EPS 659a — Data Analysis in Earth Sciences

Problem Set Seven

due Friday, Oct 29, 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.

We will continue to pursue 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 and Cycles Over 10-year Data Segments.

We ask whether the annual cycle of CO_2 at Mauna Loa has changed significantly since 1958. Translate this problem into a sequence of estimates of the quadratic+annual-cycle model in seven overlapping 10-year (120-month) data segments. In R, this calculation can be effected with a `for` loop over a collection of data-segment start points, each starting in January to preserve a uniform phase angle. You can slice out segments of the data representers to match the data segments. The regression parameters for each data segment can be saved in arrays that you must declare before the loop. You fill the values of these parameter arrays into your arrays for plotting. Plot the data-segment values against the midpoints of the data segments.

(a) compute and plot the annual-cycle amplitudes for decade-long data segments of the Mauna Loa CO_2 time series.

```
# pre-define arrays for regression parameters of seven data segments
ann1 <- 1:7
ann1_err <- 1:7
yr <- 1:7
# initialize segment counter
k=0
# the data segments are defined by their start-points
for (j in c(1,109,217,321,429,537,637)){
k=k+1
timed <- time[j:(j+119)]
time2d <- timed^2
co2d <- co2[j:(j+119)]
ann_sd <- ann_s[j:(j+119)]
```

```
ann_cd <- ann_c[j:(j+119)]
ann2_sd <- ann2_s[j:(j+119)]
ann2_cd <- ann2_c[j:(j+119)]
lmlmlm <- lm(co2d ~ timed + time2d + ann_sd + ann_cd + ann2_sd + ann2_cd)
summary(lmlmlm)
anns <- coef(summary(lmlmlm))["ann_sd","Estimate"]
annc <- coef(summary(lmlmlm))["ann_cd","Estimate"]
ann1[k] <- sqrt(anns^2+annc^2) yr[k] <- year[j+60]
anns <- coef(summary(lmlmlm))["ann_sd","Std.  Error"]
annc <- coef(summary(lmlmlm))["ann_cd","Std.  Error"]
ann1_err[k] <- sqrt(anns^2+annc^2)
}
```

One way to plot parameter estimates with their uncertainties applies the `arrows` function, see `https://www.benjaminbell.co.uk/2019/04/how-to-add-error-bars-in-r.html`, which recommends

```
arrows(x0=yr,y0=ann1+ann1_err,x1=yr,y1=ann1-ann1_err,code=3,angle=90,length=0.2)
```

Describe any significant changes in annual-cycle amplitude that are evident. Is there a trend? Is there a transition? Or is the variability consistent with random fluctuations about a fixed mean value?

(b) Compute and plot the annual-cycle phases for decade-long data segments of the Mauna Loa CO_2 time series. Express these phase changes in degrees, by multiplying the output of `atan2()` by $180/\pi$.

```
anns <- coef(summary(lmlmlm))["ann_sd","Estimate"]
annc <- coef(summary(lmlmlm))["ann_cd","Estimate"]
ann_amp <- sqrt(anns^2 + annc^2)
ann_phase[k] <- 180*atan2(anns,annc)/pi
# the phase uncertainty in radians is the relative uncertainty in coefficient amplitude
# multiply this phase uncertainty by 180/π to express uncertainty in degrees
anns <- coef(summary(lmlmlm))["ann_sd","Std.  Error"]
annc <- coef(summary(lmlmlm))["ann_cd","Std.  Error"]
ann_pherr[k] <- (180/pi)*sqrt(anns^2+annc^2)/ann_amp
```

Each degree of phase in the annual cycle translates approximately into one day. How many days has the annual cycle of CO_2 shifted since 1958?

(c) Compute and plot the semi-annual-cycle amplitudes for decade-long data segments of the Mauna Loa CO_2 time series. Describe any changes since 1958.

Problem 2: Fitting an Exponential Trend Over the Full Data Series.

Fossil-fuel consumption is closely tied to economic activity. On a global scale, economic growth has roughly followed the path of exponential growth, with aggregate GGP (Gross Global Product) expanding by a stable percentage every year (on average!). This history motivates us to model the Mauna Loa CO_2 time series with exponential-growth models.

