Name: Your name here
Due: 2024/11/06

Day 18 - Lab: Regression inference

Introduction

In this assignment, we will get some more practice with analyzing real data using time series regression models. To motivate our efforts, we will use average monthly CO_2 readings from the Mauna Loa observatory[1], which have been collected continuously since March of 1958. Is there evidence that the magnitude of the seasonal effect has increased over time, after accounting for the trend?

[1] Link to the NOAA webpage.

Regression inference

In this class, we focus on two tools for making inference about regression models: the t-test and the extra sum of squares F-test. We introduce these methods briefly here, but expect to learn more about each in a regression class. The t-test, which is obtained from the `summary` output of a model, tests

$$H_0 : \beta = 0 \quad \text{vs} \quad H_a : \beta \neq 0$$

holding all other predictor variables constant in the model. The distribution of

$$t = \frac{\hat{\beta}}{SE(\hat{\beta})}$$

under the null hypothesis is t_{df}, where df denotes the degrees of freedom associated with the residual error.

The extra sum of squares F-test is used to compare nested models. For example, consider two models:

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \epsilon_i \\ y_i &= \beta_0 + \beta_1 x_{1i} + \epsilon_i \end{aligned}$$

We say that the second model is nested within the first model, since all parameters in the second model also exist in the first model. We can use these two models to test the following hypotheses:

$$H_0 : \beta_2 = \beta_3 = 0 \quad \text{vs} \quad H_a : \text{at least one } \beta_i \text{ does not equal 0 for } i \in \{2, 3\}$$

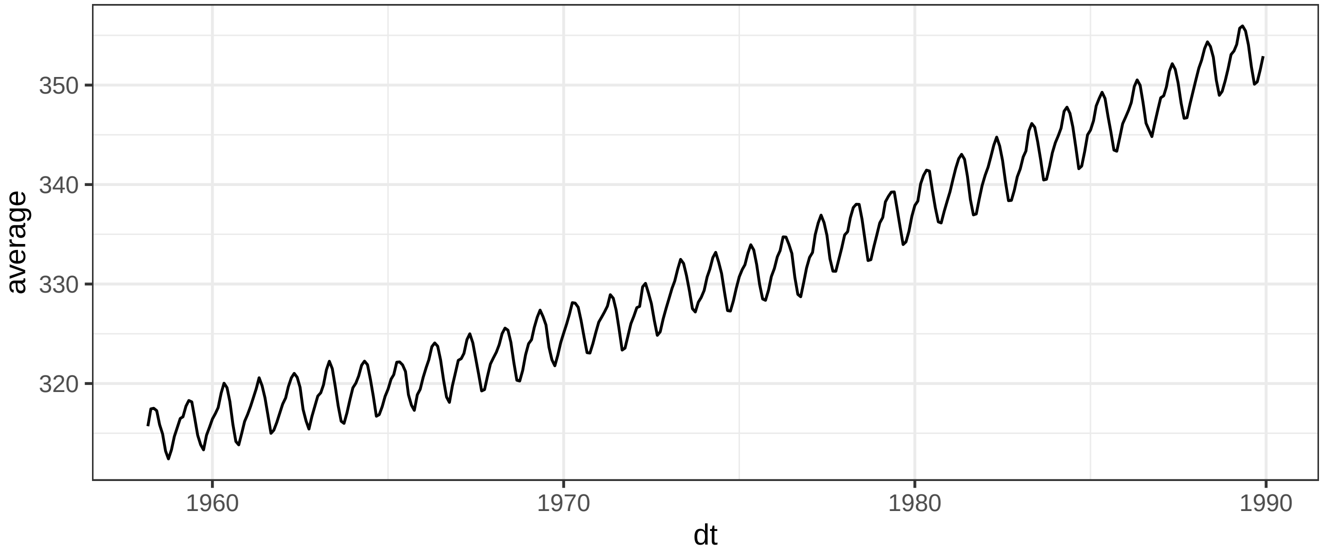
Under the null hypothesis, the distribution of the test statistic is $F_{2,df}$, which is an F distribution with two numerator and df denominator degrees of freedom, where df is the degrees of freedom associated with the residual error. To perform this test in R, we use the `anova` function: `anova(fit1, fit2)` compares models `fit1` and `fit2` in this way.

