Journal of Statistics Education

An Example of Using Linear Regression of Seasonal Weather Patterns to Enhance Undergraduate Learning

Teresa Jacobson
Josh James
Neil C. Schwertman
California State University, Chico

Journal of Statistics Education Volume 17, Number 2 (2009), www.amstat.org/publications/jse/v17n2/jacobson.html

---

Key Words: Collaborative learning; Class project; Data analysis.

Abstract

Group activities are an excellent way to enhance learning. When students are actively involved in a relevant project, understanding and retention are improved. The proposed activity introduces a timely and interesting project typical of the type encountered in statistical practice. Using the computer to successfully developing an appropriate model is a valuable educational experience that builds confidence.

I. Introduction

Whenever possible many teachers attempt to involve students in research or projects to enrich their educational environment. Such projects can serve as a capstone for highly motivated students, bringing together the theoretical and applied aspects of their studies. The purpose of this example project is to introduce statistics undergraduates to such topics as model selection, higher order polynomial regression and to demonstrate that basic statistics methods and computer software can be used to quite precisely define patterns in real data.

Weather data affords the opportunity to enhance the learning environment by providing interesting practical real world data for analysis. Driscoll (1988) had his meteorology students monitor the accuracy of the TV temperature forecasts in seven U.S. cities for six months to compare the accuracy to forecasts from the National Weather Service, noting little difference in the accuracies for temperatures. Reading (2004) developed a weather data activity for students with a focus on variations in monthly temperatures and rainfall. To encourage the use of real data in the classroom, the National Oceanographic and Atmospheric Administration (NOAA) provides interesting data and resources. Google "NOAA Education Resources".

Weather is so popular that there is now a television channel completely devoted to weather issues. Weather can have a substantial impact on the economy, especially agriculture but other areas as well. The 2005 hurricanes Katrina and Rita caused an extensive shutdown of oil and gas rigs in the Gulf of Mexico resulting in a significant spike in energy prices while the 2006 freeze in California caused hundreds of millions of dollars in losses to citrus growers and wide spread unemployment. Consequently, it is not surprising there has been significant scientific efforts in studying and predicting weather patterns, see for example, journals such as: *Bulletin of the American Meteorological Society, Journal of Meteorology, Journal of Applied Meteorology and Climatology* and *Weather and Forecasting*.

Of course in practice the meteorological study of weather patterns use rather complex models and techniques (see for example: Brenner(1986), Brunet, et.al. (2007), Gyakum(1986), Murphy (1998), Serra, et.al. (2001) and Stone and Weaver(2002)). A historic perspective of evolving weather prediction models initially developed for the National Meteorological Center by the Princeton University Institute of Advanced Study in 1932 is provided by Murphy (1998) and Shuman (1989). The purpose here, however, is not to advance meteorological practice but rather increase statistics undergraduates' appreciation of the usefulness of statistics methods and to provide practical experience using these methods and computers with real world data. A Google search of "Modeling Daily Temperatures" lists ten references that illustrate the methods currently in use.

While most of forecasting literature focuses on short term predictions of just a few days, the long term forecasts can still be useful. For example, planning weeks in advance an outdoor activity such as a hike, swim, or camping or perhaps even an outdoor wedding, the

expected maximum (or minimum) temperatures and 95% confidence limits on the range of daily temperatures could be helpful in determining the practicality of the activity. The minimum temperature and it's 95% lower bound would provide valuable information for determining when to put temperature sensitive plants in an outdoor garden.

Since weather conditions usually are quite localized, many of the studies in the scientific journals by necessity pertain to very limited locales. Like studies by Gyakum (1986), Brenner (1986) and Mass (1987)) the example project in this paper is quite localized. Similar to Driscoll (1988) and Reading (2004), this example project for statistics undergraduates is a study of the daily temperature patterns throughout the year based on the thirty year temperature averages for Chico, California U.S.A. This type of project is easily adaptable as a class activity even at the elementary level.

This activity was motivated on one particularly hot day by a common keen interest among the authors of the daily temperature trends. For background, Chico, located in the Sacramento River valley and the northern end of the California "central valley" gets very hot in the summer. Temperatures above 100F are very common but the winters are mild, rarely below 30F or -1C. To establish focus, by consensus with the students, specific questions were addressed. The weather related questions investigated in the project are:

1. What day of the year on average has the highest maximum temperature and what is that temperature?
2. What day of the year on average has the lowest maximum temperature and what is that temperature?
3. What day of the year on average has the highest minimum temperature and what is that temperature?
4. What day of the year on average has the lowest minimum temperature and what is that temperature?
5. Is the daily trend in temperatures the same for the maximum and minimum temperatures?
6. Are the random variations over thirty years in each day's maximum temperature roughly the same for each day throughout the year?
7. Are the random variations over thirty years in each day's minimum temperature roughly the same for each day throughout the year?

The statistical questions addressed are:

1. Can the daily patterns in temperature be modeled adequately using polynomial regression models with day of the year as the independent variable?
2. What criteria should be used to evaluate the adequacy of the model?

To answer these questions the steps were:

1. To access the data as described in Section II.
2. To plot the data using the JMP statistical program.
3. To analyze data using the regression analysis in JMP with successive higher order polynomial models until the changes in R-square and root mean square error (RMSE) are minimal.