(a) Fit CO_2 values $x_n = B\exp(Ct_n)$ with B and C fixed constants. B is the initial amplitude at $t_n = 0$ (i.e., at 1958) and C is the exponential growth rate. For small values of C and t_n in units of years, we can use the power-series representation of the exponential

$$\exp(Ct_n) = 1 + Ct_n + (Ct_n)^2/2! + (Ct_n)^3/3!....$$

to identify C with the instantaneous yearly growth rate. The exponential-growth model is nonlinear, but this particular model can be made linear by taking the logarithm of the data:

$$\log(x_n) = \log(B) + Ct_n = \tilde{B} + Ct_n$$

which can be fit with the R-function `lm(log(co2)~time)`. Perform this regression and plot the fitted linear model against the logarithm of CO_2.

(b) Transform this linear model fit back into the original frame of reference by taking the exponential of the model predictions. What are the values of this model at the endpoints of the time series in 1958 and 2021? Plot the model aganst the Mauna Loa CO_2 data and comment on the shortcomings of the data fit.

(c) The shortcomings of the exponential-growth model in (a) can be overcome, at least partly, by changing the model to include both a constant and an exponential term:

$$x_n = A + B\exp(Ct_n)$$

This model is also nonlinear, and cannot be converted into a linear model by a simple mathematical scaling. We must fit for the parameters with a nonlinear least-squares regression. R has this functionality, see the function `nls` described on the internets by googling "R nonlinear regression." Linear regression and nonlinear regression share the same goal of finding parameter values that minimize the residual misfit variance between the data and the model-prediction. Linear regression computes its solution in one matrix-inversion step. Nonlinear regression typically requires several matrix-inversion steps to find a solution. The `nls` algorithm uses linear-regression techniques at each iteration to get closer and closer to the final answer, just like a beginning golfer putts his/her golfball closer and closer to the hole with each stroke. You must give the algorithm a starting guess for the solution, from which it starts to converge on a proper solution. You can estimate a rough first guess for the solution by plugging a few values into the model (`astart, bstart, cstart`) to see if the prediction is anywhere close, say, within 50% of the data values.

```
# nonlinear fit of log(co2)
# subtract 1788 from time before regression, # to use pre-industrial initial condition
# the nls function in R asks for initial values for the regression parameters
# which are identified by the start parameter
time0 <- time-1788
nlm_co2 <- nls(co2 ~ a+b*exp(c*time0),start=list(a=astart,b=bstart,c=cstart))
predexp <- predict(nlm_co2)
plot(time,co2,"l")
title(main="Mauna Loa Carbon Dioxide (ppm)")
lines(time,predexp)
```

Report the values you obtain for the constant A, the initial 1958 CO_2 value B and the CO_2 growth rate C, with their uncertainties. What is the standard deviation of the residual variance? How does this compare with the value obtained for the quadratic-polynomial solution in Problem 1?

(d) Can we extrapolate an exponential-growth model back to the beginning of the Industrial Revolution? The first term of the nonlinear-regression model can be identified with the pre-industrial levels of CO_2, which are estimated to have been 260-270 ppm. How many standard deviations is this range from your value for A?

(e) What is your model's prediction for the year 1788, at the start of the French Revolution and at $t = 0$ of your model? What is the inferred growth rate of the anthropogenic CO_2 in percentage per year, and per decade?

Problem 3: Uncertainty Propagation in Linear Regression.

Assuming uniform uncorrelated uncertainties in the Mauna Loa CO_2 data, we can write the data variance as

$$\langle \delta\mathbf{d} \otimes \delta\mathbf{d}^T \rangle = \sigma^2 \mathbf{I}$$

where σ is the standard deviation of each data point and $\mathbf{I}$ is the identity matrix. This leads to the formula for the propagation of uncertainty in the least-squares regression problem $\mathbf{G}\cdot\mathbf{m} = \mathbf{d}$ that involves the $K \times K$ inner-product matrix of data representers:

$$\langle \delta\mathbf{m} \otimes \delta\mathbf{m}^T \rangle = \sigma^2 \left(\mathbf{G}^T \cdot \mathbf{G} \right)^{-1}$$