1. [1 pt] Read the Mauna Loa data in to R and filter the data to contain years prior to 1990 (not including 1990).

```
co2 <- read_csv("co2_mm_mlo.csv", skip = 40) %>%
  filter(year < 1990) %>%
  mutate(dt = ym(paste0(year, "-", month))) %>%
  mutate(
    t = 1:n(),
    month = factor(month)
  )
```

2. [2 pt] Plot the average CO_2 mole fraction (ppm) over time and describe the series in terms of trend and seasonality.

```
co2 %>% ggplot() + geom_line(aes(x = dt, y = average)) + theme_bw()
```



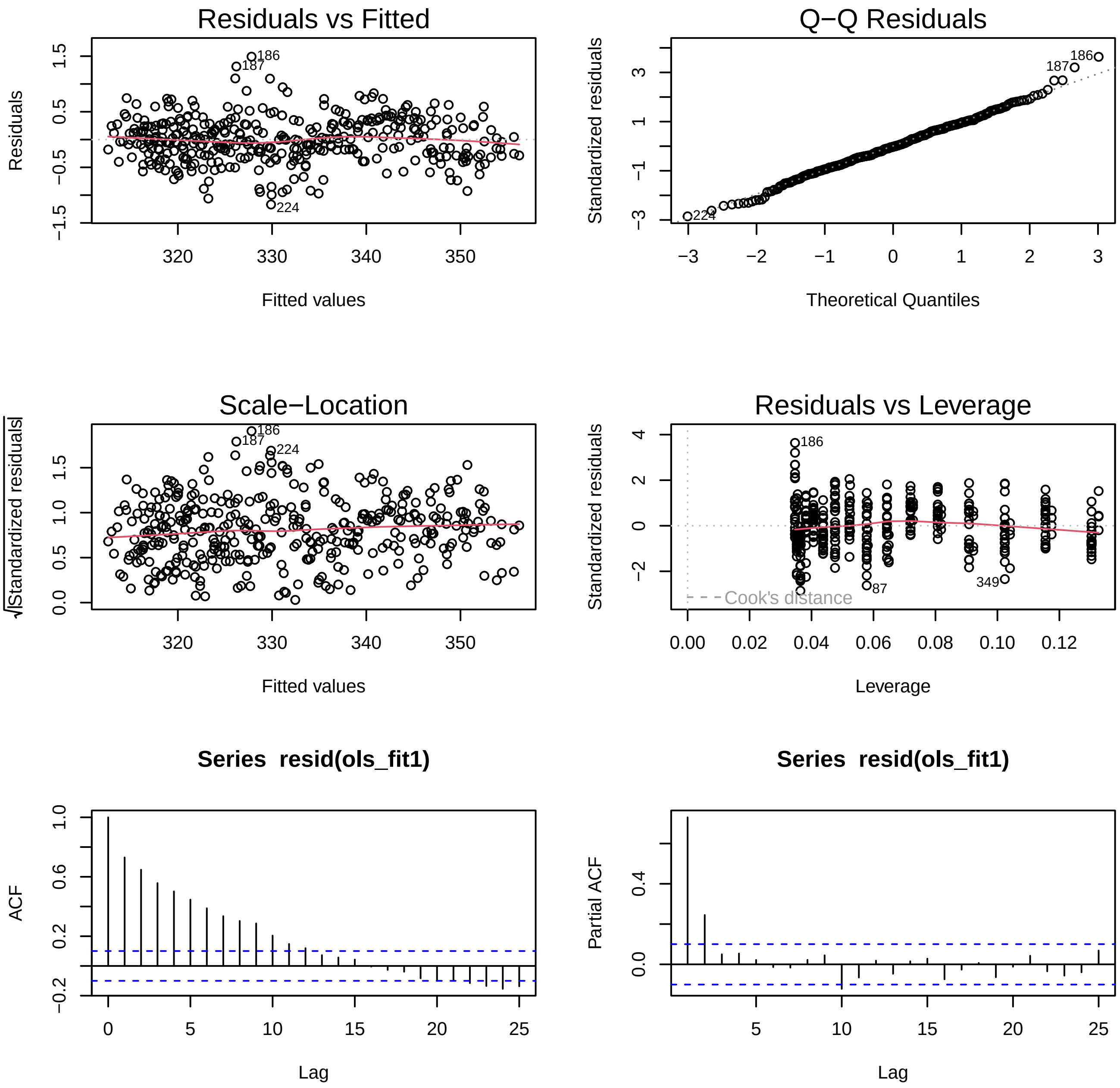
There seems to be an increasing quadratic relationship with time, and a clear seasonal effect in which CO_2 peaks during May and June.