The two undergraduate seniors' (first two authors) statistical education consisted of three semesters of upper division calculus based mathematical statistics and two semesters of applied statistics/experimental design. The primary topics covered in applied statistics were analysis of variance and the general linear model with an introduction to SAS and JMP statistical programs. Other topics included the design of experiments using blocking to decrease noise and factorial designs, orthogonal and multiple comparison procedures to enhance information, repeated measures and analysis of covariance. While the analysis of data similar to the weather data could easily be used as project for a class, for these two students it was not a requirement or extra credit for any class. It was purely voluntary to enrich their learning experience. They were asked if they would like to do a research project. No prodding was needed to keep them on track which indicates that the students were enthusiastic and enjoyed doing some actual research and found it rewarding and interesting.

The students had not been introduced to time series or non-linear modeling which could be used for a more statistically appropriate analysis of the weather data. The intent was to provide students with practice in the basic tools of research and modeling and to address the limitations of these methods.

The advisor on the project, suggested the topic, search the internet for the data and related literature. The students suggested and investigated all the models and wrote the initial version of the paper. They were asked to write the paper in a form for submission to a peer reviewed journal. This was their first attempt at such a task and the initial version was more about temperature patterns in Chico, California than about what they learned and how they benefitted from the project. Even with the limitations on statistical inference, the paper provides insight into the daily temperature patterns throughout the year in the California central valley. The subsequent revisions required more commentary from the teacher's prospective and the advisor became involved in the rewrite since both students had graduated.

The graphical displays of the data and the polynomial models used to answer the questions are provided in Figures 1 to 8 and the summary of the numerical part of the statistical analyses are in Tables I to IV. The Question 9), criteria for evaluating the models are discussed in Section III of this paper. Sections IV and V describe the analyses of the temperature and random variation data respectively. In section VI the nine questions are answered with some concluding comments.

Accessing weather data

The weather data for the western United States (excluding Hawaii and Alaska) is available at the web address: www.wrcc.dri.edu/climsum.html. This page is: "Western Regional Climate Center". Click on the desired state and city or location, then on "daily tabular data". There are 183 locations listed however many have no data. Other temperature data can be found for each state at the web address:

http://cdo.ncdc.noaa.gov/cgi-bin/climatenormals/climatenormals.pl. First click on "Product Selection", then click on "Daily Stations Normals". Next select the appropriate state and choose the desired city.

Using the daily temperatures for thirty years would be a massive data set and for just modeling the patterns, the daily averages are sufficient. Furthermore, there may be gradual climate change which could invalidate any analysis used for prediction. The purpose of this activity, however, was to introduce statistical methodology and use of statistical programs rather than temperature prediction.

3. Criteria for evaluating the models

Three criteria were initially considered for choosing the "best" model: R-square, Root Mean Square Error (RMSE) and the p-values of the polynomial coefficients. A lively discussion of the relative merit of each concluded that R-square and RMSE are quite similar and are best if the purpose of the model is prediction and that the p-value is better suited for testing separately each term in the polynomial. The p-value criterion was significant at $p<.05$ for all parameters in all models considered for both maximum and minimum temperatures as well as all models in the analyses of the variations or standard deviations. In fact, all the final models had all p-values <.0001. Consequently that criterion was deemed to be ineffective in discriminating the relative adequacy of the models and only the R-square and RMSE criteria were used to measure the relative quality of the competing models.

There are other methods for determining models such as SELECT, PRESS and the stepwise, forward selection, and backward elimination regression procedures but these were considered too advanced for undergraduates.

4. Determining the Models for Maximum and Minimum Daily Temperatures

The data provided from the website of the Western Regional Climate Center were the mean maximum and mean minimum temperatures for each day of the year over the 30 year period 1971 to 2000. However in some cases observations for a few years were missing in the averages. That daily temperatures on consecutive days are likely to be correlated since there are frequent "hot spells" and "cold spells" presents a challenge. Such data often are analyzed using time series techniques. Because of the student's background, a time series analysis was not practical. Of necessity the focus became to introduce undergraduates to modeling using regression methods and the correlations were ignored for purposes of illustration. A challenging task of a more advanced nature would be a time series analysis of weather data.

The average maximum daily temperatures throughout the year are displayed in Figure 1 and the individual 95% confidence limits for the range of the daily maximum temperature are displayed in Figure 2. To find models which closely approximated the plots, simple polynomial regression models were evaluated using both SAS with the PROC REG procedure and JMP. Using the day of the year as the independent variable, starting with just the quadratic model the order of the polynomial models was increased until an adequate fit was obtained. While SAS has many more options it is more difficult to use. Since the analyses were identical for both SAS and JMP the simpler JMP was used for evaluating all the remaining polynomial models. The R-square and RMSE for the maximum daily temperature for the various polynomial models are provided in Table I with the plots of the various polynomial models in Figure 3.

Table I indicates very little improvement in the model using a polynomial higher than degree 5. The next higher polynomial model only increased R-square by 0.00009 and reduced RMSE by only 0.014723. Therefore the fifth degree polynomial model was considered to adequately fit the data with R-square of 0.998252 and RMSE of 0.594392.

