NISS

# An Illustration of the Use of Generalized Linear Models to Measure Long-Term Trends in the Wet Deposition of Sulfate

Patricia E. Styer

Technical Report Number 18
August, 1994

National Institute of Statistical Sciences
19 T. W. Alexander Drive
PO Box 14006
Research Triangle Park, NC 27709-4006
www.niss.org

Although the information in this document has been funded wholly or in part by the United States Environmental Protection Agency under assistance agreement #CR819638-01-0 to the National Institute of Statistical Sciences, it may not necessarily reflect the views of the Agency and no official endorsement should be inferred.

# An Illustration of the Use of Generalized Linear Models To Measure Long-Term Trends in the Wet Deposition of Sulfate

**Patricia E. Styer** *

**National Institute of Statistical Sciences**

## Abstract

In this research, I present a method to measure long-term trends in the wet deposition of sulfate, adjusting for effects of season and meteorology. The methodology proposed incorporates the use of generalized linear models, specifically gamma regression models, which are a useful extension of previous efforts applying ordinary least squares regression models to precipitation monitoring data. Gamma regression models are appropriate for right-skewed, positive data and alleviate the problems introduced by fitting regression models to such data on a transformed scale. For the application presented here, the gamma regression models provide simple estimates of long-term trends in the wet deposition of sulfate. While these trend estimates are very similar to estimates produced by ordinary least squares regression models fitted to the log-transformed data, I discuss other applications where it is more advantageous to fit regression models on the untransformed scale.

**Key words:** gamma regression, environmental monitoring, lognormal distribution

---

*Supported by the EPA Cooperative Agreement CR 819638-01 and the National Science Foundation Grant DMS-9208758.

# 1 Introduction

In environmental monitoring, it is common to find data in the form of positive measurements where the variance increases with the expected value of the measurement. Researchers have noted that such data are conveniently analyzed by applying the logarithmic transformation and using statistical techniques that rely on the approximate Normality of the transformed variables.[1] In this paper, I propose a technique for analyzing positive, right-skewed, and continuous data in which the response variable is not transformed. I present an illustration using acid deposition monitoring data, and discuss the advantages of preserving the original scale of measurement. Specifically, I model the sulfate concentration in precipitation, taking into account covariates like precipitation amount, season, and local meteorological conditions.

This application was motivated by the need to assess recent trends in acid deposition and to develop methodology to evaluate the future impact of the Clean Air Act Amendments of 1990.[2] Since acid deposition is known to depend on meteorological conditions, regression-type models are useful for removing known sources of variability. Additionally, they can correct for bias if short periods of deposition monitoring overlap with confounding trends in weather. Other uses of these regression models include the comparison of trend estimates from precipitation samples collected on different time scales, and the estimation of total deposition at each monitoring site, using the regression model to fill in values when sulfate measurements are missing.

## 2 Model Formulation

I illustrate the application of a gamma regression model using sulfate concentration in precipitation as the response variable. To adjust for meteorological conditions, I include a few summaries of local weather conditions. The monitoring data used for this analysis were collected on roughly a daily basis during periods of precipitation, so I use daily-level measures of meteorology that are roughly concurrent with the precipitation activity. A reasonable model for daily sulfate concentration would then be that $S_i$, the sulfate deposition on day $i$, is gamma-distributed with density

$$f(s_i) = \frac{1}{\Gamma(\nu)} \left( \frac{\nu}{\mu_i} \right)^{\nu} s_i^{\nu-1} \exp\left( \frac{-\nu}{\mu_i} s_i \right).$$

This parameterization is standard in generalized linear models. Adopting common terminology distinguishing the shape and scale parameters, the shape parameter is $\nu$ and the scale parameter is $\mu_i \nu^{-1}$ under this parameterization. The value of $\nu$, the shape parameter, determines the amount of skewness. Smaller values of $\nu$ indicate less symmetric distributions, and, for larger $\nu$ the gamma distribution is similar to the Gaussian distribution.

The expected value for the sulfate deposition on day $i$ is

$$\mu_i = \mathrm{E}(S_i) = \exp(\mathbf{x}_i' \beta),$$

where $\mathbf{x}_i$ includes a seasonal factor, a component to permit the estimation of long-term trends, and daily-level meteorological covariates. The assumption that the $S_i$'s have constant shape parameter $\nu$ is equivalent to the condition that the coefficient of variation $\nu^{-1/2}$ is constant.[3] In other words, the ratio

of the standard deviation to the mean is constant, so the standard deviation is proportional to the mean. This assumption specifying how the standard deviation is related to the mean is analogous to the assumption in ordinary least squares regression that the standard deviation is constant for all levels of the mean response.

