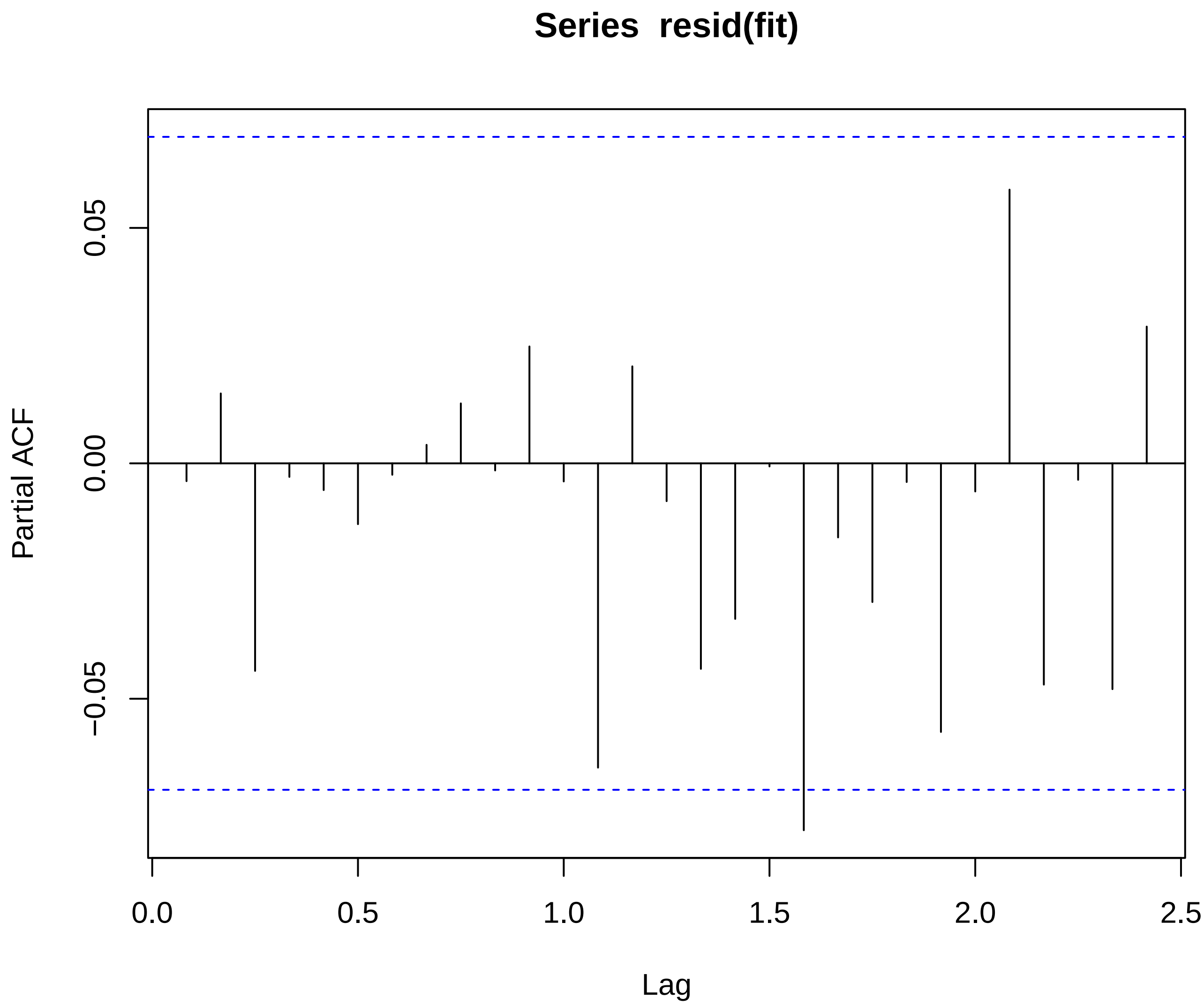
- [3 pt] Define the model (what kind of model is fit and whether it contains regression coefficients)
- [4 pt] Proof that this model accounts for the serial correlation in the errors
- [4 pt] A plot that visualizes the observed series, fitted series, and forecasted series (with 95% prediction intervals) in a single figure
- [4 pt] An answer to the research question
- [2 pt] Good formatting (written response, output is reasonably clean, no callout box errors, complete sentences, spelling, figures of reasonable size, etc.)

[2] Link to the NOAA webpage containing the data. You do not need to download anything from this website. You should use the `co2_mm_mlo.csv` file.