While the objective of the project was to introduce polynomial regression modeling for real data, one non-linear model was investigated. One student observed the data plot in Figure 1 seemed to follow a Beta type distribution with (day/366) on the horizontal axis. The Beta model was temperature, $T(x)=A\left(\frac{x}{366}\right)^{\alpha}\left(1-\frac{x}{366}\right)^{\beta}+C$ where x is the day of the year starting Jan. 1. To estimate these constants for Chico, California, the maximum average high temperature occurs on both July 27 and 28, the 209th and 210th day of the year and x = 209.5 was used. Taking the derivative of T(x), $T'(x)=0$ and solving the ratio $\frac{\beta}{\alpha}=\frac{366-x}{x}=.747$. Using this ratio and the minimum highest average daily temperature as the initial value for C (C= 53.2), an iterative process was used to minimize the sum of all the squared differences between the actual mean high temperature and the T(x) for all 366 days. The model $T(x)=864.1\left(\frac{x}{366}\right)^{2.595}\left(\frac{1-x}{366}\right)^{1.925}+54.71$

seems to provide a very good fit (see Figure 4). While the R-square for this non-linear model is 0.989 and the root mean square error is 1.456 the polynomial linear model provided a better fit to the data.

Figure 5 is the plot of the average minimum daily temperatures with the individual 95% confidence limits on the range of the daily minimum temperatures. The average minimum daily temperature plot, as would be expected, followed a similar pattern to the maximum temperature plot. The R-squares and RMSE for the various polynomial models are included in Table II and the plots of these models are provided in Figure 6. The inclusion of the sixth degree term increased R-square by 0.003323 to 0.996717 and RMSE was reduced by 0.213329 to 0.511967. This model was considered to be quite adequate fit of the data. Due to the similarity in the patterns for maximum and minimum temperatures it would be reasonable to reconsider an order six polynomial for maximum temperature but this was not done.

Since the nonlinear Beta type model was less satisfactory than the higher degree polynomial models and due to time constraints, the nonlinear model was not investigated for the average daily minimum temperature data. The non-linear model requires the estimation of four parameters while the polynomial models requires the estimation of six or seven parameters. The small increase in the number of parameters for the linear models is well justified by the improvement in the model's fit as well as the simplicity of the linear models compared to the non-linear approach.

5. Modeling the Variations in Maximum and Minimum Daily Temperatures

The Western Regional Climate Center website also provides, for each day, the standard deviations of the maximum and minimum daily temperatures over the same approximate thirty year period. The standard deviations of the maximum temperatures follow a particularly complex pattern. The "transition periods" from winter to summer and from summer to winter had the largest daily variation in maximum daily temperatures. Throughout the entire periods from March 23 to June 23 and from September 19 to October 19, and only in these periods were the standard deviations at least 8.000. On May 19 the maximum of all the estimated standard deviation was 9.571 while a second local maximum of 8.807 occurred on September 30.

Based on the students' knowledge of polynomials and the graphical display of the daily standard deviations in maximum temperatures it was estimated that a model of at least degree seven would be necessary to satisfactorily describe the pattern. Figure 7 shows the fit of the polynomial models. The sixth order polynomial was used to compare the erratic pattern of the maximum standard deviations to the corresponding pattern of the best fitting polynomial for the minimum standard deviations. Table III provides the R-squares and RMSE for the polynomial models of degrees 6, 7 and 8. Including the eighth degree term increased R-square by 0.020975 to 0.885433 and reduced RMSE by 0.028903 to 0.33539. For such an erratic pattern this seemed quite reasonable.

The basic underlying trend in the daily variations as measured by the standard deviations of the daily minimum temperatures had a much less erratic pattern, being greatest during the winter and more stable or consistent during the late summer to October. During the period from August 3 to October 13 the standard deviations in the minimum daily temperature were consistently less than 5.0 degrees with 4.570 the minimum of all the estimated standard deviations occurring on August 30. From November 1 to March 23 the standard deviations in minimum daily temperatures was consistently greater than 6.0 with the largest standard deviation in the data, 7.490 occurring on January 3. For model comparison, Table IV provides the R-square and RMSE for the polynomial models of degrees two to six. Figure 8 is the plot of the standard deviations of the daily minimum temperatures with the graphs of the polynomial models. By including the sixth degree term the R-square only increased by 0.024551 to 0.977235 but the RMSE decreased substantially by 0.057555 to 0.13090. Hence the sixth degree model was considered the best.

6. Answering the questions and concluding comments.

The study provided answers to the nine questions in the Introduction that were used to focus the project. The first four weather questions were answered immediately upon accessing the data and the other weather questions were easily answered from the figures. Specifically, for Chico California

1. The highest average maximum temperature of 95.1F occurs on July 27 and 28.
2. The lowest average maximum temperature of 53.0F or occurs on January 2 and 3.
3. The average highest minimum temperature of 60.8F occurs during the period July 21 to 28.
4. The average lowest minimum temperature of 33.6F occurs on December 27 and 28.
5. The pattern in the daily maximum and minimum temperatures is roughly the same with the highest maximum and minimum daily occurring on the same days and the lowest maximum and minimum temperatures only six days apart.
6. Figure 7 clearly shows that the basic pattern of random variation in maximum temperature varies greatly during the year, being much larger during the transition times, Spring and Fall.
7. Figure 8 show that the basic pattern of random variation in minimum daily temperature is much less erratic and much more consistent during the summer.
8. With R-squares greater than 0.99 and root mean square errors (RMSE) just about .5 the polynomial regression was able to model both maximum and minimum temperature patterns very well.
9. The three criteria used to evaluate the models are described in Section 3 and were in complete agreement with each other. The two used to evaluate the differences between models were R-square and RMSE.

