Stat ST485/685, Project 7
Due: Thursday, December 16

(101 points total) In this project, you are asked to find, analyze, and use an ARIMA(p, d, q) model for the data set so2.txt. This is data from monitoring atmospheric sulfur dioxide levels from *S. Mazumdar and N. Sussman, Relationships of Air Pollution to Health: Results from the Pittsburg Study, Arch. Env. Health., 38: 17-24, 1983.*

- You should carry out the analysis by following the steps below. Your solution should be numbered accordingly.
- Your solution will be graded for legibility and clarity. You should spend effort on presentation.
- Include your code at the end of your solution.
- The cover page will be graded for completeness.

1. **(3 points)** Complete the cover page.
2. **(15 points)** Part 1: Determination of d.

   You should consider a plot of the data and fit a cubic polynomial using least squares then compare relative sizes of the coefficients. Use the results to choose a d.

   Note that overdifferencing, or choosing d too high, will result in significant loss of points.

   Do not subtract the least squares polynomial from the data. You are using differencing to remove any trend..

   For your answer:

   (a) Present a plot of the original data.
   (b) Make observations about possible trends.
   (c) Report the results of the least squares polynomial fit and the relative sizes of coefficients.
   (d) Specify d and explain your choice.
   (e) For the chosen d, display a plot of the mean-centered differenced data.
3. **(44 points)** Part 2: Determination of p and q for the mean-centered differenced data.

   You should begin by using a plot of the sample acf/pacf to make observations about possible orders of dependency. Then, you must use MLE to fit an ARMA(p, q) model for **at least four** combinations of p and q. Compare the plots of the ARMA(p, q) model together with sample acf/pacf values, plots of the model residuals, and the aic or aicc values to choose p and q.

   Hint: The best model has $p > 1$ and $q > 1$. Some of the assigned points will depend on how close you get to the optimal values.

   For your answer:

   (a) Show the plot of the sample acf/pacf for the mean-centered differenced data.
   (b) Give observations on possible orders of dependency.
   (c) For the ARMA(p, q) estimated using MLE, show plots of the model acf/pacf values together with the sample acf/pacf and plots of the model residuals for **four** choices of p and q.

   Display plots for only four choices even if you try more. If you try more, display results for values that help justify your final choice.

(d) Give the aic or aicc values for each of the estimated models in (c).

(e) Specify the p and q values you choose and give the reason.

4. **(36 points)** Part 3: Use MLE to fit the ARMA(p, q) model for the chosen p and q and analyze the model.

   You have already displayed the original and mean-centered differenced data in 1. You are working with that data!

   For your answer,

   (a) Specify p, d, and q.

   (b) Give the estimated coefficients for the MLE fit.

   (c) Give the value of the AIC or AICC.

   (d) Plot the model and sample acf/pacf values together.

   (e) Use the plot from (d) to assess the quality of the model fit.

   (f) Plot the standardized model residuals.

   (g) Plot the sample acf/pacf for the standardized model residuals.

   (h) Assess the plots from (f) and (g) with respect to the hypothesis that the model residuals behave like iid noise.

   (i) Evaluate the Ljung-Box and McLeod-Li statistics and indicate if they support rejection of the hypothesis that the model residuals behave like iid noise.

   (j) Using (h) and (i), give a final assessment on the validity of the hypothesis that the model residuals behave like iid noise.

   (k) Use the results from (e) and (j) to give a summary evaluation about the quality of the fitted model.

   In 3., you compare the plots of model/sample acf/pacf and model residuals for different p and q to choose best values for p and q. In this question, you are asked to assess how well the model for the chosen p and q fits the data. The model corresponding to the best value of p and q may or may not be a good model!

5. **(3 points)** Part 4: Use the estimated model to make a forecast.

   For your answer,

   (a) Plot the data together with prediction of values for 10 time steps past the last time of the data together with the confidence bounds.