PREDICTION OF PRECIPITATION IN NEW YORK CITY

A PROJECT REPORT
Presented to the Department of Mathematics and Statistics
California State University, Long Beach

In Partial Fulfillment of the Requirements for the Degree
Bachelor of Science in Mathematics
Option in Statistics

Faculty Reviewer:

Kagba Suaray, Ph.D.

[redacted]
B.S., 2014, California State University, Long Beach

May 2013

Prediction of Precipitation in New York City

[redacted]

ABSTRACT

BACKGROUND: This paper examines the relationship between weather measurements on a given day and precipitation on the following day. Previous studies past millennia have found that there truly exists such a relationship, yet their accuracies are still low present days. It is important to extend the studies by determining which and how much factors that can be observed today affect the probability of precipitation on tomorrow.

METHODS: A sample of 1917 New York's daily observations from National Oceanic and Atmospheric Administration (NOAA)'s National Climatic Data Center (NCDC) were gathered and analyzed through the use of software.

RESULTS: This paper found that departure from normal temperature, average due point, the observation of natural freezing, daily temperature range, and the chronological position of the given day have relationship with the probability of precipitation on the following day.

CONCLUSIONS: Even though a statistically significant model was found, it is important to note that there are many other variables which cannot be observed on the surface that can change the probability of precipitation. This paper focuses on effective variables among those which can be easily obtained on surface.

Table of Contents

Human beings have attempted to predict weather for ages. They constructed countless weather stations with super computers over the world, and yet, people are often mad at the forecasters for not precisely predicting the weather. In fact, due to the nature of limited quantities of observable data, it is impossible to make a spot-on forecast. In agriculture, weather forecasting is crucial because weather directly affects profit and loss. For most other people in urban area, one of the most common concerns is whether they should bring umbrellas to work tomorrow or not.

Knowing future weather depends upon knowing what the weather is doing now. In order to make most precise prediction of a certain place, one should look into more than weather readings taken at the area; readings from nearby places and around the globe, as well as weather balloon observations give a better picture of what is going to happen in the area.[1]

Method

For simpler approach, we examine and attempt to find the odds it is going to either rain or snow in New York City on the following day, based on the weather readings measured on the given day. We are going to use SAS 9.3 to find the relationship.

Data Collection

The local climatological data was collected and used to predict if it will be necessary to bring an umbrella for tomorrow based on today's weather conditions in New York. The data was collected by National Oceanic and Atmospheric Administration (NOAA)'s National Climatic Data Center (NCDC). It contains 1917 samples which were observed daily from January 1st, 2008 to March 31st, 2013, at Belvedere Castle in Central Park, New York, New York.

Data Preparation

The raw data originally had total of 24 different variables (See Appendix A), but some modifications were necessary for the data to be imported to SAS for interpretation since some of the data were non-numerical or non-informative. For example, one of the columns gives a type of significant weather on the given day and its values were RA for rain, SN for snow, HZ for haze and so forth.

To make the modification, the data was first imported to Microsoft Excel 2013. Eleven additional columns were generated and seven were eliminated. The first three columns contain chronological data (year, month and date). In order to simplify and generalize these data, a column was generated to inform if a given day was among first half of the year or not; this column is binary. Another column was added by calculating the difference between maximum and minimum temperatures ("daily temperature range") on a given day. Also, from the "significant weather" column, seven columns were generated. The seven columns have binary data (1 for true, 0 for false). For example, if it rained on the given day, it has a value of 1 on the "rain" column. "Umbrella" column was generated by combining "rain" and "snow" columns to inform if an umbrella was necessary on the given day. The "umbrella" column was copied and moved upward by one cell to generate "umbrella tomorrow" column to inform if an umbrella was necessary on the following day. This column represents our responsive variable. At this point, our sample size is reduced by 1, since otherwise, this column would have a missing value for the very last entry; the information whether an umbrella was necessary on April 1st, 2013 was not included in the original data.

Then, the first three "chronological" columns were removed as well as "sunrise time", "sunset time", and "significant weather" columns. Additionally, "amount of precipitation" column was also eliminated since its information is somewhat repeated in the "umbrella" column. For the same reason, "rain" and "snow" columns were eliminated. In fact, we are specifically interested in if it's actually necessary to prepare an umbrella for tomorrow, not how much it is going to rain. Either it's going to rain little or much, an umbrella is necessary for both cases.

The data now contains 25 explanatory variables and one responsive variable ($n = 1916$). They are labeled X_1, X_2, …, X_{27}, and Y. The below are the brief descriptions of the variables.

- X_1: Maximum temperature (ºF)
- X_2: Minimum temperature (ºF)
- X_3: Average temperature (ºF)
- X_4: Departure from normal temperature (ºF)
- X_5: Average dew point (ºF)
- X_6: Average wet bulb temperature (ºF)
- X_7: Heating degree day (ºF)
- X_8: Cooling degree day (ºF)
- X_9: Average pressure at station (inHg)
- X_{10}: Average pressure at sea level (inHg)
- X_{11}: Resultant wind speed (mph)
- X_{12}: Resultant wind direction (degrees)
- X_{13}: Average wind speed (mph)
- X_{14}: 5 second maximum wind speed (mph)
- X_{15}: 5 second maximum wind direction (degrees)
- X_{16}: 2 minute maximum wind speed (mph)
- X_{17}: 2 minute maximum wind direction (degrees)
- X_{18}: Was it heavily foggy? (1=True, 0=False)
- X_{19}: Was it foggy? (1=True, 0=False)
- X_{20}: Was it misty? (1=True, 0=False)
- X_{21}: Was it hazy? (1=True, 0=False)
- X_{22}: Was it freezing? (1=True, 0=False)
- X_{23}: Was an umbrella necessary? (1=True, 0=False)

X_{24}: Daily temperature range (°F)
X_{25}: Was it among the first half of the year? (1=True, 0=False)
Y: Was an umbrella necessary on the following day? (1=True, 0=False)

The data is now imported to SAS (See Appendix B: `CODE 1`), and one shall further examine each explanatory variable to decide which ones should be actually kept before the model building begins.

Preliminary Model Investigation

The binary responsive variable limits our choice of model, and multiple logistic regression model was chosen to interpret the data. One expects the final model to have the following form:

$$E\{\hat{Y}\} = \hat{\pi} = \frac{\exp(\boldsymbol{X'w})}{1 + \exp(\boldsymbol{X'w})}$$

, where $\boldsymbol{w}$ is a matrix of subsets of $\boldsymbol{b}$.