While the temperature questions are interesting the statistical questions 8) and 9) were much more important from the educational prospective. The primarily purpose of this paper was not to analyze the temperature pattern of Chico, California but rather to demonstrate an example project and the vast analysis potential of readily available weather data. Projects such as this one, afford ample opportunities to discuss with students many of the vexing questions and problems that can occur while modeling real data. One of the most obvious is the time scale. Fortunately for this data both the minimum high and low daily temperatures occurs very close to the first day of the year (January 2 and December 28) and the day of the year was a natural time scale. In other data the determination of a time scale can have significant ramifications which should be discussed. For data based on a yearly cycle it is important that the model provide values close together at the end and beginning of the time scale. For the weather example in this paper, days one and 365 are really only one day apart and should have nearly identical average temperatures. Students could be asked how to address this requirement. With some thought students may suggest, due to the cyclic nature of yearly data, a trigonometric model to ensure a smooth transition from the end to the beginning of the time scale. In this project, simple trigonometric models were considered initially but not pursued due to the complexity and time constraints. Furthermore, the simpler the higher order polynomial models seemed to adequately describe the patterns.

A second area of discussion is the importance of independence of the data. It was assumed there would be dependencies caused by hot or cold spells that last for several days. It is essential to point out that the lack of independence in this data precludes statistical inference with the methods used. A general discussion about handling dependencies such a paired data analysis or the Geisser-Greenhouse correction factor should be mentioned as possibilities. Neither of these could be used for this project since the thirty daily temperatures for each day of the year was not available. It is instructive to point out to the students that more complex techniques such as time series and multivariate analysis are available and would allow statistical inference from the data.

The Editor suggested that since the maximum and minimum daily temperature patterns are closely related raises the question: Should the same degree polynomial be used for both? Specifically, students could discuss how much should the similarity in patterns and intuition influence the modeling process? How important is it that the model make sense from a practical standpoint? It could be pointed out that the efficacy of using the same degree polynomial for both maximum and minimum temperature models could be tested by the classical linear model technique of finding the total regression sum of squares (RSS) for the full model (allowing different degree polynomials for each) and a restricted RSS by using the same degree polynomial for each. The difference between the two RSS's divided by the appropriate degrees of freedom and the MSE results in a statistic which is a pseudo F because of the dependencies. This statistic, nevertheless, provides some measure of the efficacy of using polynomials of the same order for both maximum and minimum temperature models.

Upon graduating, statistical students, whether in graduate school or in industry, are likely to need to model and analyze data. Projects like this afford practice in using statistical programs such as JMP and should aid in the understanding and appreciation of statistical methodology and the limitations. The students were enthusiastic and particularly enjoyed analyzing local data that had meaning to them. The modeling of real data can enhance confidence and promote statistical maturity.