3. [2 pt] Fit two regression models using ordinary least squares: one that includes a time by month interaction and time squared, and one that includes time, time squared, and month (no interaction).

```
ols_fit1 <- lm(average ~ t*month + I(t^2), co2)
ols_fit2 <- lm(average ~ t + month + I(t^2) , co2)
```

4. [4 pt] Assess the independence, normality, constant variance, and linearity assumptions for the interaction model.

```r
par(mfrow = c(3, 2))
plot(ols_fit1)
acf(resid(ols_fit1))
pacf(resid(ols_fit1))
```



> **Independence:** We expect a violation of independence because this is a time series course. :) Observations are collected sequentially over time, and we expect observation collected close together in time to be more similar than observations further apart. This is exactly what we see in the ACF and PACF plots of the residuals, which suggest an AR(2) process is present.
> **Linearity:** The squared term seemed to solve any potential linearity issues. The residuals vs fitted values plot shows no apparent trends, suggesting this assumption is reasonably satisfied.
> **Constant variance:** There is no evidence of fanning in the residuals vs fitted plot, and no trends in the scale-location plot. Therefore, we have no reason to question this assumption.
> **Normality:** The point follow the hypothesized quantile line remarkably well in the QQ plot, again suggesting no reason to question this assumption.

5. [2 pt] Ignore any violations of the assumptions for now. Is the additive model nested within the interaction model? Why or why not?

> Yes - all terms in the additive model are also in the interaction model.

6. [2 pt] Use an extra sums of squares F-test to address the research question and write up a conclusion that references: the test statistic, distribution of the test statistic under the null hypothesis, and the p-value. Ignore the violation of independence for now.

```r
anova(ols_fit2, ols_fit1)
```

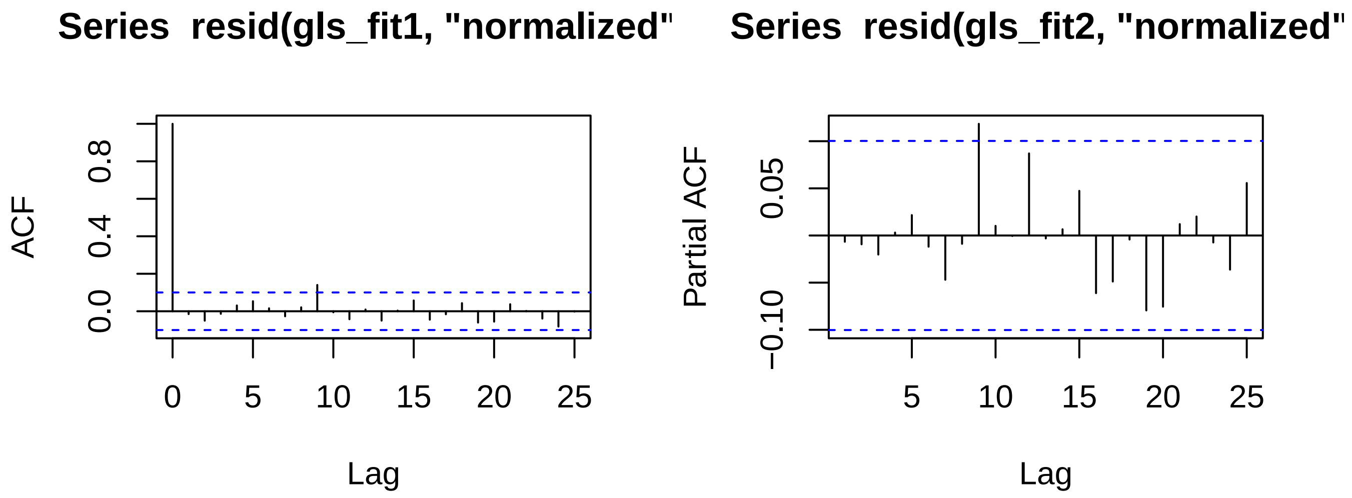
```
Analysis of Variance Table

Model 1: average ~ t + month + I(t^2)
Model 2: average ~ t * month + I(t^2)
  Res.Df    RSS Df Sum of Sq      F   Pr(>F)
1    368 67.246
2    357 62.155 11    5.0917 2.6587 0.002751 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

> There is strong evidence to suggest that the interaction model provides a better explanation of the data than does the additive model ($F = 2.6587$ with 11 numerator and 357 denominator degrees of freedom, yielding a corresponding p-value of 0.0028).

7. [4 pt] Refit both regression models using GLS and incorporate an AR(2) correlation structure in each model **using maximum likelihood as the method**. Reassess the independence assumption for the interaction model using the new fit.