The logistic procedure was taken with all 25 explanatory variables (see `CODE 2`), and SAS returned the following:

`OUTPUT 2`

Note: The following parameters have been set to 0, since the variables are a linear combination of other variables as shown.

X8 =	- 65 * Intercept + 1 * X3 + X7
X24 =	1 * X1 - 1 * X2

The second equation was expected since it is how the "daily temperature range" column (X_{24}) was calculated and generated. The first equation, on the other hand, does not seem clear at first, but it was revealed that cooling degree days (X_8) is actually a function of average temperature (X_3) and heating degree days (X_7) by its very own definition. Since it is impractical to keep multiple variables that have same information, X_1, X_2, X_7 and X_8 were eliminated. Again, with the remaining 21 explanatory variables, the logistic procedure was taken (see `CODE 3`):

`OUTPUT 3`

Testing Global Null Hypothesis: BETA=0

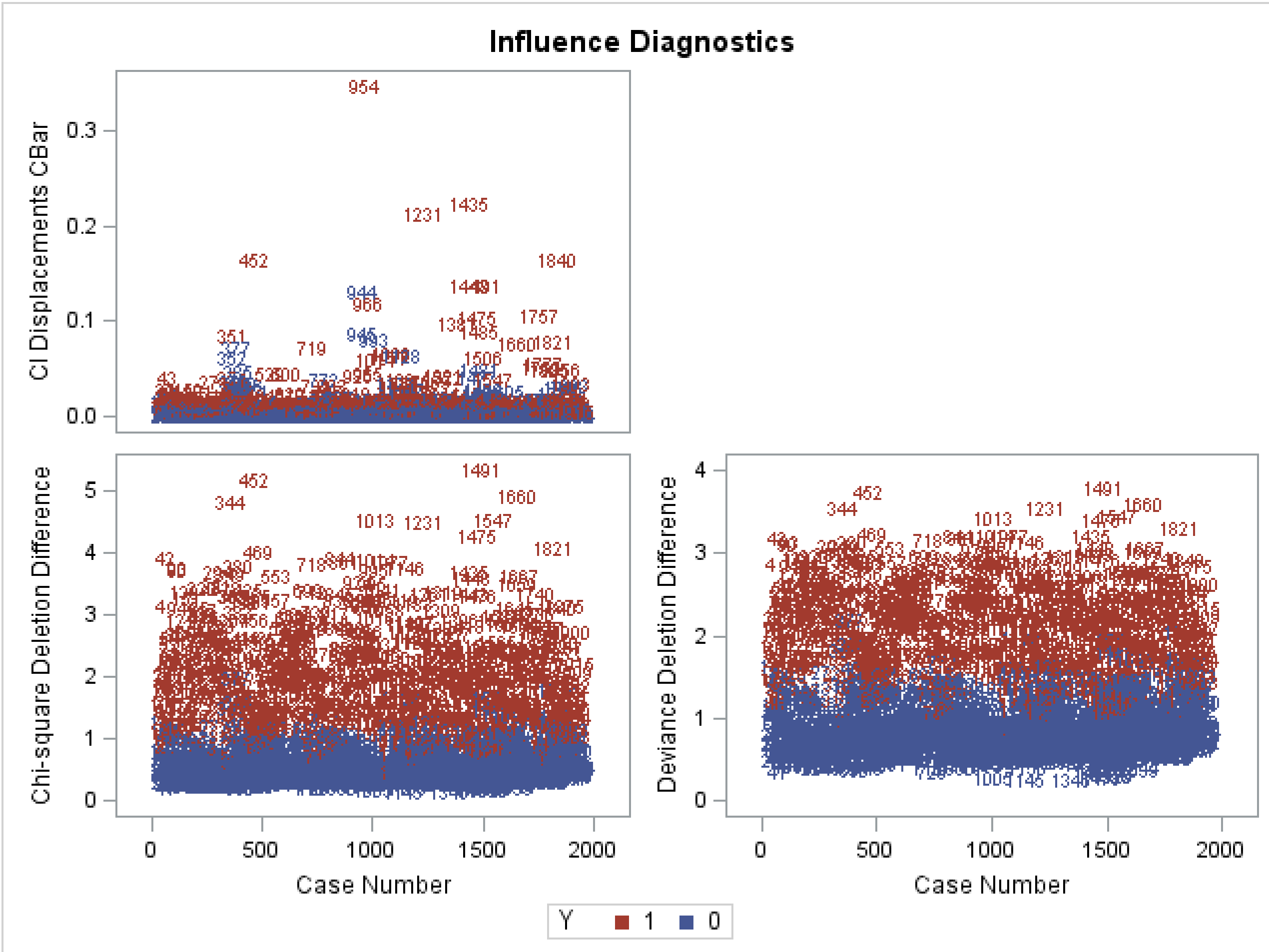
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	126.3209	21	<.0001
Score	124.7523	21	<.0001
Wald	117.2393	21	<.0001

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	0.4514	0.7993	0.3189	0.5723
X3	1	0.00740	0.0107	0.4777	0.4895
X4	1	-0.0195	0.00811	5.7597	0.0164
X5	1	-0.0121	0.0121	1.0076	0.3155
X6	1	0.00894	0.0127	0.4963	0.4811
X9	1	-0.0277	0.0270	1.0517	0.3051
X10	1	-0.0167	0.0152	1.2118	0.2710
X11	1	0.0173	0.0213	0.6567	0.4177
X12	1	0.00644	0.00650	0.9795	0.3223
X13	1	-0.0498	0.0404	1.5223	0.2173
X14	1	-0.00873	0.0188	0.2147	0.6431
X15	1	0.00336	0.000900	13.9157	0.0002
X16	1	0.0609	0.0350	3.0299	0.0817
X17	1	-0.00070	0.000887	0.6251	0.4292
X18	1	-0.3205	0.4857	0.4355	0.5093
X19	1	0.6149	0.3381	3.3084	0.0689
X20	1	0.1207	0.1587	0.5780	0.4471
X21	1	-0.2269	0.1658	1.8733	0.1711
X22	1	-0.5040	0.4255	1.4028	0.2363
X23	1	-0.5739	0.1477	15.1071	0.0001
X24	1	0.0307	0.0120	6.5912	0.0102
X25	1	-0.2721	0.1142	5.6784	0.0172

As shown, only 5 out of 21 explanatory variables had *p*-values of less than 0.05. As expected, not all explanatory variables were appropriate to be kept in the model. Also, notice the "was an umbrella necessary?" variable (X_{23}) has a significantly large Wald chi-square value (= 15.1071). This variable is decided to be dropped since it won't be useful when the model building begins; it may prevent other important predictor variables from being *interesting*. In fact, it is a common sense to expect another rainy day after a rainy day. Finding odds of raining more unexpectedly is more interested.