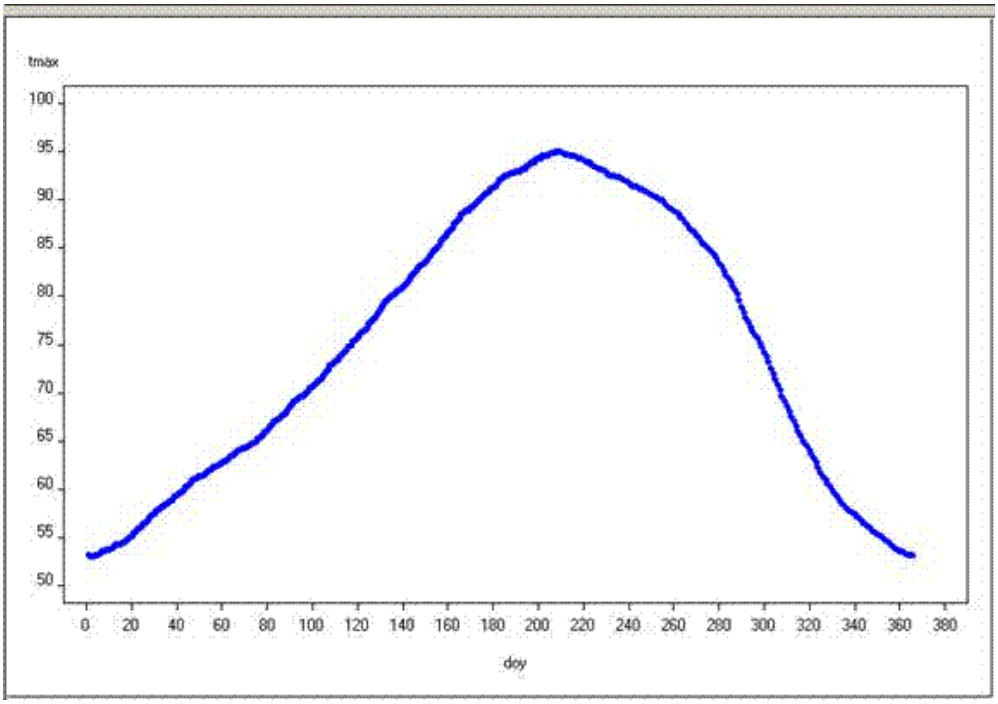


Figure 1 Daily Maximum Temperatures

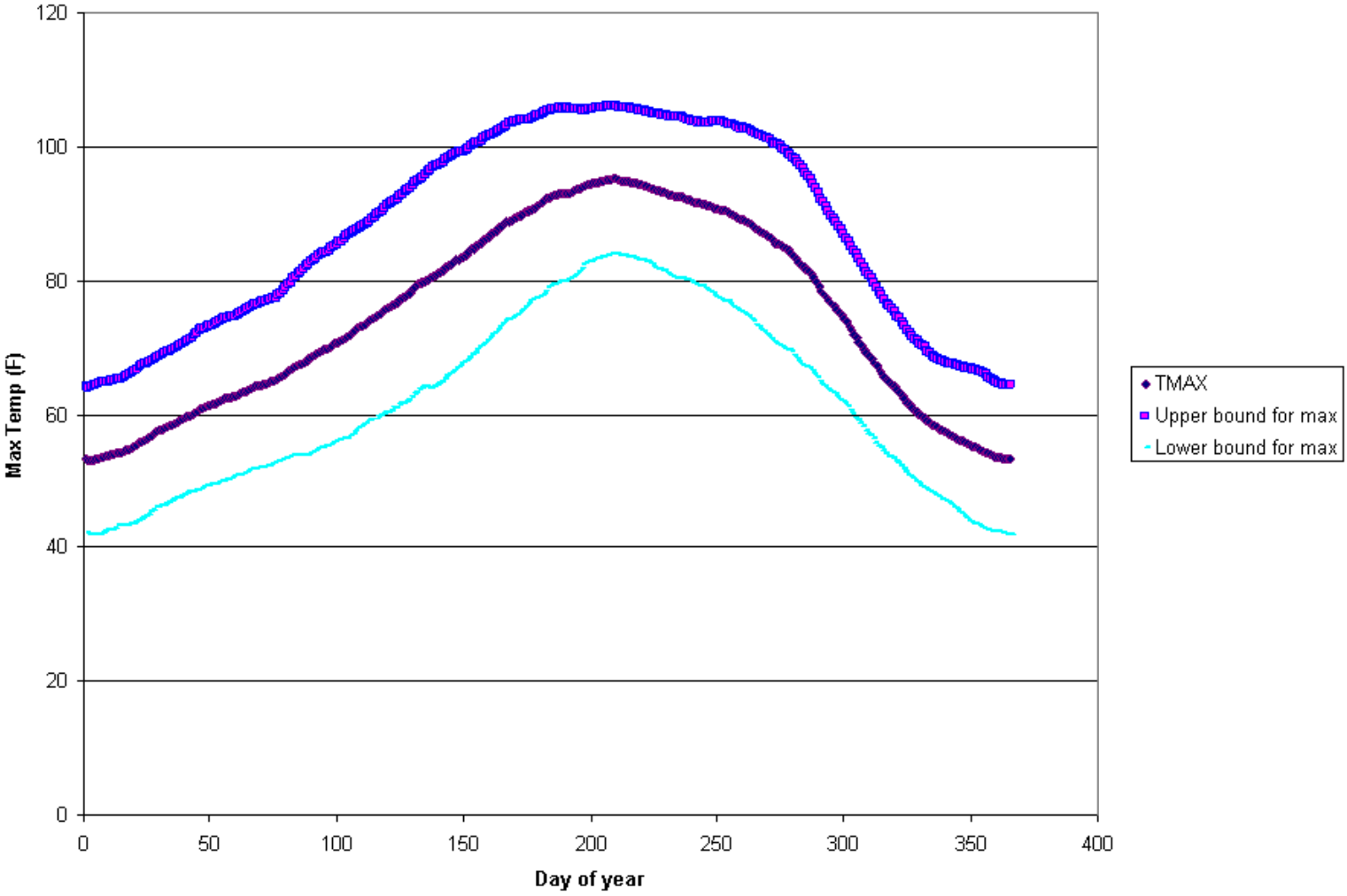
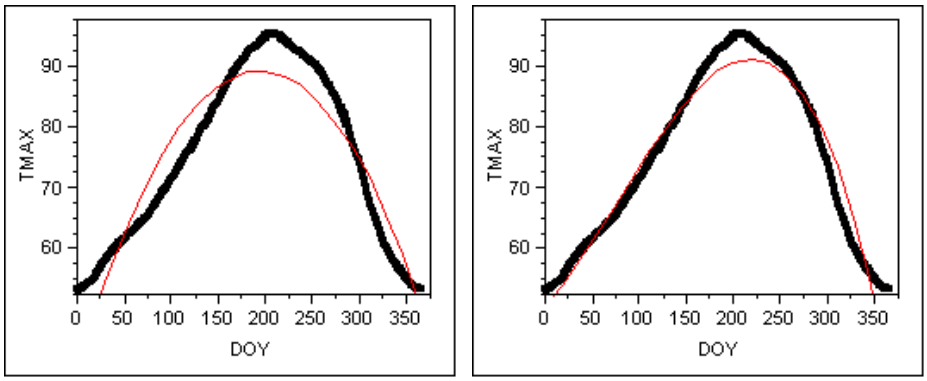
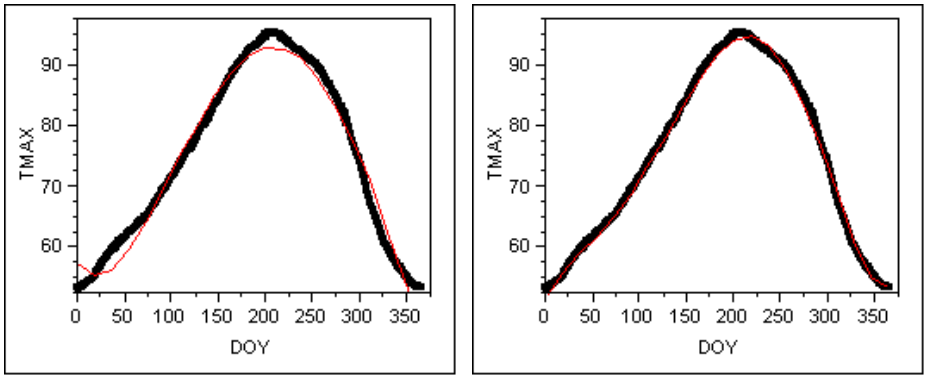


