Written Report – 6.419x Module 4

Name: ptoche

Long calculations and code snippets are in the appendix at the end of this report. I have also included the autograded section 4-3 for completeness.

The Mauna Loa CO_2 Concentration

In 1958, Charles David Keeling (1928-2005) from the Scripps Institution of Oceanography began recording carbon dioxide (CO_2) concentrations in the atmosphere at an observatory located at about 3,400 m altitude on the Mauna Loa Volcano on Hawaii Island. The location was chosen because it is not influenced by changing CO_2 levels due to the local vegetation and because prevailing wind patterns on this tropical island tend to bring well-mixed air to the site. While the recordings are made near a volcano (which tends to produce CO_2), wind patterns tend to blow the volcanic CO_2 away from the recording site. Air samples are taken several times a day, and concentrations have been observed using the same measuring method for over 60 years. In addition, samples are stored in flasks and periodically reanalyzed for calibration purposes. The observational study is now run by Ralph Keeling, Charles's son. The result is a data set with very few interruptions and very few inhomogeneities. It has been called the "most important data set in modern climate research."

The data set for this problem can be found in `CO2.csv`. *It provides the concentration of CO_2 recorded at Mauna Loa for each month starting March 1958. More description is provided in the data set file. We will be considering only the CO_2 concentration given in column 5. The goal of the problem is to fit the data and understand its variations. You will encounter missing data points; part of the exercise is to deal with them appropriately.*

Let C_i be the average CO_2 concentration in month i ($i = 1, 2, \ldots$, counting from March 1958). We will look for a description of the form:

$$C_i = F(t_i) + P_i + R_i$$

where:

- $F\colon t \mapsto F(t)$ *accounts for the long-term trend.*
- t_i *is time at the middle of the ith month, measured in fractions of years after Jan 15, 1958. Specifically, we take*

  $$t_i = \frac{i + 0.5}{12}, i = 0, 1, \ldots,$$

  where $i = 0$ corresponds to Jan, 1958, adding 0.5 is because the first measurement is halfway through the first month.
- P_i *is periodic in i with a fixed period, accounting for the seasonal pattern.*
- R_i *is the remaining residual that accounts for all other influences.*

The decomposition is meaningful only if the range of F is much larger than the amplitude of the P_i and this amplitude in turn is substantially larger than that of R_i.

Pre-Processing data

You may notice that there are some inhomogeneities in data and the CO_2 concentration at these points is recorded as -99.99. Before proceeding, we must clean the data. One simple way to do this is to drop all missing values from the table. For the purpose of the problems below, use this simple method of dropping all the missing values.

Other methods include forward filling–fill missing values with previous values, and interpolation.

Fitting a Linear Model (autograded)

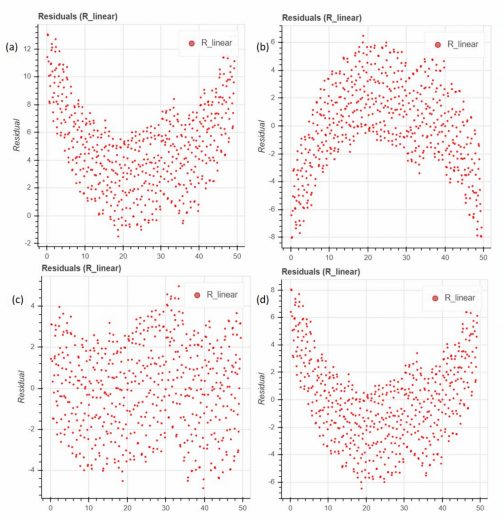
Now, fit the training data to a simple linear model. Plot the data and the fit.

1.

What are the values of $\hat{\alpha}_0$ and $\hat{\alpha}_1$?

2.

Plot the residual error $R_{linear}(t) = C(t) - F_1(t)$. Which of the following plots matches the residual plot generated?



3.

Report the root mean squared prediction error $RMSE$ and the mean absolute percentage error $MAPE$ with respect to the test set for this model.

Fitting a Quadratic Model (autograded)

In the problems below, we will fit quadratic and cubic models to the data and compute again $RMSE$ and $MAPE$. We will then evaluate which of these models is the lowest degree model that still captures the trend of the data best sufficiently.