```
# clean data
library(forecast)
ml <- read_csv("co2_mm_mlo.csv", skip = 40)
ml_clean <- ml %>%
  dplyr::select(year, month, average) %>%
  mutate(dt = ym(paste0(year, "-", month)))
```

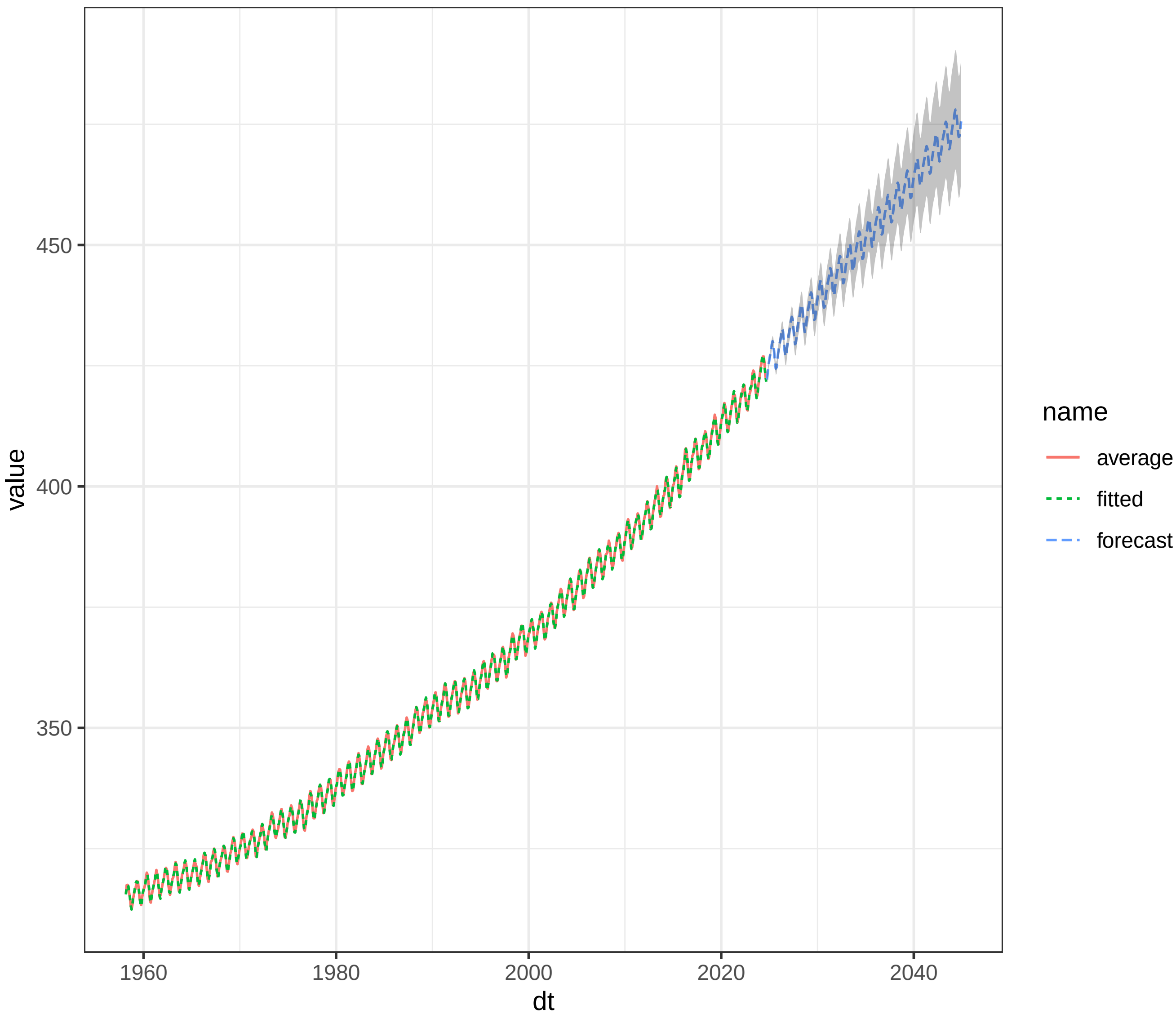
```
ml_clean %>%
  ggplot() +
  geom_line(aes(x = dt, y = average)) +
  theme_bw()

ml_ts <- ts(
  ml_clean$average,
  start = c(1958, 3),
  freq = 12
)

library(forecast)
fit <- auto.arima(ml_ts)
saveRDS(fit, "ml_fit.rds")
```

```
fit <- readRDS("ml_fit.rds")
forecast <- forecast(fit, h = 3+12*20, level = 95)
# plot
forecast_df <- tibble(
  year = c(rep(2024, 3), rep(2025:(2025+19), each = 12)),
  month = c(10:12, rep(1:12, 20)),
  value = forecast$mean,
  lwr = c(forecast$lower),
  upr = c(forecast$upper)
) %>% mutate(name = "forecast", dt = ym(paste0(year, "-", month)))

obs_df <- ml_clean %>%
  mutate(
    fitted = fitted(fit)
  ) %>%
  dplyr::select(dt, fitted, average) %>%
  pivot_longer(-dt)

pacf(resid(fit))
```




```r
ggplot() +
  geom_line(
    data = obs_df,
    aes(x = dt, y = value, col = name, linetype = name)
  ) +
  geom_line(
    data = forecast_df,
    aes(x = dt, y = value, col = name, linetype = name)
  ) +
  geom_ribbon(
    data = forecast_df,
    aes(x = dt, ymin = lwr, ymax = upr), alpha = .3
  ) +
```

```
theme_bw()
```




```
forecast_df %>% filter(lwr > 450)
```

```
# A tibble: 90 x 7
   year month value   lwr   upr name     dt
  <dbl> <int> <dbl> <dbl> <dbl> <chr>    <date>
1  2036     4  457.  450.  464. forecast 2036-04-01
2  2036     5  458.  451.  465. forecast 2036-05-01
3  2036     6  457.  450.  465. forecast 2036-06-01
4  2037     3  458.  451.  466. forecast 2037-03-01
5  2037     4  460.  452.  467. forecast 2037-04-01
```

```
 6  2037     5  460.  453.  468. forecast 2037-05-01
 7  2037     6  460.  452.  468. forecast 2037-06-01
 8  2037     7  458.  450.  466. forecast 2037-07-01
 9  2038     1  459.  451.  467. forecast 2038-01-01
10  2038     2  460.  452.  468. forecast 2038-02-01
# i 80 more rows
```