Figure 2 Daily Maximum Temperatures with Confidence Limits



Tmax Quadratic Model **Tmax Cubic Model**



Tmax Quartic Model **Tmax Quintic Model**

Figure 3 Maximum Temperature Models

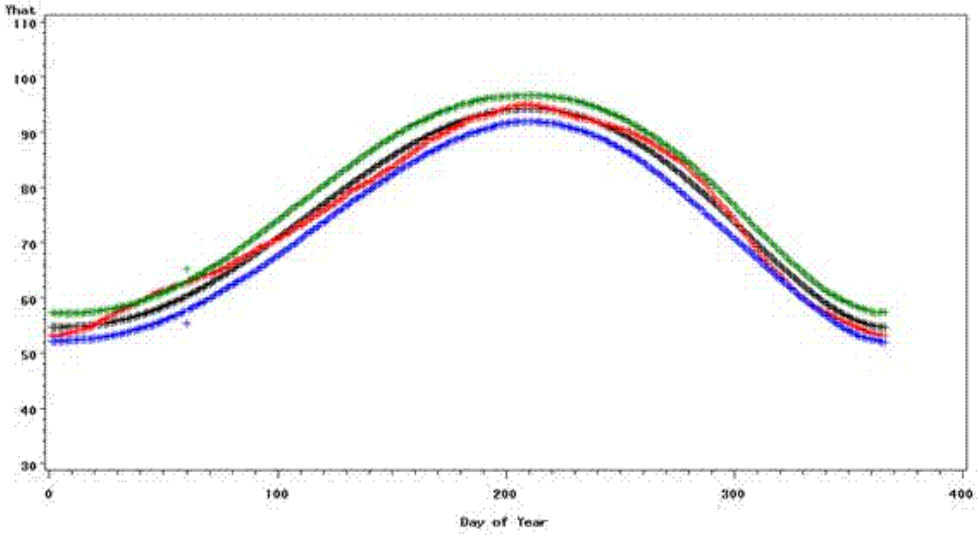


Figure 4 Tmax Beta Model with Confidence Limits

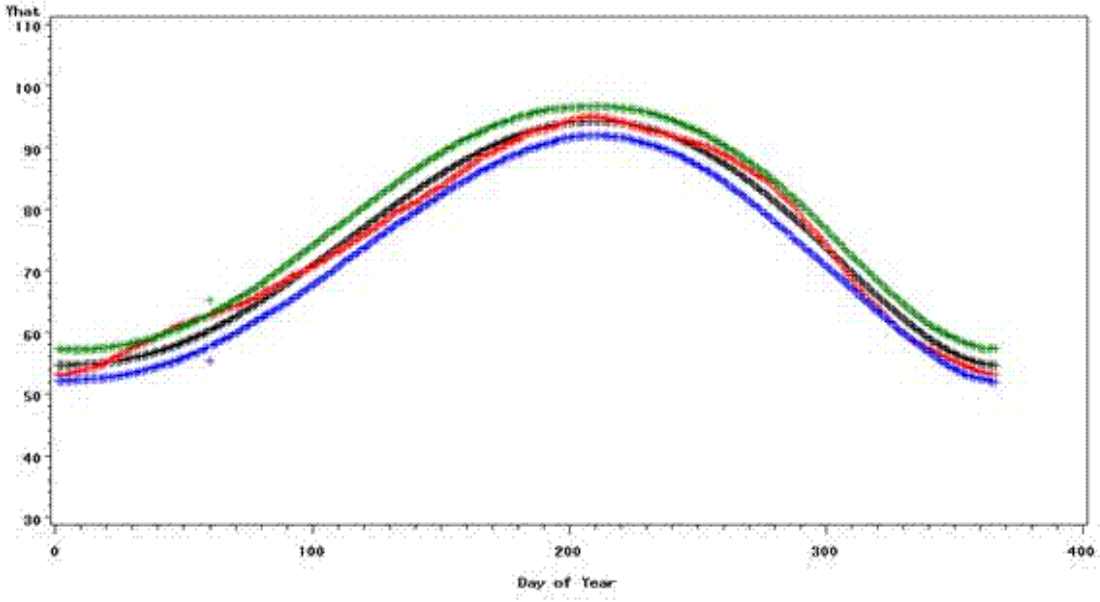
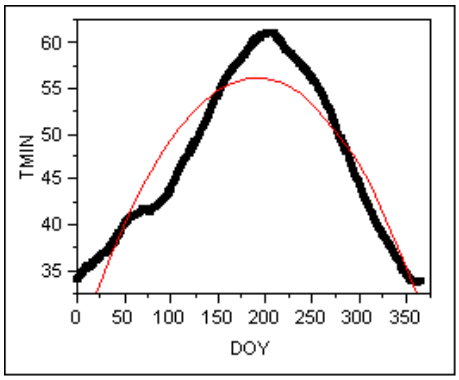
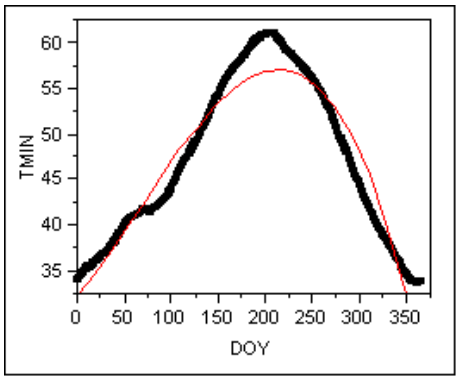


