EPS 659a — Data Analysis in Earth Sciences

Problem Set Six

due Monday, Oct 18, 2021

Carbon Dioxide at Mauna Loa, Hawaii since 1958: Trends and Cycles.

On the Canvas server there is a dataset of carbon-dioxide values measured monthly at Mauna Loa, Hawaii USA in csv format (`co2_maunaloa.csv`). There are four columns, two for integer years and months, another for digital years and the last column for parts-per-million CO_2 in the atmosphere. These numbers were monthly-averaged from air parcels taken atop Mauna Loa volcano.[1]

We will be looking at linear-model fits to the CO_2 time series to explore its evolution over the Professor's lifetime. Start your R-scripts with the following code to read the data file:

```
MaunaLoa <- read.csv("co2_maunaloa.csv")
names(MaunaLoa)
year <- MaunaLoa$YEAR
month <- MaunaLoa$MONTH
time <- MaunaLoa$TIME
rtime <- time-1990 co2 <- MaunaLoa$CO2
plot(time,co2,"l")
abline(lm(co2~time))
title(main="Mauna Loa Carbon Dioxide (ppm)")
```

reads and plots the data from March 1958 to August 2021, a total of 762 monthly values.

Problem 1: Fitting Trends Over the Entire Data Series.

(a) Fit the CO_2 data with a linear-trend analysis (`lm_co2 <- lm(co2~time)`). What is the linear model's increase of CO_2 concentration per-year and per-decade? Explore the data-fit residuals versus time and their QQ plot with the R command `plot(lm_co2)` and comment on how well the linear model succeeds.

(b) What is the nominal start-point of the trend (March value in 1958) and the endpoint (mid-2021 value) according to the linear-trend model? (Use the R command `predict(lm_co2)` to find these values.) How do these compare with the data itself near the endpoints of the time series?

(c) What is the standard deviation of the data-misfit residual? What is the F variance ratio for the model (referenced to one degree of freedom for the trend parameter – the degree-of-freedom for the constant is not treated as a fitted parameter by R)

(d) Plot the data against the linear-trend model and identify its shortcomings. You can obtain a time-series of predicted data with the R-function `predict`, e.g.,

```
lm_co2 <- lm(co2~time)
summary(lm_co2)
pred_co2 <- predict(lm_co2)
plot(time,co2,"l")
lines(time,pred_co2)
```

[1]The motivation for using Mauna Loa was that its summit is free of vegetation, and at altitude high enough to be free of local effects of low-altitude plant growth. In the mid-20th-century Mauna Loa was thought to average global CO_2 effectively. Later measurements have shown that Earth's CO_2-levels vary geographically in a dynamic annual cycle, even at high altitude.

(e) Fit the CO_2 data x_n with a quadratic polynomial of time t_n, so that

$$x_n = A + Bt_n + Ct_n^2$$

The quadratic model can be fit in R with the same command as the linear model, with altered syntax

```
time2 <- time^2
lm2_co2 <- lm(co2~time+time2)
summary(lm2_co2)
pred2_co2 <- predict(lm2_co2)
plot(time,co2,"l")
lines(time,pred2_co2)
```

What is the nominal start-point of the trend (that is, the value in 1958) and the endpoint (2021 value) according to the quadratic-trend model? How do these compare with the data itself near the endpoints of the time series?

(f) for the quadratic model, what are the yearly increases in CO_2 at the start and end of the time series? You can estimate these by differencing points 13 and 1 and points 762 and 750 in the "predicted" time series `pred2_co2` computed above.

(g) What is the standard deviation of the data-misfit residual? What is the F variance ratio for the model (referenced to two degrees of freedom for the trend and quadratic parameter – the degree-of-freedom for the constant is not treated as a fitted parameter by R)

(h) Plot the data against the quadratic-trend model and identify the model's shortcomings.

(i) Try a cubic-polynomial fit

```
time2 <- time^2
time3 <- time^3
lm3_co2 <- lm(co2~time+time2+time3)
summary(lm3_co2)
```

What is the standard deviation of the data-misfit residual? How significant is the cubic parameter? If your estimates of significance are confusing, try referencing your time ordinate to the late 20th century to avoid corrolating large numbers:

```
tim <- time - 1990
tim2 <- tim^2
tim3 <- tim^3
lm3_co2 <- lm(co2~tim+tim2+tim3)
summary(lm3_co2)
```

Problem 2: Fitting Trends and Cycles Over the Entire Data Series.

Generate two annual-cycle data-representers `ann_s,ann_c` using the sine and cosine functions in R with its fixed-constant value `pi`. Then add these two data representers to the linear model, along with the quadratic polynomial

```
pi2 <- 2*pi
ann_s <- sin(pi2*time)
ann_c <- cos(pi2*time)
lm2a_co2 <- lm(co2 ~ time + time2 +ann_s + ann_c)
```

(a) What is the amplitude of the annual cycle in CO_2, measured in ppm? Recall that you must sum the two amplitudes with the sum-of-squares Pythagoras formula. Use the R-function `coef` to extract the values. The uncertainty of the amplitude can likewise be computed from the parameter uncertainties. Report the amplitude with its uncertainty

```
coef_s <- coef(summary(lm2a_co2))["ann_s","Estimate"]
coef_c <- coef(summary(lm2a_co2))["ann_c","Estimate"]
coef_ann <- sqrt(coef_s^2 + coef_c^2)
dcoef_s <- coef(summary(lm2a_co2))["ann_s","Std.  Error"]
dcoef_c <- coef(summary(lm2a_co2))["ann_c","Std.  Error"]
dcoef_ann <- sqrt(dcoef_s^2 + dcoef_c^2)
```

(b) When within the year is the maximum of the annual cycle of CO_2? If `coef_s=0` and `coef_c` >0, then the maximum occurs at the start of the year, because the maximum of the $\cos(2\pi t)$ function occurs at $t = 0$. For all other cases the timing of the annual maximum can be ascertained from the phase delay $\phi = \tan^{-1}(a_s/a_c)$, where a_s, a_c are the amplitude of the sine and cosine terms, respectively. In R one can use the function `atan2(coef_s,coef_c)` to obtain the phase angle in radians, then divide this phase angle by 2π to obtain a fraction of the cycle, e.g., $\phi = \pi$ implies a half-cycle delay, so that the annual-cycle CO_2-maximum occurs at the end of June. Make sure that you look at the plotted annual cycle in the data to confirm that your computed maximum time is consistent with reality.

(c) Subtract the quadratic model from the data (`resid <- co2 - pred2_co2`) to obtain a trend residual. Then subtract the quadratic+annual-cycle model from the data to obtain *its* residual, then plot the two residuals against each other. What are their respective standard deviations?

(d) Add a semi-annual cycle to the linear model for Mauna Loa CO_2 as a harmonic modification to the annual cycle.

```
pi4 <- 4*pi
ann2_s <- sin(pi4*time)
ann2_c <- cos(pi4*time)
lm2a2_co2 <- lm(co2 ~ time + time2 + ann_s + ann_c + ann2_s + ann2_c)
```

What is the amplitude of the semi-annual cycle of CO_2, with its uncertainty? In what way does the semi-annual cycle change the shape of the annual cycle? For instance, how does it change the timing and the amplitudes of the annual maximum and minimum of CO_2? (Qualitative description desired, rough quantitiative estimates OK)

(e) Subtract the two-cycle model from part (d) from the CO_2 data and plot the result. What is the standard deviation of the residual?