# 3 Data Sources

To monitor the wet deposition of sulfate, precipitation samples have been collected in several networks of stations located throughout the United States. Typically, a covered bucket is set out for some specified period of time and the bucket opens to collect the sample during periods of precipitation. The precipitation amount is recorded alongside the bucket. Several chemical constituents are measured from the sample, including the sulfate concentration. For this analysis, I use monitoring data from four sites in the Multistage Atmospheric Power Product Pollutions Study (MAP3S) network, located in Whiteface, New York, Ithaca, New York, College Station, Pennsylvania, and Charlottesville, Virginia. In this network, the sample bucket is changed daily during periods of intermittent precipitation and after each episode of heavy precipitation.[4] The data span about 12 years at each of the sites, from late 1976 through 1988. The reported sulfate and precipitation values used in this analysis were obtained from the Acid Deposition System (ADS) and include a quality control flag.[5] I exclude observations that failed the ADS quality control checks.

The regression models in this paper include meteorological covariates mea-

suring temperature and wind conditions. As suggested in other studies, I use 850 millibar wind measurements, selected to be concurrent with the precipitation event, as potential measures of both transport and local weather conditions.[6,7] The wind data used here are from a subset of the National Meteorological Center (NMC) gridded upper air data. The particular data set employed has a grid resolution of two degrees latitude by four degrees longitude,[8] so I selected the grid point closest to the monitoring site. Information about wind is incorporated using the zonal ($U$) and meridional ($V$) components, which provide information about both wind speed and direction.

The regression equations also include a measure of seasonally adjusted temperature, represented by $T$. In this analysis, the temperature variable is also extracted from the 850 millibar measurements using NMC gridded upper air data. For each MAP3S site analyzed, I used the 850 millibar temperature measurements from the closest grid point, taken to be the measure of daily temperature for that precipitation event. Specifically, I calculate $T_i$ as the daily temperature for the $i^{th}$ precipitation event minus the average daily temperature for the month during which the $i^{th}$ precipitation event began. The average daily temperatures for each month are estimated from all of the years for which monitoring data are available. Temperature deviations should be positively associated with sulfate concentration even after adjusting for season and wind direction. Generally, warmer weather indicates an increase in photochemical oxidants for sulfur. Also, higher temperatures during the day can increase the intensity of thermal inversion created by the diurnal cycle of solar heating and nighttime cooling.

In addition to the temperature and wind covariates, I also include a covariate showing the number of days since the last precipitation event ($D$). Since precipitation activity cleanses the air of pollutants, sulfate concentration should be lower during periods of frequent precipitation episodes and should increase as the time span between precipitation episodes increases. And finally, I include an index variable corresponding to year to provide an estimate of long-term trend in deposition. The yearly index variable is represented by $Y$ in the regression equation.

The full specification of the regression model is the following: daily sulfate concentration is assumed to be gamma-distributed with mean

$$\begin{aligned}\mu_i &= \exp(a_1 + \sum_{m=2}^{12} a_m M_{mi} + b_p \log P_i + b_u U_i + b_v V_i \\ &\quad + b_t T_i + b_d D_i + b_y Y_i),\end{aligned}$$

where, for the $i^{th}$ precipitation event, $M_{mi}$ is an indicator variable for the month in which the precipitation event begins, $P_i$ is the total precipitation amount in millimeters, $U_i$ and $V_i$ are the zonal and meridional components of wind, $T_i$ is a measure of temperature in degrees Celsius, $D_i$ is the number of days since the last precipitation event, and $Y_i$ is the index variable for year. The interpretation of the regression coefficients will be illustrated in some detail in the results section below.

# 4 Results and Diagnostics

Table 1(a) shows the estimated regression coefficients for four sites in the MAP3S network as produced by the *glm* function in the S programming environment.[9]

The reported standard errors are computed using the usual asymptotic covariance matrix, assuming the observations are independent. This assumption will be investigated in section 4.2 below. While the primary interest is the estimate of long-term trend for all four sites, I will illustrate the interpretation of all of the regression coefficients by focusing on the site at Whiteface Mountain, New York.