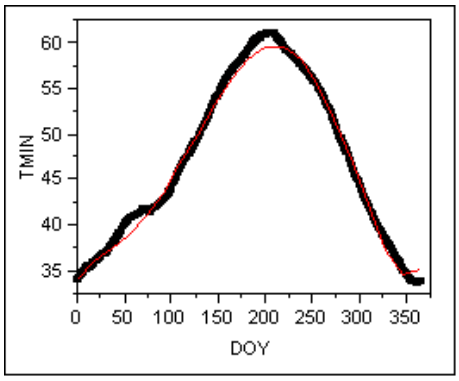
Figure 5 Tmin Daily Temperatures with Confidence Limits



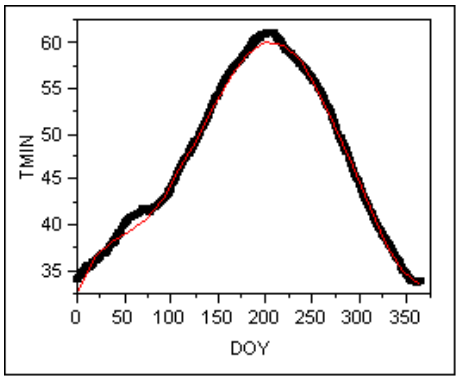
Tmin Quadratic Model



Tmin Cubic Model

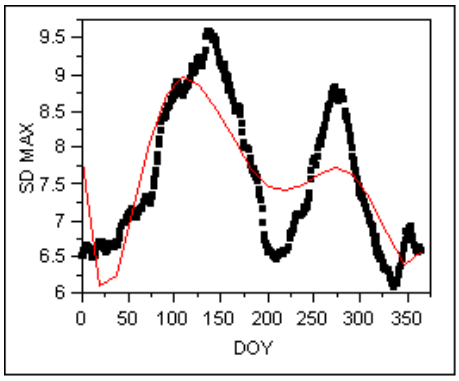


Tmin Quintic Model

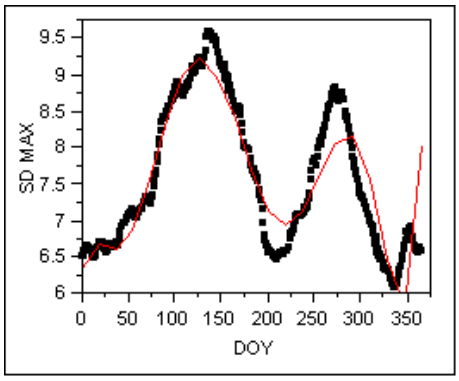


Tmin Sextic Model

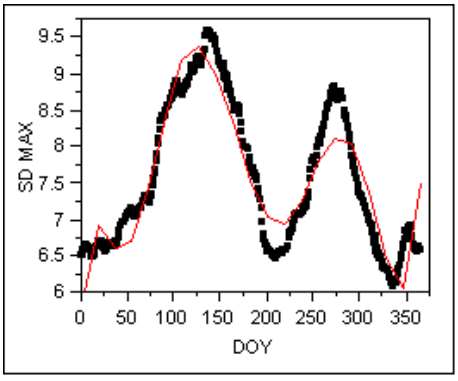
Figure 6 Minimum Temperature Models



Daily Standard Deviations Maximum Temperatures Sextic Model



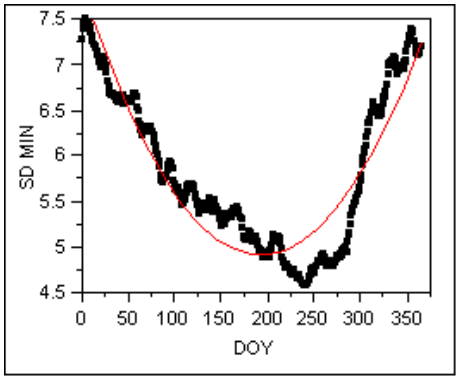
Daily Standard Deviations Maximum Temperatures Septic Model



