HW4 Project Problem C for Math 420

This project concerns simple linear regression-based methods for quantitative prediction of weather data. Weather forecasters use very sophisticated models to predict future weather from global observations made in the recent past. Here we will consider one type of observation, daily high temperatures, at a few locations in the U.S.

Choose a training data set (e.g., 1995–2003) and a separate test data set (e.g., 2005–2008) to use for the predictive models described below. Use the training data set to find the parameters of the model that yield the best predictions for the training data set in the least-squares sense, and compute the RMS (root-mean-square) prediction error for both the training data set and the test data set. As you add more terms to your model, the prediction error should always decrease for the training data set, but if it stops decreasing for the test data set, that is a sign that your model has gotten too complicated and you are "overfitting" the data. Another criterion for overfitting that you can calculate as you go, when you are selecting a set of p predictors from a larger fixed set of potential predictors, is Colin Mallows' C_p, a *penalized sum of squared errors* that you can read about by Googling "Mallows Cp". (It will also be discussed in class.)

1. First consider the high temperature T_n at a particular location on day n. A reasonable prediction for tomorrow's temperature T_{n+1} is today's temperature T_n; compute the RMS error in that prediction (i.e., the square root of the average of $(T_{n+1} - T_n)^2$ over n) as a baseline. Then try the linear model

   $$T_{n+1}^{(p)} = a + bT_n$$

   where $T_{n+1}^{(p)}$ represents the prediction for T_{n+1}. Try also polynomial models of increasing degree, and see when you reach a point of diminishing returns.

2. Another direction to go from the linear model above is to predict tomorrow's high temperature in terms of today's and yesterday's values:

   $$T_{n+1}^{(p)} = a + bT_n + cT_{n-1}.$$

   Try using the data from 2 days ago, 3 days ago, etc., again until you find diminishing returns.

3. Another potentially useful predictive piece of data is tomorrow's "normal" temperature based on averaged historical data. See if you can improve your predictions by building this quantity into your model.

4. Weather typically propagates from west to east, so to predict tomorrow's (time $n + 1$) temperature in one location, it may be useful to know today's (time n) temperature in a different location to the west. See if you can use data from other locations to improve your predictions. Do the data support the premise that measurements from locations to the west have more predictive power than data from locations to the east?

5. Predictors can be chosen from those suggested above, or you can build your own, for example by plotting residuals versus possible new predictors. (You might also find that a certain predictor is

useful only in certain seasons, or only when certain weather conditions are present. Try to find a way to use indicator functions for the specified conditions multiplied by your predictor to make your predictor more useful.) After assembling as many predictor variables as you like, make a case for your “best” predictive model for tomorrow’s high temperature at a given location using data that is available today. All of the predictive models you try should be linear combinations of the predictors you assemble. Your answer may depend on the location, but try also to draw general conclusions – for example, which terms are most important to include in your predictive model in all cases? Make sure that you can interpret and build a narrative explaining the terms in your model. Simplicity is also a virtue – if a complicated model yields only slightly smaller errors than a simpler model, it may be best to go with the simpler one. (Your test and training disipline, or use of C_p will also lead you to this conclusion.) Can you replicate your conclusions with other training/test data sets?

6. How does the accuracy of your models compare with professional weather forecasts?

Data Here are two possible sources of historical data:

```
http://cdiac.ornl.gov/epubs/ndp/ushcn/access.html
http://www.wunderground.com/history/
```

For the first site, select a state, click “MAP SITES”, click a station name, then click “Get Daily Data” in the balloon on the map. Go to the bottom of the new page that opens, and select the data you want to download. Stations with reasonably complete data in recent years include Livermore CA, Boulder CO, Rolla MO, Beltsville MD.

The second site has data collected at airports. Enter an airport or city name, click “Submit”, then in the new page click “Custom”, select a date range, and click “Go”. The data appears at the bottom of the resulting page; to download as text, click “Comma Delimited File”. Suggested airports are San Francisco, Denver, St. Louis, and College Park.

For most of the locations at which data is available, “normal” daily temperatures based on data from 1971–2000 are available here:

```
http://cdo.ncdc.noaa.gov/cgi-bin/climatenormals/
climatenormals.pl?directive=prod_select2&prodtype=CLIM84
```