## 4.1 Parameter Interpretation

The parameter $a_1$ is an intercept term and is of little interest. The parameters $a_2$ through $a_{12}$ estimate the difference between the logarithm of the expected sulfate levels for January and the successive months, with all other variables held constant. The estimated coefficient, $\hat{a}_7$, for example, is the difference between July and January. Hence, the expected sulfate concentration is greater in July by a factor of exp(1.162); i.e., it is about three times the predicted amount for January given the same values of the other covariates. In general, these coefficients show how sulfate levels vary seasonally, typically reaching a peak in July or August. The estimated coefficient for precipitation amount, $\hat{b}_p$ is negative, indicating that the expected sulfate concentration decreases as precipitation increases. If precipitation were doubled, for example, with all other covariates held constant, the fitted sulfate value at Whiteface would decrease by a factor of $2^{-0.309}$. Hence, the fitted concentration for the event with more precipitation would be about 81 percent of the fitted value for the smaller precipitation event. The coefficient for the temperature variable $\hat{b}_t$ is positive, where an observed increase of 1 degree Celsius increases the expected

sulfate concentration by a factor of exp(.036), or, roughly, about 3.7 percent. The interpretation of the wind components are presented graphically in Figure 1. These plots show the estimated effect of wind direction given constant wind speeds of 1 m $s^{-1}$ (Figure 1(a)) and 8 m $s^{-1}$ (Figure 1(b)). With 0° as north, the southwest quadrant is associated with winds with low sulfate concentration. Winds from the west and northwest are associated with high sulfate concentration. With a wind speed of 1 m $s^{-1}$, for example, the predicted sulfate concentration on a day with wind from the northwest would be about one percent larger than if the wind were from the south. The second plot shows how the magnitude of the effect changes with the wind speed. For a wind speed of 8 m $s^{-1}$, the predicted sulfate concentration would be about 10 percent higher for the same conditions. The positive coefficient for day count, $\hat{b}_d$, indicates that sulfate concentration is expected to increase for each day without rain that precedes a precipitation event. Specifically, each rainless day increases the expected sulfate concentration by a factor of exp(0.031), or about three percent. And finally, the long-term trend coefficient $\hat{b}_y$ is negative, showing an observed decrease in sulfate concentration through time. At Whiteface, the fitted values decrease about 3.3 percent per year ($\exp(0.032) \times 100\%$).

In general, the estimated coefficients and standard errors are similar among all four sites, though the Virginia site looks somewhat different. Here, for example, the wind components, the number of days between precipitation events, and the long-term trend do not have a significant effect on sulfate concentration. In general, the sign of the coefficients agree with common hypotheses for the behavior of sulfate concentration discussed in the *Data* section.

## 4.2 Regression Diagnostics

As stated above, the standard errors reported in Table 1(a) are calculated under the assumption of independence. Since sulfate observations are time series, it is necessary to check the adequacy of the independence assumption. To investigate the possibility of short-term serial correlation, I examined the standardized deviance residuals from the fitted models. Though small and inconsistent, there is some evidence of positive autocorrelation among the residuals from the four MAP3S sites analyzed here. To adjust for the possibility of short-term serial correlation, I applied a jackknifing technique, leaving out observations in blocks defined by calendar month. The jackknifed standard errors tend to be close to the usual asymptotic estimates. Table 1(b) shows the jackknifed estimates and standard errors for the yearly trend term. The jackknifed estimates are very similar to the standard estimates.

It is possible to perform several informal checks for departures from the assumed gamma model using diagnostics that are similar to the usual diagnostic plots for ordinary least squares regression. These include various plots using standardized deviance residuals, plots to check the assumption that the coefficient of variance is constant, plots to check to adequacy of the logarithmic link function, and plots to check that the terms in the linear predictor do not need to be transformed. Checks for influential points and points with high leverage also follow by analogy with ordinary least squares regression. Overall, the gamma model with a logarithmic link function appears to fit the data well. Checks that the coefficient of variation is constant for each site support the assumption that the standard deviation is proportional to the mean. Examination of the partial

residual plots show that the scale chosen for each covariate in the linear predictor is satisfactory. While there are no extreme outliers, there are a small number of influential points at each site. Table 1(c) shows the estimated coefficients and standard errors with these observations deleted; there is very little change in the estimate of long-term trend.