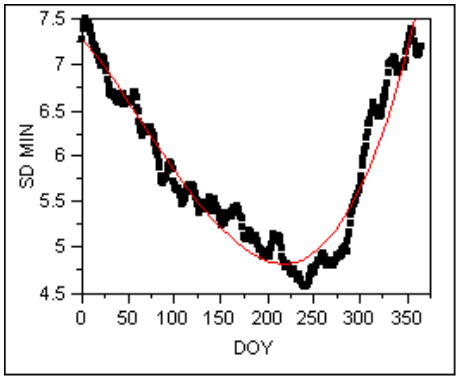
Daily Standard Deviations Maximum

Temperatures Octic Model

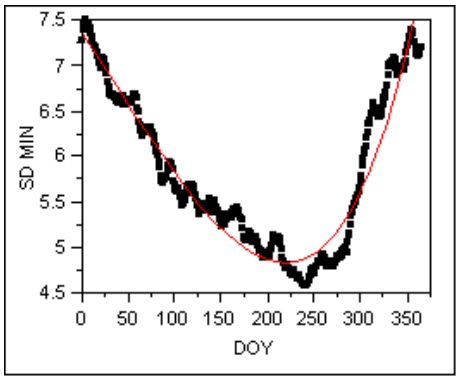
Figure 7 Models of Standard Deviation of Daily Maximum Temperatures



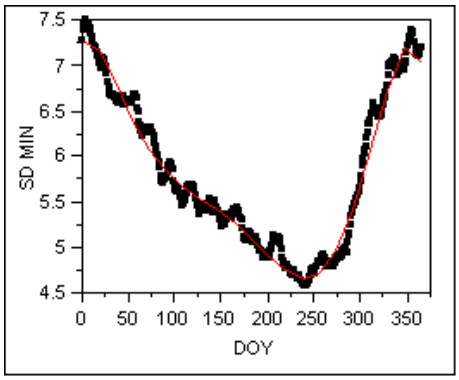
Daily Standard Deviations Minimum Temperatures Quadratic Model



Daily Standard Deviations Minimum Temperatures Cubic Model



Daily Standard Deviations Minimum Temperatures Quartic Model



Daily Standard Deviations Minimum Temperatures Sextic Model

Figure 8 Models of Standard Deviation Minimum Temperature Models

Table I
Maximum Temperature

Model	R-Square	Root Mean Square
Quadratic	0.88407	4.820916
Cubic	0.957602	2.919457
Quartic	0.982139	1.897503
Quintic	0.998252	0.594392
Sextic	0.998342	0.579669

Table II
Minimum Temperature

Model	R-Square	Root Mean Square
Quadratic	0.868455	3.223
Cubic	0.926276	2.4162
Quartic	0.975827	1.3855
Quintic	0.993394	0.7253
Sextic	0.996717	0.512

Table III
Standard Deviations of Maximum Temperatures

Model	R-Square	Root Mean Square
Sextic	0.690793	0.549455
Septic	0.864458	0.364293
Octic	0.885433	0.33539

Table IV
Standard Deviations of Minimum Temperatures

Model	R-Square	Root Mean Square
Quadratic	0.869937	0.311158
Cubic	0.931936	0.225404
Quartic	0.932711	0.224427
Quintic	0.952684	0.188456
Sextic	0.977235	0.130901

References

Brenner I.S. (1986). "Biases in MOS (Model Output Statistics) Forecasts of Maximum and Minimum Temperatures in Phoenix, Arizona", Weather and Forecasting, Vol. 1(3), 226-229.

Brunet M., Sigro J., Jones P.D., Saladie O., Aguilar, Moberg A., Della-Marta P.M., Lister D., Walter A. (2007). "Annual and Seasonal changes in the distribution of daily maximum and minimum temperature data in temperature extreme indices thoughout the 1901-2005 period over mainland Spain.", Geophysical Research Abstracts 9, 07167.

Driscoll D.M. (1988). "A Comparison of Temperature and Precipitation forecasts issued by Telecasters and National Weather Service", Weather and Forecasting, Vol.3(4), 285-295.

Gyakum J.R. (1986). "Experiments in Temperature and Precipitation Forecasting in Illinois", Weather and Forecasting, Vol. 1(1), 77-88.

Mass C.F. (1987). "The 'Banana Belt' of Coastal Regions in Southern Oregon and Northern California", Weather Forecasting, Vol. 2(3), 253-258.

Murphy A.H. (1998). "The Early History of Probability Forecasts: Extensions and Classifications", Weather and Forecasting, Vol. 13(1), 5-15.

Reading C. (2004). "Student Description of Variation While Working with Weather Data", Statistics Education Research Journal, Vol. 3(2), 84-105.

<>

Serra C. , Burgueno A. , Lana X. (2001). "Analysis of Maximum and Minimum daily temperatures recorded at Fabra Observatory Barcelona NE Spain in the period 1917-1998", International Journal of Climatology, Vol. 21, 617-636.

Shuman (1989). "History of numerical weather prediction at the National Meteorological Center", Weather and Forecasting, Vol. 4, 286-296.

Stone, Weaver (2002), "Daily maximum and minimum temperature trends in a climate model", Geophysical Research Letters, Vol. 29(9),

70-71.

Teresa Jacobson
California State University, Chico
Email: kalany@gmail.com

Josh James
California State University, Chico
Email: joshjames5@hotmail.com

Neil C. Schwertman
Professor Emeritus of Statistics
California State University, Chico
Email: NSchwertman@csuchico.edu