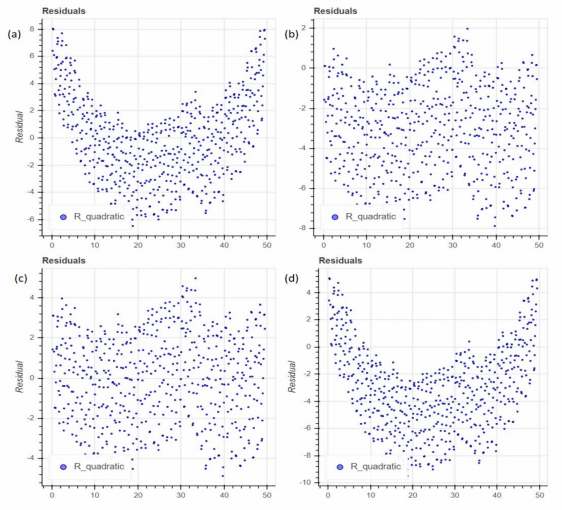
Now, fit the data to a quadratic model $F_2(t) \sim \beta_0 + \beta_1 t + \beta_2 t^2$.

1.

What are the values of $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_2$?

2.

Plot the residual error $R_{quadratic}(t) = C(t) - F_2(t)$. Which of the following plots matches the residual plot so generated?



3.

Report the root mean squared prediction error $RMSE$ and the mean absolute percentage error $MAPE$ with respect to the test set for the quadratic model.

Fitting a Cubic Model (autograded)

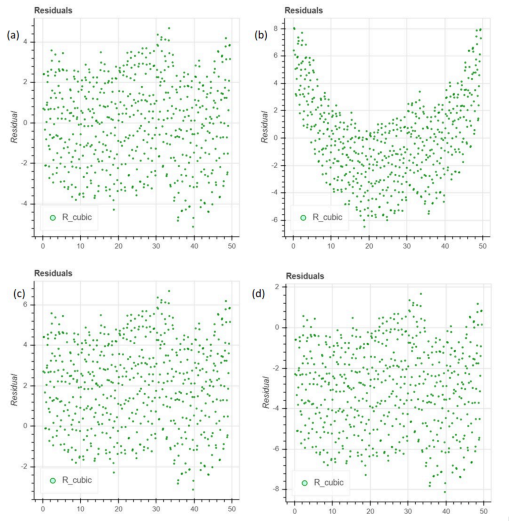
Now, fit the data to a cubic model $F_3(t) \sim \gamma_0 + \gamma_1 t + \gamma_2 t^2 + \gamma_3 t^3$.

1.

What are the values of $\hat{\gamma}_0$, $\hat{\gamma}_1$, $\hat{\gamma}_2$, and $\hat{\gamma}_3$?

2.

Plot the residual error $R_{cubic}(t) = C(t) - F_3(t)$. Which of the following plots matches the residual plot so generated?



3.

Report the root mean squared prediction error $RMSE$ and the mean absolute percentage error $MAPE$ with respect to the test set for the cubic model.

Model Selection (autograded)

Based on these plots, which of the following is true?

1. *R_{linear} shows a systematic concave upward trend, $R_{quadratic}$ does not have a clear systematic trend, and R_{cubic} closely resembles $R_{quadratic}$ in terms of both trend and magnitude.*
2. *R_{linear} shows a systematic concave downward trend, $R_{quadratic}$ does not have a clear systematic trend, and R_{cubic} resembles $R_{quadratic}$ in terms of trend but is shifted upwards in terms of magnitude.*
3. *R_{linear} shows a systematic concave downward trend, $R_{quadratic}$ shows a systematic concave upward trend while R_{cubic} does not show any clear systematic trend.*
4. *R_{linear}, $R_{quadratic}$, and R_{cubic} do not show a clear systematic trend and closely resemble each other in terms of trend and magnitude.*

Based on the residual plots drawn above and the prediction errors reported, what is the lowest degree polynomial that seems to be sufficient to represent the data?

1. *Linear Model (degree $= 1$).*
2. *Quadratic Model (degree $= 2$).*
3. *Cubic Model (degree $= 3$).*
4. *None of the above are sufficient.*

Fitting a Periodic Signal (autograded)

Consider $F_n(t)$ to be the polynomial trend chosen in the last problem as sufficient to represent the trend in the data. We will now extract the periodic component which appears in the data.

First, remove the deterministic trend $F_n(t)$ from the time series and compute the average residual $C_i - F_n(t_i)$ for each month. Namely, collect all the residuals (from removing terministic trend) for Jan (resp. Feb, Mar, etc) and average them to get one data point for Jan (resp. Feb, Mar, etc). The collection of these points can be interpolated to form a periodic signal P_i. Report the values of the periodic signal P_i for the month of January and February.