### 4.3 Long-Term Trend

It is of some interest to determine if the long-term trend is best summarized as a constant percent-per-year change in observed sulfate deposition.[10] To investigate this question, I refit the models described above without any specification for the shape of the long-term trend. Instead, I fit each year as a separate factor so that the long-term trend on the linear scale is fit as an unconstrained step-function. Figure 2 shows the estimated change in the year effects, relative to 1978, obtained from this alternate approach. The constant percent-per-year trend estimate in the basic model is roughly equivalent to smoothing the year-to-year changes with a single straight line. More complicated summaries are also possible. For example, if there is a sudden change in the rate of sulfate deposition due to the Clean Air Act Amendments, the trend can be parameterized by a step function or with a change-point in the slope.

## 5 Comparison with Lognormal Distribution

If the same covariates are fit assuming a lognormal distribution, i.e., with a Normal-theory regression model fit to log-transformed data, the estimated

coefficients are almost identical to the coefficients obtained from the gamma regression model. The final line of Table 1 lists the long-term trend estimates from the lognormal model. A formal comparison of the log likelihood functions shows that the gamma regression models are better at the Whiteface, NY and Virginia sites, while the lognormal model does better at Ithaca, NY. There is virtually no difference at the Pennsylvania site (Table 2).

The advantages of fitting the gamma regression model, though, evolve with other applications. For example, for weekly-level sulfate data, or for values obtained by aggregating to an even longer time scale, the logarithmic transformation tends to be too strong, creating a left-skewed distribution of observations. Also, if the actual fitted values on the original scale of measurement are of interest, the use of transformed data creates the need for an awkward back-transformation step. Though there are reasonable methods for obtaining fitted values on the original scale from an analysis on a transformed data,[11,12] the use of the gamma regression model alleviates the problem. Furthermore, if it is desirable for the model to have an easily interpretable physical meaning, which is the case for the model for sulfate concentration, the use of generalized linear models, like the gamma regression model, can preserve the interpretability of the model.

# 6 Summary and Conclusions

At three of the four sites analyzed here, estimated sulfate concentration in precipitation shows a steadily decreasing trend. A constant percent-per-year

decrease appears to be adequate to describe the trend for the period from the late 1970's to the late 1980's. As further reductions in sulfur dioxide emissions occur, it may be desirable to include a more complicated description of the long-term trend. Gamma regression models provide a flexible way to model sulfate concentration in a parsimonious and easily interpretable manner. While the models require some distributional assumptions, there are no apparent discrepancies for the four sites analyzed here. Finally, gamma regression models can be used for applications that require fitted values or parameter estimates on the original scale of the data.

# 7 References


1. Georgopoulos, P.G. and Seinfeld, J.H., 'Statistical distributions of air pollutant concentrations', *Environ. Sci. Technol.*, **16**,401A-416A (1982).

2. National Acid Precipitation Assessment Program 1992 Report to Congress, Washington, D.C., (1993).

3. McCullagh, P. and Nelder, J. A., *Generalized Linear Models*, Chapman & Hall, New York (1989).

4. MAP3S, 'The MAP3S/RAINE Precipitation Chemistry Network: Statistical overview for the period 1976-1980', *Atmospheric Environment* **16**, 1603-1631 (1982).

5. Watson, C.R. and Olsen, A.R., *Acid Deposition System (ADS) for statistical reporting: system design and user's code manual*, EPA-600-8-84-023, U.S. Environmental Protection Agency, Research Triangle Park, N.C., (1984).

6. Maxwell, C., Eynon, B.P., Endlich, R.M., 'Statistical Analysis of Precipitation Chemistry Measurements over the Eastern United States. Part IV: The Influence of Meteorological Factors', *Journal of Applied Meteorology*, **27**, 1352-1358, (1988).

7. Berge, E., 'Time-Trends of Sulfate and Nitrate in Precipitation in Norway (1972-1982)', *Atmospheric Environment*, **22**, 333-338 (1988).

8. Atmospheric Sciences, University of Washington and Data Support Section, NCAR, *National Meteorological Center Grid Point Data Set: Version II*. University of Washington, (1990).

9. Chambers, J.M. and Hastie, T. J., *Statistical Models in S*, Chapman & Hall, New York, (1993).

10. Sirois, A., 'Temporal variation of sulfate and nitrate concentration in precipitation in Eastern North America: 1979-1990', *Atmospheric Environment*, **27A**, 945-963, (1993).