This formula can be decomposed into an expression that involves the principal components of model space, via the singular-value decomposition (SVD) of $\mathbf{G}$, that is,

$$\mathbf{G} = \sum_{k=1}^{K} \lambda_k \mathbf{u}_k \otimes \mathbf{v}_k^T$$

where the $\{\lambda_k\}$ are the singular values of $\mathbf{G}$, ordered so that $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_K \geq 0$. The $\{\mathbf{u}_k\}$ are the left-singular vectors (principal components in data space), and the $\{\mathbf{v}_k\}$ are the right-singular vectors (principal components in model space). Plugging the SVD into the uncertainty formula,

$$\langle \delta\mathbf{m} \otimes \delta\mathbf{m}^T \rangle = \sigma^2 \left(\mathbf{G}^T \cdot \mathbf{G} \right)^{-1} = \sum_{k=1}^{K} \frac{\sigma^2}{\lambda_k^2} \mathbf{v}_k \otimes \mathbf{v}_k^T$$

after some algebra that we discussed in class.

(a) Consider the linear regression for the Mauna Loa CO_2 time series as posed in problem 2(d) of Problem Set 6. Construct the matrix $\mathbf{G}$ for a quadratic polynomial in time, plus cyclic annual and semi-annual components. In order to scale the data representers for `time, time2` to be comparable in amplitude with the cyclic data representers `ann_s`, etc, divide the time ordinate by 2000, that is, compute

```
data_representer <- (time-1990.0)/25.0
data_representer2 <- data_representer^2
constant <- rep(1.0,length(data_representer))
gmatrix <- cbind(constant,data_representer,data_representer2,
                                          ann_c,ann_s,ann2_c,ann2_s)
```

Compute the SVD of `gmatrix` and plot the $k = 1, 2, \ldots 7$ singular values against k. What is the condition number of `gmatrix`, that is, the ratio of its largest and smallest singular values?

(b) Print out the principal component $\mathbf{v}_1$ of model space with the largest singular value. Print out the principal component $\mathbf{v}_K$ of model space with the smallest singular value. How are they related?

(c) Compute the singular value decomposition for the regression in part (a) for the case where the time ordinate is NOT referenced to the midpoint of the data series. In other words, scale the data representers thusly:

```
data_representer <- time/2000.0
data_representer2 <- data_representer^2
```

Compute the SVD of this `gmatrix` and plot the $k = 1, 2, \ldots 7$ singular values against k. How do they differ from part (a)? What is the condition number of `gmatrix`, that is, the ratio

of its largest and smallest singular values? To understand the importance of this result, plot the data representers for the constant, linear-time, and quadratic-time against each other for time since 1958.

Problem 4: Bootstrap estimates of parameter uncertainty.

In this problem you will apply the bootstrap to the regressions against the Mauna Loa CO_2 data. The R language does not have a bootstrap function in its basic package, but there is a "boot" library that you can download that will perform the bootstrap on a linear regression. The key commands for this problem are in the Rmd-file template. The command

```
library(boot)
```

retrieves the bootstrap software from the internet. One guide for the library can be found at

```
https://www.statmethods.net/advstats/bootstrapping.html
```

Bootstrap usage involves first specifying a statistic of the regression, in this case the coefficients of the model parameters, which are returned in a vector

```
model <- function(formula,data,indices){
d <- data[indices,]
fit <- lm(formula,data=d)
return(coef(fit))
}
```

We then assemble a "data frame" in R that contains all the data and the data representers

```
co2matrix <- data.frame(data=co2,g1=constant,g2=data_representer,
              g3=data_representer2,g4=ann_c,g5=ann_s,g6=ann2_c,g7=ann2_s)
```

Here we bootstrap with 1000 replications, specifying the regression formula EXCEPT for the constant term `g1`, which R automatically includes.

```
results <- boot(data=co2matrix, statistic=model, R=1000, formula=data~g2+g3+g4+g5+g6+g7)
results
plot(results)
results$t0
```

Compute and report the bootstrap regression coefficients, and their standard deviations for the quadratic-time fit, with both annual and semi-annual cycles of CO_2.