OUTPUT 4



Data Refinement

Before the actual procedure of model selection, it is first necessary to check if there are any (1) potential outliers and/or (2) multicollinearity among the explanatory variables.

Outliers

Detecting potential outliers, or influential observations, is crucial since they have noticeably larger impact on various estimates than most of other observations.[2] In this case where the model is logistic, case-deletion diagnostics will be employed to notice the effect of individual cases on the analysis. Specifically, the Pearson chi-square and the deviance statistics were used to spot influential changes where i^{th} observation is deleted. Unlike standard regression situations, generalized guidelines of decision rule are not available for logistic regression since the distribution of the delta statistics is unknown except under certain restrictive assumptions.[3] Instead, the judgment will be made on the basis of a subjective visual assessment. Also, `LABEL` option was also employed because otherwise, even if an influential observation was detected visually, it would be nearly impossible to actually spot which exact observation it is due to its large sample size (n = 1916) (see `CODE 4`). `OUTPUT 4` shows that both Pearson chi-square and deviance difference vs. case number plots (bottom left and right) suggest that 452nd or 1491st observations are influential. Either of them would change the Pearson chi-square statistic by more than 5 and deviance by almost 4. The two cases were inspected and it was shown that both of the cases had zero values for "average pressure at sea level" (X_{10}). Since it is physically and geographically impossible to actually observe 0 inHg of air pressure, the values were assumed to be entry errors. Therefore the two cases were eliminated (n = 1914) (see `CODE 5`).

Multicollinearity

Variables with multicollinearity should be eliminated or fixed before model building because of its potential effects on regression coefficients and their standard errors.[3]

OUTPUT 6

Pearson Correlation Coefficients, N = 1916
Prob > |r| under H0: Rho=0

	X9	X10	X11	X13	X14	X16	X12	X15	X17
X9	1.00000	0.56986	-0.00445	0.00910	-0.00333	0.01060	-0.01615	0.00138	-0.01133
		<.0001	0.8457	0.6906	0.8843	0.6430	0.4798	0.9519	0.6200
X10	0.56986	1.00000	0.03627	-0.00201	0.00149	-0.00211	0.06235	0.01416	0.00941
	<.0001		0.1125	0.9298	0.9480	0.9263	0.0063	0.5358	0.6807
X11	-0.00445	0.03627	1.00000	0.73493	0.60323	0.61890	0.08206	0.05870	0.06377
	0.8457	0.1125		<.0001	<.0001	<.0001	0.0003	0.0102	0.0052
X13	0.00910	-0.00201	0.73493	1.00000	0.76654	0.82414	0.09274	0.09731	0.08230
	0.6906	0.9298	<.0001		<.0001	<.0001	<.0001	<.0001	0.0003
X14	-0.00333	0.00149	0.60323	0.76654	1.00000	0.91748	0.11481	0.23978	0.18233
	0.8843	0.9480	<.0001	<.0001		<.0001	<.0001	<.0001	<.0001
X16	0.01060	-0.00211	0.61890	0.82414	0.91748	1.00000	0.05259	0.11515	0.09083
	0.6430	0.9263	<.0001	<.0001	<.0001		0.0213	<.0001	<.0001
X12	-0.01615	0.06235	0.08206	0.09274	0.11481	0.05259	1.00000	0.61491	0.62510
	0.4798	0.0063	0.0003	<.0001	<.0001	0.0213		<.0001	<.0001
X15	0.00138	0.01416	0.05870	0.09731	0.23978	0.11515	0.61491	1.00000	0.79514
	0.9519	0.5358	0.0102	<.0001	<.0001	<.0001	<.0001		<.0001
X17	-0.01133	0.00941	0.06377	0.08230	0.18233	0.09083	0.62510	0.79514	1.00000
	0.6200	0.6807	0.0052	0.0003	<.0001	<.0001	<.0001	<.0001	

r

OUTPUT 7

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	0.42219	0.11060	3.82	0.0001	0
X3	1	-0.00239	0.00223	-1.07	0.2843	12.71784
X4	1	0.00330	0.00173	1.91	0.0567	1.29027
X5	1	0.00408	0.00237	1.72	0.0855	17.45980
X6	1	-0.00145	0.00254	-0.57	0.5692	15.77471
X10	1	0.00353	0.00286	1.24	0.2168	1.36186
X16	1	-0.00571	0.00279	-2.05	0.0410	1.24690
X19	1	-0.11424	0.06195	-1.84	0.0653	1.48158
X20	1	0.06533	0.02920	2.24	0.0254	1.72902
X21	1	0.01452	0.03610	0.40	0.6876	1.13920
X22	1	0.17217	0.08950	1.92	0.0546	1.47818
X24	1	-0.00960	0.00248	-3.88	0.0001	1.54522
X25	1	0.07619	0.02418	3.15	0.0017	1.28477

Some subsets of the explanatory variables are suspected to have high multicollinearity due to their nature. "Average pressure at the station" (X_9) and "average pressure at sea level" (X_{10}) almost always have very similar values, because both variables have same type of information measured with different methods. Similarly, X_{11}, X_{13}, X_{14} and X_{16} have information on speed of wind measured at different time intervals and methods, and X_{12}, X_{15}, and X_{17} have information on direction of wind at different time intervals and methods. To check multicollinearity among these variables, a correlation matrix was generated (see `CODE 6`). As shown in `OUTPUT 6`, the result confirms the theory. To resolve this, seven explanatory variables (X_9, X_{11}, X_{13}, X_{14}, X_{12}, X_{15} and X_{17}) were eliminated.

Moreover, the "fog" column (X_{19}) completely includes information in the "thick fog" column (X_{18}), and in order to avoid repentance of information, X_{18} was also eliminated.