11. Duan, N., 'Smearing estimate: A nonparametric retransformation method', *Journal of the American Statistical Association*, **78**,605-610, (1983).

12. Styer, P.E. and Stein, M.L., 'Acid deposition models for detecting the effect of changes in emissions: An exploratory investigation utilizing meteorological variables', *Atmospheric Environment*, **26A**, 3019-3028, (1992).

**Table 1** Results from regression models

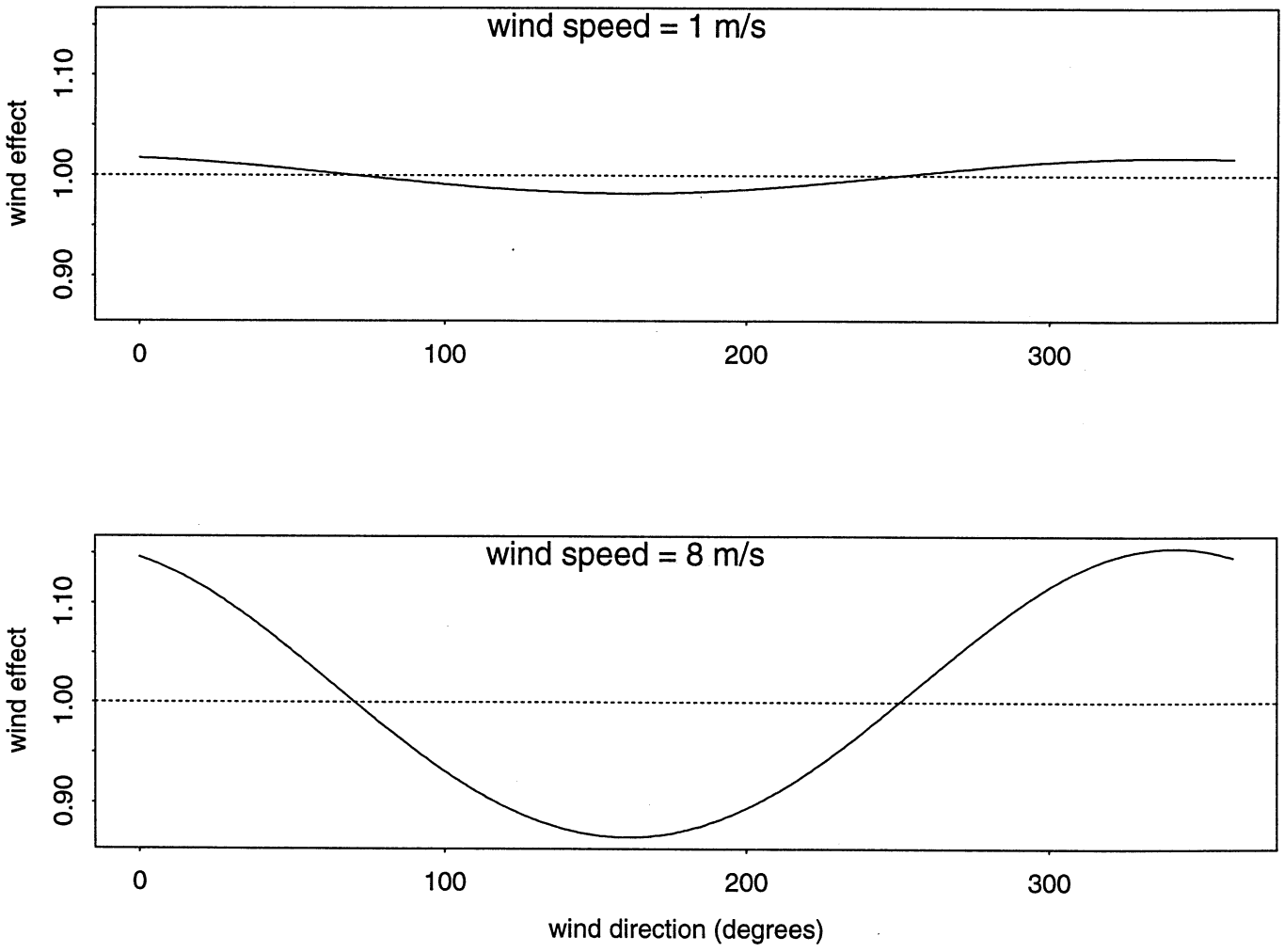
(a) Estimates and standard errors for gamma models