The final model (12 points)

1. (3 points)

Plot the periodic signal P_i. (Your plot should have 1 data point for each month, so 12 in total.) Clearly state the definition of P_i, and make sure your plot is clearly labeled.

2. (2 points)

Plot the final fit $F_n(t_i) + P_i$. Your plot should clearly show the final model on top of the entire time series, while indicating the split between the training and testing data.

3. (4 points)

Report the root mean squared prediction error $RMSE$ and the mean absolute percentage error $MAPE$ with respect to the test set for this final model. Is this an improvement over the previous model $F_n(t_i)$ without the periodic signal?

4. (3 points)

What is the ratio of the range of values of F to the amplitude of P_i and the ratio of the amplitude of P_i to the range of the residual R_i (from removing both the trend and the periodic signal)? Is this decomposition of the variation of the CO_2 concentration meaningful?

Autocovariance Functions

1. (4 points)

Consider the $MA(1)$ model,

$$X_t = W_t + \theta W_{t-1}$$

where $\{W_t\} \sim W \sim \mathcal{N}(0, \sigma^2)$. Find the autocovariance function of $\{X_t\}$. Include all important steps of your computations in your report.

2. (4 points)

Consider the $AR(1)$ model,

$$X_t = \phi X_{t-1} + W_t$$

where $\{W_t\} \sim W \sim \mathcal{N}(0, \sigma^2)$. Suppose $|\phi| < 1$. Find the autocovariance function of $\{X_t\}$. (You may use, without proving, the fact that $\{X_t\}$ is stationary if $|\phi| < 1$.) Include all important steps of your computations in your report.

CPI and BER Data Analysis (autograded)

The goal of this problem is to analyze the CPI and BER data for the last decade. The CPI (consumer price index, the price of a "market basket of consumer goods and services" - a proxy for inflation) is released monthly by the Bureau of Labor Statistics, and is given in `CPI.csv`*. The file* `T10YIE.csv` *lists (during most of the same time period) the break-even rate (BER) , or the difference in yield between a fixed rate and inflation adjusted 10 year treasury note. This difference can be interpreted as what the market views will be the inflation rate for the next 10 years, on average.*

There is more than a decade of data in `CPI.csv`*. For your results to the problems below, report the mean squared prediction error for 1 month ahead forecasts starting September 2013. For example, to predict the CPI in May 2015, you can use all the data before May 2015. You should perform all of your model fitting on the months prior to September 2013, and use the remaining months for evaluation.*

Detrend CPI

First, we will try to predict the monthly CPI without using the BER.

Plot the monthly CPI value as a time series (e.g. take the first CPI value of each month as that month's CPI). What is the most obvious trend that you can observe in the data?

We will now detrend this data into $CPI_t = T_t + R_t$. Below, we fit a linear trend to the training data. (You could also experiment with a different way of detrending.)

1.

Fit a linear trend T_t to the training data:

$$T_t = \alpha_0 + \alpha_1 t$$

(Recall the training data is to be set to all months prior to and not including September 2013. Use the remaining months for evaluation later.) Verify your result by plotting the linear trend on top of the data. Report $\hat{\alpha}_0$ and $\hat{\alpha}_1$ up to 3 significant digits.

2.

Subtract the linear trend from the data to get the residuals R_t. Visualize R_t. Do you see obvious further trend or seasonality in R_t? Report the maximum absolute value of the residuals over the training data.

AR Model: Determine the Lag

As there seems to be no other clear trend present in this residual, we can take the linear trend to be sufficient and move to the next step, i.e. to deseasonalize the data. However, since the visualization of the residual also seems to indicate no clear seasonality, we directly proceed to fitting an AR Model on the residual R_t.

Determine the order p of the AR model by examining the auto-correlation and partial auto-correlation functions of the residuals. Start with a single AR term and add other terms as necessary.

AR Model: Find the Parameters

Using the lag p from the previous problem, fit an AR(p) model to the training data. Recall from the lecture the AR(p) model:

$$X_t = \phi_1 X_{t-1} + \ldots + \phi_p X_{t-p} + \epsilon_t$$

Report the coefficients $\hat{\phi}_1$ and $\hat{\phi}_2$. Plot the final model predictions and the actual observed data together to visualize the fit.