The correlation matrix produced by `PROC CORR` is useful to confirm multicollinearity between two variables, but it is not helpful enough to detect any other underlying multicollinearity, since it is possible to have data in which no pair of variables has a high correlation, but several variables together may be highly interdependent.[4] Instead, `PROC REG` with `VIF` option gives variance inflation factors, which are helpful to detect multicollinearity among more than two variables. Using `PROC REG` while the data is logistic may seem inappropriate, but multicollinearity is a property of the explanatory variables, not the dependent variables. `PROC REG` is used only for its `VIF` option, which `PROC LOGISTC` does not have (see `CODE 7`).

The result shown in `OUTPUT 7` indicates presence of multicollinearity among "average temperature" (X_3), "average dew point" (X_5) and "average wet bulb" (X_6) with variance inflation factors greater 10. (Recall that coefficient estimates and test statistics shown in this figure are not the interests here.)

Again, the variables (X_3 and X_6) were dropped to resolve the multicollinearity. Fortunately, there are still enough number of explanatory variables for model building. Another `VIF` procedure was run to recheck any multicollinearity among remaining variables (see `CODE 8`) and no more multicollinearity was detected as shown in `OUTPUT 8`.

`OUTPUT 8`

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	0.36355	0.09323	3.90	<.0001	0
X4	1	0.00312	0.00172	1.82	0.0696	1.27226
X5	1	0.00089616	0.00079156	1.13	0.2577	1.94024
X10	1	0.00357	0.00250	1.43	0.1537	1.03890
X16	1	-0.00541	0.00277	-1.95	0.0515	1.23008
X19	1	-0.11013	0.06188	-1.78	0.0753	1.47824
X20	1	0.08130	0.02710	3.00	0.0027	1.48936
X21	1	0.01571	0.03608	0.44	0.6633	1.13802
X22	1	0.18275	0.08922	2.05	0.0407	1.46855
X24	1	-0.01125	0.00220	-5.11	<.0001	1.22275
X25	1	0.07650	0.02414	3.17	0.0016	1.28061

In summary, in this model refinement stage, two observations (i = 452, 1491) were eliminated due to its influence on the analysis, and 10 explanatory variables were eliminated due to their high multicollinearity. Now the data has 1914 observations and 10 explanatory variables (listed above) and no other influential outlying observations or multicollinearity are present.

Model Selection

For logistic regression models, the AIC_p and SBC_p criteria are easily adapted to select "best" subsets of parameters.[3] The subsets with smaller values for these criteria are more favorable.

Unfortunately, in SAS 9.3, unlike `PROC REG`, `PROC LOGISTIC` does not include automated AIC_p and SBC_p criteria in its `SELECTION` option. There are 10 explanatory variables and there are 2^{10} = 1024 possible models to compare. Finding the best model by *direct* comparison is an unrealistic task. One of the possible ways to resolve this problem is to use the `STEPWISE` option with `SLE` and `SLS` close to 1 (see `CODE 9`). The result will give a sequence of models starting with an empty subset and finishing with a full subset in a way of maximizing the likelihood at each step.[5] The interpretation of this result is different from the "usual" `STEPWISE` (in which case `SLE` and `SLS` have values of 0.15, 0.30, etc.); instead of getting a single stepwise result, the entire sequence is obtained. The interpretation of the result is similar to that of `PROC REG` with `STEPWISE BEST=1` option.

OUTPUT 9

Summary of Stepwise Selection

Step	Effect Entered	Effect Removed	DF	Number In	Score Chi-Square	Wald Chi-Square	Pr > ChiSq
1	**X20**		1	1	26.7064		<.0001
2	**X24**		1	2	15.3020		<.0001
3	**X25**		1	3	6.2446		0.0125
4	**X16**		1	4	6.7234		0.0095
5	**X4**		1	5	4.1013		0.0429
6	**X10**		1	6	2.6353		0.1045
7	**X22**		1	7	1.3235		0.2500
8	**X19**		1	8	2.8329		0.0924
9	**X5**		1	9	1.3802		0.2401
10	**X21**		1	10	0.1674		0.6824

The result in `OUTPUT 9` should be interpreted as following: step 1 indicates $\{X_{20}\}$ is the best subset when $p = 2$, and step 2 indicates $\{X_{20}, X_{24}\}$ is the best subset when $p = 3$, and so forth, based on their score statistics.

It is necessary to note that "mist" (X_{20}) and "daily temperature range" (X_{24}) seem to have abnormally large influence on the responsive variable, which was not noticeable in `OUTPUT 3`. However, these variables are decided to be kept since dropping these might result in omitting too many important variables. If the final model shows an unexpected result because of these two variables, one shall then come back to this point and drop them.

`PROC LOGISTIC` with `ODS OUTPUT` statement generates a table that contains AIC_p and SBC_p statistics for each subset of the sequence.

OUTPUT 10

Obs	Step	Criterion	Equals	InterceptOnly	InterceptAndCovariates
1	0	-2 Log L	=	2451.541798	2451.542
2	1	AIC		2453.541798	2429.139
3	1	SC		2459.098749	2440.253
4	1	-2 Log L		2451.541798	2425.139
5	2	AIC		2453.541798	2415.668
6	2	SC		2459.098749	2432.339
7	2	-2 Log L		2451.541798	2409.668
8	3	AIC		2453.541798	2411.412
9	3	SC		2459.098749	2433.640
10	3	-2 Log L		2451.541798	2403.412
11	4	AIC		2453.541798	2406.610
12	4	SC		2459.098749	2434.395
13	4	-2 Log L		2451.541798	2396.610
14	5	AIC		2453.541798	2404.520
15	5	SC		2459.098749	2437.862
16	5	-2 Log L		2451.541798	2392.520
17	6	AIC		2453.541798	2403.737
18	6	SC		2459.098749	2442.636
19	6	-2 Log L		2451.541798	2389.737
20	7	AIC		2453.541798	2404.433
21	7	SC		2459.098749	2448.889
22	7	-2 Log L		2451.541798	2388.433
23	8	AIC		2453.541798	2403.531
24	8	SC		2459.098749	2453.543
25	8	-2 Log L		2451.541798	2385.531
26	9	AIC		2453.541798	2404.150
27	9	SC		2459.098749	2459.719
28	9	-2 Log L		2451.541798	2384.150
29	10	AIC		2453.541798	2405.983
30	10	SC		2459.098749	2467.109
31	10	-2 Log L		2451.541798	2383.983

`OUTPUT 10` is the result of the generated table (see `CODE 10`). As shown, the AIC_p statistic is minimal (= 2403.531) in step 8 (X_{20}, X_{24}, X_{25}, X_{16}, X_4, X_{10}, X_{22}, X_{19} entered), and the SBC_p statistic is minimal (= 2460.694) in step 2 (X_{20}, X_{24} entered).