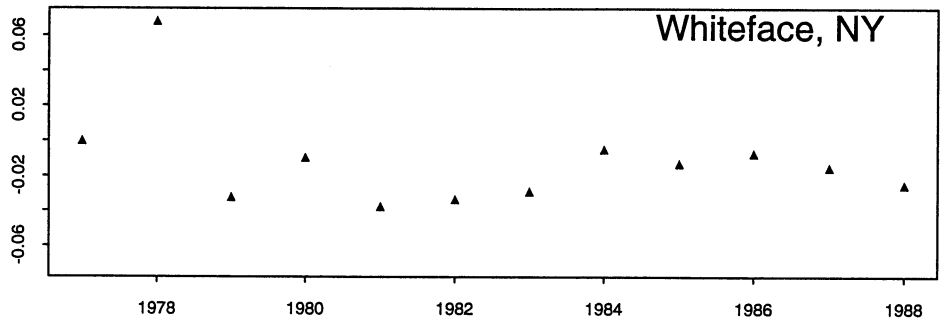
| Coefficients | Whiteface, NY | Ithaca, NY | College Sta., PA | Charlottesville, VA |
|---|---|---|---|---|
| Intercept ($a_1$) | 2.895(0.127) | 3.327(0.118) | 3.245(0.088) | 3.436(0.151) |
| Feb-Jan ($a_2$) | 0.331(0.127) | 0.310(0.108) | 0.358(0.097) | 0.343(0.142) |
| Mar-Jan ($a_3$) | 0.624(0.115) | 0.738(0.103) | 0.646(0.092) | 0.478(0.140) |
| Apr-Jan ($a_4$) | 1.026(0.118) | 1.096(0.104) | 1.111(0.094) | 0.765(0.142) |
| May-Jan ($a_5$) | 1.160(0.115) | 1.310(0.106) | 1.388(0.094) | 0.798(0.141) |
| Jun-Jan ($a_6$) | 1.223(0.114) | 1.370(0.098) | 1.482(0.090) | 1.070(0.141) |
| Jul-Jan ($a_7$) | 1.162(0.116) | 1.430(0.102) | 1.564(0.091) | 1.108(0.135) |
| Aug-Jan ($a_8$) | 1.381(0.114) | 1.597(0.102) | 1.539(0.096) | 1.109(0.137) |
| Sep-Jan ($a_9$) | 0.925(0.116) | 1.194(0.105) | 1.265(0.096) | 0.927(0.149) |
| Oct-Jan ($a_{10}$) | 0.744(0.116) | 0.781(0.100) | 0.819(0.089) | 0.591(0.148) |
| Nov-Jan ($a_{11}$) | 0.428(0.114) | 0.481(0.101) | 0.588(0.091) | 0.508(0.148) |
| Dec-Jan ($a_{12}$) | 0.294(0.115) | 0.210(0.103) | 0.347(0.090) | 0.233(0.143) |
| log precip ($b_p$) | -0.309(0.021) | -0.312(0.020) | -0.280(0.015) | -0.264(0.024) |
| temperature ($b_t$) | 0.036(0.004) | 0.033(0.004) | 0.045(0.004) | 0.020(0.007) |
| u ($b_u$) | -0.023(0.003) | -0.010(0.003) | -0.007(0.003) | -0.001(0.005) |
| v ($b_v$) | 0.014(0.008) | 0.013(0.003) | 0.017(0.003) | 0.001(0.005) |
| day count ($b_d$) | 0.031(0.007) | 0.024(0.007) | 0.023(0.006) | 0.004(0.006) |
| yearly trend ($b_y$) | -0.032(0.008) | -0.036(0.006) | -0.029(0.005) | -0.010(0.008) |
| (b) Jackknifed estimates and standard errors, gamma models | | | | |
| yearly trend ($b_y$) | -0.032(0.008) | -0.035(0.006) | -0.031(0.006) | -0.011(0.010) |
| (c) Estimates with influential points deleted, gamma models | | | | |
| yearly trend ($b_y$) | -0.034(0.006) | -0.036(0.006) | -0.028(0.005) | -0.010(0.008) |
| (d) Estimates and standard errors from lognormal model | | | | |
| yearly trend ($b_y$) | -0.038(0.007) | -0.035(0.006) | -0.028(0.006) | -0.015(0.008) |