```r
gls_fit1 <- gls(average ~ t*month + I(t^2), co2, correlation = corARMA(p=2), method =
gls_fit2 <- gls(average ~ t + month + I(t^2), co2, correlation = corARMA(p=2), method
par(mfrow = c(1,2))
acf(resid(gls_fit1, "normalized"))
```

```
pacf(resid(gls_fit2, "normalized"))
```



> Things look much better; the serial autocorrelation is no longer present at lags 1 or 2. There is a strange artifact at lag 9, but I wouldn't worry about it. It is fairly small in magnitude.

8. [2 pt] Recreate the extra sums of squares F-test using the GLS models and comment on the differences.

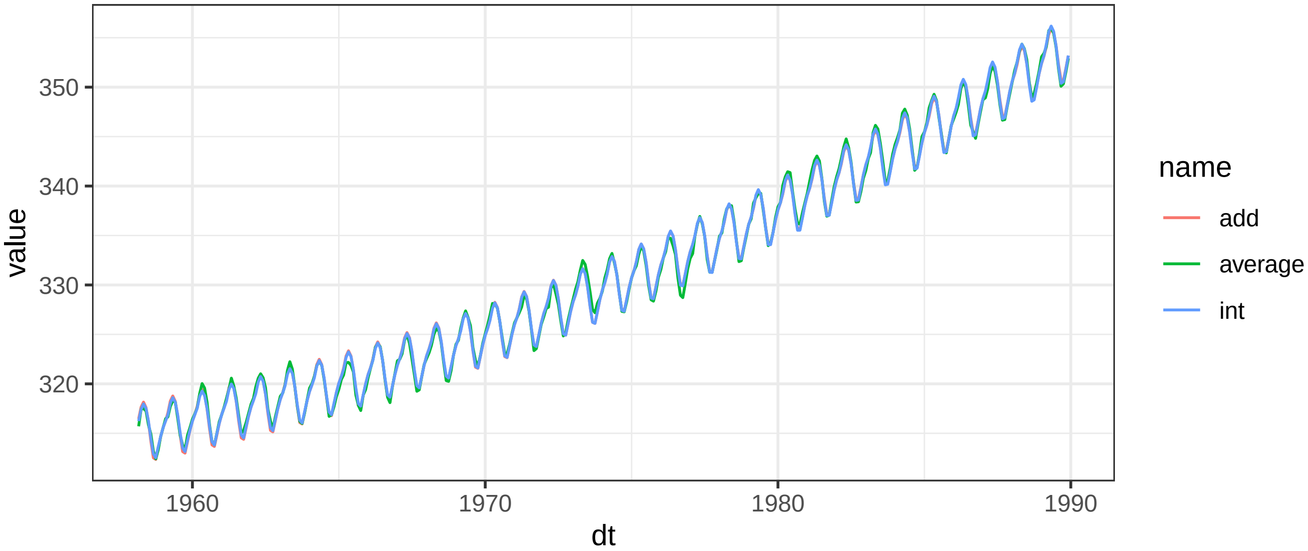
```
anova(gls_fit1, gls_fit2)
```

```
         Model df      AIC      BIC    logLik   Test  L.Ratio p-value
gls_fit1     1 28 129.6363 240.1081 -36.81815                        
gls_fit2     2 17 140.7732 207.8453 -53.38659 1 vs 2 33.13688   5e-04
```

> The test statistic changes all together! It is beyond the scope of this course, but extra sums of squares F-tests with models fit using GLS are extremely complicated, and an F-statistics cannot be obtained. Instead, we use a new test called a likelihood-ratio test. The p-value gets at the same idea though - the interaction model is better than the additive model.

9. [2 pt] Plot the raw time series, fitted values from the interaction model, and fitted values from the additive model on a single plot and comment on the differences.

```
co2 %>%
  dplyr::select(dt, average) %>%
  mutate(
    int = fitted(gls_fit1),
    add = fitted(gls_fit2)
  ) %>%
  pivot_longer(average:add) %>%
  ggplot() +
  geom_line(aes(x = dt, y = value, col = name)) +
  theme_bw()
```



The two models are pretty much the same, yet the p-value suggests that the interaction model is much better. This is a word of caution with p-values! Once sample sizes get large, the SEs get so small that it is very easy to find statistically significant results, even if the differences are not practically different.