One now compares these two subsets with the results of the forward selection (see `CODE 11`), the backward elimination (see `CODE 12`), and the stepwise selection (see `CODE 13`). The significance level is set to 0.1 for both `SLE` and `SLS` for all methods.

`OUTPUT 11`

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	2453.542	2404.520
SC	2459.099	2437.862
-2 Log L	2451.542	2392.520

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-0.0969	0.2388	0.1649	0.6847
X4	1	-0.0151	0.00746	4.0895	0.0431
X16	1	0.0290	0.0119	5.9013	0.0151
X20	1	-0.4138	0.1031	16.1097	<.0001
X24	1	0.0477	0.0100	22.6286	<.0001
X25	1	-0.2940	0.1007	8.5256	0.0035

`OUTPUT 13`

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	2453.542	2404.520
SC	2459.099	2437.862
-2 Log L	2451.542	2392.520

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-0.0969	0.2388	0.1649	0.6847
X4	1	-0.0151	0.00746	4.0895	0.0431
X16	1	0.0290	0.0119	5.9013	0.0151
X20	1	-0.4138	0.1031	16.1097	<.0001
X24	1	0.0477	0.0100	22.6286	<.0001
X25	1	-0.2940	0.1007	8.5256	0.0035

`OUTPUT12`

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	2453.542	2404.339
SC	2459.099	2448.794
-2 Log L	2451.542	2388.339

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-0.1273	0.2398	0.2820	0.5954
X4	1	-0.0164	0.00752	4.7827	0.0287
X16	1	0.0304	0.0120	6.4351	0.0112
X19	1	0.4774	0.2856	2.7950	0.0946
X20	1	-0.4290	0.1066	16.1944	<.0001
X22	1	-0.7295	0.3952	3.4075	0.0649
X24	1	0.0482	0.0101	23.0009	<.0001
X25	1	-0.2872	0.1009	8.0955	0.0044

As `OUTPUT 11, 12,` and `13` show (not in order), the stepwise selection and the forward selection methods give same results with strong *p*-values (<0.05) of the Wald statistics for all parameters: X_4, X_{16}, X_{20}, X_{24}, and X_{25}. The backward elimination method give a result with the subset of X_4, X_{16}, X_{19}, X_{20}, X_{22}, X24, and X_{25}, where two of the variables have *p*-values of the Wald statistics greater than 0.05. Unfortunately, none of these results is identical to either of the results of AIC_p and SBC_p criteria. Moreover, the number of variables in the subset of each criterion vary too much; the SBC_p criterion has only two, where the backward elimination criterion has seven. Therefore, it is tough to predict which variables should be actually kept. One assumed that this is due to the two variables (X_{20} and X_{24}) not dropped from the whole set, which had abnormally large influences, as previously mentioned. Thus, the entire procedures so far in this model selection stage were repeated (a) after X_{20} is dropped, (b) after X_{24} is dropped, and (c) after both X_{20} and X_{24} are dropped (See `CODE 14`). The interpretation of these outputs is trivial and not shown. `TABLE1` is the summary of all three possible cases.

TABLE 1

Criterion	AICp	SBCp	Forward Selection	Backward Elimination	Stepwise Selection
a) X_{20} dropped	X_4, X_5, X_{24}, X_{25}	X_4, X_5, X_{16}, X_{22} X_{24}, X_{25}	X_4, X_5, X_{22}, X_{24}, X_{25}	X_4, X_5, X_{22}, X_{24}, X_{25}	X_4, X_5, X_{22}, X_{24}, X_{25}
b) X_{24} dropped	X_{16}, X_{20}, X_{25}	X_{20}	X_{16}, X_{20}, X_{25}	X_{16}, X_{20}, X_{25}	X_{16}, X_{20}, X_{25}
c) X_{20} and X_{24} dropped	X_5, X_{22}, X_{25}	X_5, X_{25}	X_5, X_{22}, X_{25}	X_5, X_{22}, X_{25}	X_5, X_{22}, X_{25}

As shown in TABLE 1, all three cases have much better consistency of predictor selections for multiple criteria. Specifically, each of case (b) and (c) has same subsets for all criteria except the SBC_p criterion. However, the number of predictor variables are too small to make the final model informative and useful. The case (a), on the other hand, also has good consistency and the numbers of variables in the subsets are not too small. Therefore, one decided to accept this case: drop the X_{20}. Further analytic investigation is necessary to choose a single best subset by comparing each model.

PROC LOGISTIC procedures are taken for each of the three possible subsets: {X_4, X_5, X_{24}, X_{25}}, {X_4, X_5, X_{16}, X_{22}, X_{24}, X_{25}}, and {X_4, X_5, X_{22}, X_{24}, X_{25}} (see CODE 15, CODE 16, CODE 17, respectively).

OUTPUT 15

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	2451.378	2409.163
SC	2456.934	2436.945
-2 Log L	2449.378	2399.163

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	50.2151	4	<.0001
Score	49.7274	4	<.0001
Wald	48.5110	4	<.0001

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	0.5296	0.1945	7.4157	0.0065
X4	1	-0.0122	0.00775	2.5001	0.1138
X5	1	-0.0105	0.00306	11.8500	0.0006
X24	1	0.0585	0.00991	34.8149	<.0001
X25	1	-0.4228	0.1080	15.3275	<.0001

OUTPUT 16

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	2451.378	2407.599
SC	2456.934	2446.494
-2 Log L	2449.378	2393.599

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	55.7787	6	<.0001
Score	55.3070	6	<.0001
Wald	53.6644	6	<.0001