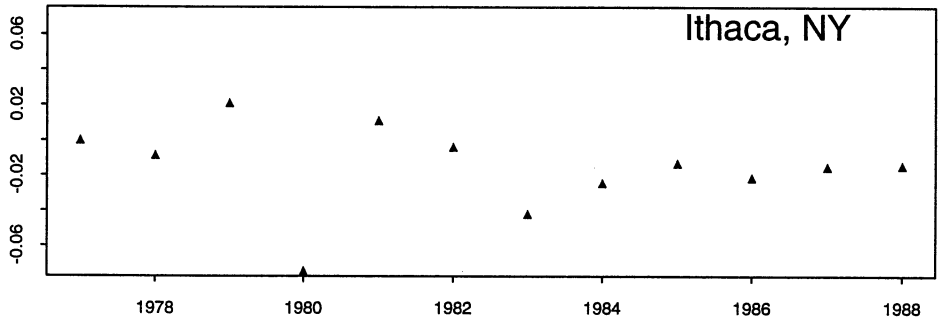
**Table 2** Negative loglikelihood values for the gamma and lognormal regression models

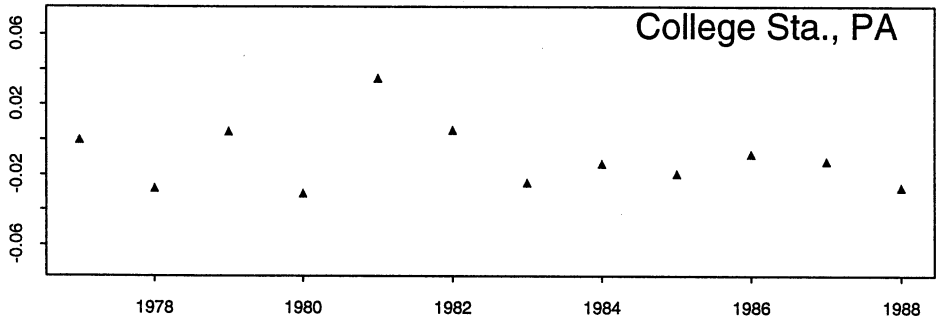
| Model | Whiteface, NY | Ithaca, NY | College Sta., PA | Charlottesville, VA |
|---|---|---|---|---|
| Gamma | 3948.5 | 3637.1 | 4135.2 | 2510.4 |
| Lognormal | 3957.2 | 3613.3 | 4135.9 | 2510.4 |

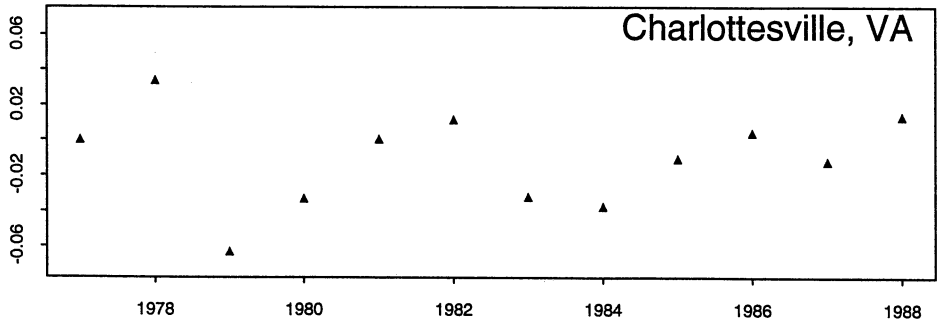
## List of Figures


wind speed = 1 m/s
wind effect
0.90
1.00
1.10
0
100
200
300
wind speed = 8 m/s
wind effect
0.90
1.00
1.10
0
100
200
300
wind direction (degrees)


Whiteface, NY
0.06
0.02
-0.02
-0.06
1978
1980
1982
1984
1986
1988


Ithaca, NY
0.06
0.02
-0.02
-0.06
1978
1980
1982
1984
1986
1988


College Sta., PA
0.06
0.02
-0.02
-0.06
1978
1980
1982
1984
1986
1988


Charlottesville, VA
0.06
0.02
-0.02
-0.06
1978
1980
1982
1984
1986
1988