Mean Squared Error

We will now evaluate our model using the mean squared prediction error.

The final model is the sum of the linear trend T_t and the AR(p), with all parametered determined using only the training data (i.e. earlier than September 2013).

Using this final model, compute the 1-month-ahead forecasts for the validation data (i.e. starting September 2013). Recall that the 1-month-ahead forecasts is the prediction that uses all the data before that month. For example, to predict the CPI in May 2015, use all the data before May 2015. (The model itself, however, is determined only using the training data.)

Report the root mean squared prediction error $RMSE$ for 1-month-ahead forecasts for the validation data starting September 2013. (Recall that the 1-month-ahead forecasts is the prediction that uses all the data before that month. For example, to predict the CPI in May 2015, you can use all the data before May 2015.)

Converting to Inflation Rates

Inflation Rate from CPI

Note that CPI is not a rate by itself while BER is. In order to make the different data sets comparable and use them together, we must first transform the data. One way to do this is to calculate monthly inflation rates from the different data sets given to us:

How might you calculate monthly inflation rates from the CPI data?

The monthly inflation rate can be calculated as the percentage change of CPI per month:

$$IR_t = \frac{CPI_t - CPI_{t-1}}{CPI_{t-1}}$$

where t indexes the months.

Plot the monthly inflation rate calculated by the formula above and report below the value for February, 2013.

Another way to get the monthly inflation rate is to compute the difference of the logarithms:

$$IR_t = \ln(CPI_t) - \ln(CPI_{t-1})$$

where t indexes the months. Plot the monthly inflation rate calculated by the formula above and report below the value for February, 2013.

1. (9 points)

Repeat the model fitting and evaluation procedure from the previous page for the monthly inflation rate computed from CPI. Your response should include:

- ***(1 point)** Description of how you compute the monthly inflation rate from CPI and a plot of the monthly inflation rate. (You may choose to work with log of the CPI.)*
- ***(2 points)** Description of how the data has been detrended and a plot of the detrended data.*
- ***(3 points)** Statement of and justification for the chosen AR(p) model. Include plots and reasoning.*
- ***(3 points)** Description of the final model; computation and plots of the 1 month-ahead forecasts for the validation data. In your plot, overlay predictions on top of the data.*

2. (3 points)

Which AR(p) model gives the best predictions? Include a plot of the RMSE against different lags p for the model.

Inflation Rate from BER (3 points)

BER data is not only reported on a daily basis but the values are calculated over a 10 year period. In order to use this data to get the "monthly" inflation rate one must

- *Choose a representative value of BER for each month (e.g. the average or the value on the last day of the month);*
- *"Deannualize" the monthly representatives above to convert to the monthly inflation rate. This deannualization of BER can be done by using the formula*

$$BER_t^{monthly} = (BER_t^{yearly} + 1)^{1/12} - 1$$

where t indexes the months.

Use the average value of BER over each month as the monthly representative, and deannualize this value to find the monthly inflation rate. Report the monthly inflation rate for Feb, 2013.

Overlay your estimates of monthly inflation rates and plot them on the same graph to compare. (There should be 3 lines, one for each datasets, plus the prediction, over time from September 2013 onward.)

External Regressors and Model Improvements

External Regressors

Next, we will include monthly BER data as an external regressor to try to improve the predictions of inflation rate. Here we only consider to add one BER term in the AR(p) model of CPI inflation rate. In specific, we model the CPI inflation rate X_t by

$$X_t = \sum_{i=1}^{p} \phi_i X_{t-i} + \psi Y_{t-r} + W_t$$

where Y_t is the BER inflation rate at time t, $r \geq 0$ is the lag of BER rate w.r.t. CPI rate, and W_t is white noise.

1. (4 points)

Plot the cross correlation function between the CPI and BER inflation rate, by which find r, i.e., the lag between two inflation rates. (As only one external regressor term is involved in the model, we only consider the peak in the CCF plot.)

2. (3 points)

Fit a new AR model to the CPI inflation rate with these external regressors and the most appropriate lag. Report the coefficients.

3. (3 points)

Report the mean squared prediction error for 1 month ahead forecasts.

Improving Your Model. (5 points)

What other steps can you take to improve your model from part III? What is the smallest prediction error you can obtain? Describe the model that performs best. You might consider including MA terms, adding a seasonal AR term, or adding multiple daily values (or values from different months) of BER data as external regressors.