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	0.2595	0.3012	0.7421	0.3890
X4	1	-0.0136	0.00778	3.0445	0.0810
X5	1	-0.00974	0.00322	9.1762	0.0025
X16	1	0.0174	0.0125	1.9317	0.1646
X22	1	-0.6582	0.3286	4.0123	0.0452
X24	1	0.0582	0.00996	34.1814	<.0001
X25	1	-0.4235	0.1084	15.2697	<.0001

OUTPUT 17

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	2453.542	2410.785
SC	2459.099	2444.126
-2 Log L	2451.542	2398.785

Testing Global Null Hypothesis: BETA=0			
Test	**Chi-Square**	**DF**	**Pr > ChiSq**
Likelihood Ratio	52.7571	5	<.0001
Score	52.4292	5	<.0001
Wald	50.9777	5	<.0001

Analysis of Maximum Likelihood Estimates					
Parameter	**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept	1	0.5595	0.1961	8.1445	0.0043
X4	1	-0.0133	0.00776	2.9202	0.0875
X5	1	-0.0107	0.00307	12.1417	0.0005
X22	1	-0.6213	0.3267	3.6166	0.0572
X24	1	0.0571	0.00991	33.2203	<.0001
X25	1	-0.4052	0.1079	14.0987	0.0002

The result shows that the second subset has a minimal AIC_p statistic as well as a largest chi-square statistic. However, the third subset is the only one that has the chi-square *p*-values less than 0.1 for all parameters, and therefore it is the only suitable option. Thus, this subset, $\{X_4, X_5, X_{22}, X_{24}, X_{25}\}$, is chosen.

Model Specification

Once the subset is decided, it is necessary to test whether either higher-order variables or interaction terms are to be entered.

Polynomial Regression

Occasionally, the first-order logistic model may not give *good* fit of the data.[3] In order to confirm if the first-order model is enough, higher-order polynomial models need to be compared. For simplicity, only a second-order polynomial model is generated. The X_{22} and X_{25} predictors are binary variables, implying that higher-order of these variables are non-significant, since their values won't change. Therefore, only the other three variables are squared and entered to the model to be tested. Centering their values in advance was taken to avoid multicollinearity (see `CODE 18`), since otherwise, they often would be highly correlated.[3]

`OUTPUT 18`

Analysis of Maximum Likelihood Estimates					
Parameter	**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept	1	0.5622	0.2127	6.9887	0.0082
X4	1	-0.0111	0.00835	1.7681	0.1836
X5	1	-0.0111	0.00317	12.1712	0.0005
X22	1	-0.6084	0.3276	3.4483	0.0633
X24	1	0.0579	0.00993	33.9539	<.0001
X25	1	-0.4081	0.1084	14.1695	0.0002
X4CSQ	1	-0.00021	0.000697	0.0904	0.7637
X5CSQ	1	0.000087	0.000148	0.3473	0.5556
X24CSQ	1	-0.00065	0.00130	0.2468	0.6193

The parameters that has postfix "CSQ" in the `OUTPUT 18` are the squared values of centered variables. The result suggests that none of the variables should be entered in quadratic forms as they all have very high chi-square *p*-values. Therefore, no quadratic terms are added.

Interaction Effects

To test if any cross-product terms may be helpful for model building, all possible two-factor interaction terms are added to the model and then the likelihood ratio test is taken. However, the binary variables are hard to interpret when they are interacted with other variables – either quantitative or qualitative. They will always return either same value or zero. Thus, again, the binary variables (X_{22} and X_{25}) are not considered to be interacted in this case (see `CODE 19`).

One wishes to apply the likelihood ratio G^2 test where the full model is

$$\mathbf{X'\beta_F} = \beta_0 + \beta_4 X_4 + \beta_5 X_5 + \beta_{22} X_{22} + \beta_{24} X_{24} + \beta_{25} X_{25} + \beta_{45} X_4 X_5 + \beta_{424} X_4 X_{24} + \beta_{524} X_5 X_{24}$$

, and the reduced model is

$$\mathbf{X'\beta_R} = \beta_0 + \beta_4 X_4 + \beta_5 X_5 + \beta_{22} X_{22} + \beta_{24} X_{24} + \beta_{25} X_{25}$$

, and the test is

H_0: $\beta_{45} = \beta_{424} = \beta_{524} = 0$
H_1: not all β_k in H_0 equal zero

`OUTPUT 19a`

Summary of Fit			
Deviance	2398.7847	Pearson ChiSq	1919.8884
Deviance / DF	1.2572	Pearson ChiSq / DF	1.0062
Scaled Dev	2398.7847	Scaled ChiSq	1919.8884

`OUTPUT 19b`

Summary of Fit			
Deviance	2393.1376	Pearson ChiSq	1916.1518
Deviance / DF	1.2562	Pearson ChiSq / DF	1.0059
Scaled Dev	2393.1376	Scaled ChiSq	1916.1518

, and the test statistic is

$$G^2 = -2[\log_e(R) - \log_e(F)] \\ = 2398.7847 - 2393.1376 \\ = 5.6471$$

The G^2 statistic is smaller than the cut-off point $\chi^2(0.95; 3) = 7.8147$, therefore H_0 is accepted; none of the interaction terms should be added. This is also evident by the following table:

`OUTPUT 19c`

Type III (Wald) Tests			
Source	**DF**	**ChiSq**	**Pr > ChiSq**
X4	1	3.3061	0.0690
X5	1	5.8973	0.0152
X22	1	4.6295	0.0314
X24	1	0.6808	0.4093
X25	1	14.3352	0.0002
x4x5	1	0.0002	0.9892
x4x24	1	2.6060	0.1065
x5x24	1	1.5506	0.2130

Note that all the interaction terms have high chi-square *p*-values greater than 0.1. Moreover, the chi-square *p*-value for X_{24} has changed ridiculously so that it is no longer significant to the model; in the reduced model, the value was less than 0.0001 whereas it is now 0.4093. Even if any of these values were significantly low, it is actually preferable not to have interaction terms if the regression model is logistic.[3] The final model is expected to explain how each individual meteorological factors affects the odds of raining. When interaction terms are entered into a logistic regression model, the odds ratio for a given explanatory variable could be no longer independent from the other variables, therefore the coefficient of the variable has to be interpreted differently,[6] and this will destroy the purpose of this paper. In fact, many researchers prefer to logistic interaction interpret results in terms of probability.[7]

In this model validation stage, neither additional higher order nor interaction terms are entered. In fact, avoiding either of them is desirable for simple interpretation of the model. Coincidently, and fortunately, none of these criteria met the conditions.

In summary, the decided model (not final; to be validated) is a multiple logistic regression model with five explanatory variables ($p = 6$) that has the following form:

$$\mathrm{E}\{Y\} = [1 + \exp(0.5596 - 0.0133X_4 - 0.0107X_5 \\ - 0.6213X_{22} + 0.0571X_{24} \\ - 0.4053X_{25})]^{-1}$$

, or equivalently,

$$\mathrm{logit}(Y) = -0.5596 + 0.0133X_4 + 0.0107X_5 \\ + 0.6213X_{22} - 0.0571X_{24} \\ + 0.4053X_{25}$$

<u>Model Validation</u>

To make the final decision about the model, it has to be validated. First, the model should be tested for goodness of fit before it is accepted to use based on the current data. If it is decided to be appropriate, then the model should be tested using different set of data to confirm the consistency of the model.

Goodness of Fit Test

Two types of goodness of fit test are suitable for a logistic response function: the Pearson chi-square and the deviance tests.

If the model is adequate, both the Pearson chi-square statistic and the deviance divided by degrees of freedom are expected to be close to 1.[8] To calculate this, `PROC LOGISTIC` with `SCALE=NONE` and `AGGREGATE` option was applied (see `CODE 20`).

`OUTPUT 20`

Deviance and Pearson Goodness-of-Fit Statistics				
Criterion	**Value**	**DF**	**Value/DF**	**Pr > ChiSq**
Deviance	2322.5114	1849	1.2561	<.0001
Pearson	1857.8652	1849	1.0048	0.4378

Unfortunately, as seen in `OUTPUT 20`, the deviance goodness of fit test strongly indicates that the model is not adequate (not to be confused; large deviance means that the fitted model is incorrect). On the other hand, the Pearson chi-square statistic indicates the model is very well-fit. This extreme contradiction indicates that the data is too sparse to use either statistic and therefore the *p*-values are not valid and should be ignored.[8] This is most likely because of continuous predictors (X_4, X_5, and X_{24}) which violate an assumption of these statistics;to use either of these statistics, there should be sufficiently many replicates in each X's. This is very unlikely since this model has continuous predictors. Thus, this overdispersion should not be corrected and should be left as is.

Instead, the Homer-Lemeshow goodness of fit test is more appropriate. `PROC LOGISTIC` with `LACKFIT` option conveniently returns the result (see `CODE 21`).

`OUTPUT 21a`

Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	**DF**	**Pr > ChiSq**
9.9110	8	0.2713

Again, the null hypothesis is where the model is adequate; a large *p*-value implies the model is well-fit. Fortunately, `OUTPUT 21a` indicates that the fitted model is satisfactory, and one concludes that the model is appropriate based on the current data.

Collection of New Data

If possible, the best way to check the model is by collecting a new set of data.[3] The database of NCDC has the climatological data of New York from January, 2005 (the current data set starts from 2008). For new data, the data from January, 2005 to December, 2007 was collected in the same way and imported to SAS (see `CODE 22`).

`PROC LOGISTIC` with same parameters were applied to compare point estimates, standard errors, and chi-square statistics (see `CODE 23`). If the results are relatively close, then the model is applicable under broader circumstances.[3]

`OUTPUT 23`

Analysis of Maximum Likelihood Estimates					
Parameter	**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept	1	0.5595	0.1961	8.1445	0.0043
X4	1	-0.0133	0.00776	2.9202	0.0875
X5	1	-0.0107	0.00307	12.1417	0.0005
X22	1	-0.6213	0.3267	3.6166	0.0572
X24	1	0.0571	0.00991	33.2203	<.0001
X25	1	-0.4052	0.1079	14.0987	0.0002

Analysis of Maximum Likelihood Estimates					
Parameter	**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept	1	0.3923	0.2619	2.2441	0.1341
x4	1	-0.0132	0.00970	1.8481	0.1740
x5	1	-0.0117	0.00414	8.0074	0.0047
x22	1	-0.9462	0.5416	3.0523	0.0806
x24	1	0.0836	0.0136	37.8207	<.0001
x25	1	-0.5549	0.1482	14.0219	0.0002

For convenience, the output from the original data is shown above that from the new data. The signs of parameter estimates are the same but standard errors are substantially larger in the validation set. Because of this larger standard errors, the Wald chi-square statistics became smaller, which made their *p*-values higher, which made the parameters less significant.

$$\chi^2 = Z^{*2} = \frac{b_k}{s\{b_k\}}$$

However, with smaller sample size (n = 1094), these are expected. The general comparisons *among* the variables remain the same. Therefore, informally, the fitted model is decided to be adequate.

Results and Summary

The ultimately accepted fitted model is the following:

$$\mathrm{E}\{Y\} = [1 + \exp(0.5596 - 0.0133X_4 - 0.0107X_5 - 0.6213X_{22} + 0.0571X_{24} - 0.4053X_{25})]^{-1}$$

, where X_4 is departure from normal in ºF, X_5 is average dew point in ºF, X_{22} indicates if freezing was observed (1 = True), X_{24} is daily temperature range in ºF, and X_{25} indicates if the given day is among first half of the year (1 = True). The response variable Y indicates if it will either rain or snow on the following day (1 = True).

`OUTPUT 21b`

Odds Ratio Estimates			
Effect	**Point Estimate**	**95% Wald Confidence Limits**	
x4	0.987	0.968	1.006
x5	0.988	0.980	0.996
x22	0.388	0.134	1.122
x24	1.087	1.059	1.117
x25	0.574	0.429	0.768

Instead of actual point estimates of b_k, the odds ratio for each predictors, calculated by $\exp(b_k)$, are more straightforward for simpler interpretation. `OUTPUT 21b` implies that every unit increase of departure from normal, the odds of precipitation decreases by 1 - 0.987 = 0.013 = 1.3 percent. The unit increase of average dew point has almost same effect. If freezing is observed, the odds of precipitation decreases by 1 - 0.388 = 61.2 percent. Each unit increase of the difference between maximum and minimum temperature (temperature range) increases the odds by 8.7 percent, and if it is among first half of the year (January through June), the odds decreases by 42.6 percent. These odds ratios implies effects of each predictors when other predictors are held constant. Or, alternatively, one can input values of each predictors observed on a given day directly into the model to find the probability of precipitation on the following day.

For example, on a hypothetical day, if the departure from normal is 3 ºF, the average dew point is 43 ºF, freezing was not observed, the temperature range is 23 ºF, and the day is of May, the odds of precipitation is:

$$\begin{aligned} \mathrm{E}\{Y\} = [1 + \exp(0.5596 &- 0.0133(3) \\ &- 0.0107(43) - 0.6213(0) \\ &+ 0.0571(23) - 0.4053(1))]^{-1} \\ &= 0.2727 \end{aligned}$$

Thus, one does not expect it would rain on the following day.

Discussion

As mentioned in the beginning, there are countless important factors to predict participation other than the five predictors in the model. In fact, even if every possible various data are collected on a given day, they might not be enough for spot-on forecast, since the actual precipitation may have something to do more than a set of five surface weather measurements observed on a single given day. For example, recent state and condition of atmosphere such as movement of clouds are measured 7 to 10 miles above the ground, by radiosondes (weather balloons). These values themselves are hard to interpret and not easily obtained.

Note that the actual models that weather prediction centers use nowadays even include absorption and reflection of solar radiation and infrared radiation, and how the atmosphere changes are calculated at every point in a 3D grid of points.[9] Presumably, these models are far more complicated than a single line of logistic regression model with five predictors and have to be run by super computers for their high complexibility. And more importantly, even with these enormous amount of data and super computers, weather forecasts often are wrong and many people always doubt the weather forecasts. Edward Lorenz, the pioneer of butterfly effect and chaos theory, published a paper *Deterministic Nonperiodic Flow* in Journal of the Atmospheric Sciences, in which he stated that most statistical models in meteorology are not appropriate.

The fitted model in this paper should be used as a basic reference of by how much *some easily-accessible* variables affect precipitation, and it should not be overgeneralized to precisely calculate precipitation probabilities.

References

[1] "Answers: Understanding weather forecasts." *USA TODAY*[Tysons Corner] 02 08 2006, Weather n. pag. Web. 23 Apr. 2013.

[2] D. A. Belsley, E. Kuh, and R. E. Welsch. Regression Diagnostics: Identifying Influential Data and Sources of Colinearity, 1980, Wiley, New York.

[3] Kutner, Michael H., Chris Nachtsheim, and John Neter. *Applied linear regression models*. Boston New York: McGraw-Hill/Irwin, 2004. Print.

[4] Allison, Paul D. *Logistic regression using the SAS system : theory and application*. Cary, N.C: SAS Institute, 1999. Print.

[5] Shtatland, Ernest S., Ken Kleinman, and Emily M. Cain. "STEPWISE METHODS IN USING SAS® PROC LOGISTIC AND SAS® ENTERPISE MINERTM FOR PREDICTION." *SAS Users Group International Proceedings 28*. Seattle: SAS, 2003. Web. 9 May. 2013. <http://www2.sas.com/proceedings/sugi28/258-28.pdf>.

[6] Lottes, Ilsa L., Alfred DeMaris, and Marina A. Adler. "Using and Interpreting Logistic Regression: A Guide for Teachers and Students." *Teaching Sociology*. 24.3 (1996): 284-98. Print.

[7] "Stata FAQ: How Can I Understand a Continuous by Continuous Interaction in Logistic Regression?." *UCLA: Statistical Consulting Group*. Institute for Digital Research and Education. Web. 15 May 2013. <http://www.ats.ucla.edu/stat/mult_pkg/faq/general/citingats.htm>.

[8] SAS Institute Inc. 2011. SAS/STAT® 9.3 User's Guide. Cary, NC: SAS Institute Inc.

[9] "How are weather forecasts made?." *WeatherQuestions*. Weather Street, 1 Jan 2011. Web. 15 May 2013. <http://www.weatherquestions.com/How_are_weather_forecasts_made.htm>.

APPENDIX A: A sample of raw data extracted from National Oceanic and Atmospheric Administration (NOAA)'s National Climatic Data Center (NCDC)

LCD Daily Form

QUALITY CONTROLLED LOCAL CLIMATOLOGICAL DATA

(final)

NOAA, National Climatic Data Center

Month: 07/2009

Station Location: CENTRAL PARK (94728)
NEW YORK, NY

Lat. 40.778 Lon. -73.969

Elevation(Ground): 130 ft. above sea level

Date	Temperature (Fahrenheit) Max.	Min.	Avg.	Dep From Normal	Avg. Dew pt.	Avg Wet Bulb	Degree Days Base 65 Degrees Heating	Cooling	Sun Sunrise LST	Sunset LST	Significant Weather	Snow/Ice on Ground(In) 1200 UTC Depth	1800 UTC Water Equiv	Precipitation (In) 2400 LST Snow Fall	2400 LST Water Equiv	Pressure(inches of Hg) Avg. Station	Avg. Sea Level	Wind: Speed=mph Dir=tens of degrees Resultant Speed	Res Dir	Avg. Speed	max 5-second Speed	max 5-second Dir	max 2-minute Speed	max 2-minute Dir	Date
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26
01	81	66	74	-1	65	68	0	9	0428	1931	BR HZ	0	M	0.0	0.00	29.57	29.72	3.0	07	5.2	22	060	15	050	01
02	76	66	71	-4	67	68	0	6	0429	1931	RA FG BR	0	M	0.0	0.60	29.60	29.75	2.9	06	4.2	15	070	12	100	02
03	78	65	72	-3	61	65	0	7	0429	1931		0	M	0.0	0.00	29.68	29.83	2.1	27	5.0	21	280	14	270	03
04	79	66	73	-2	55	62	0	8	0430	1931		0	M	0.0	0.00	29.70	29.85	5.0	27	7.2	22	280	14	260	04
05	79	61	70	-5	50	59	0	5	0430	1930		0	M	0.0	0.00	29.72	29.86	1.5	27	4.5	18	210	13	240	05
06	83	63	73	-2	54	62	0	8	0431	1930		0	M	0.0	0.00	29.62	29.76	0.5	16	3.7	18	240	13	220	06
07	79	61	70	-5	61	64	0	5	0432	1930	RA BR	0	M	0.0	0.13	29.60	29.74	1.9	04	5.2	23	330	14	060	07
08	77	58*	68	-8	55	60	0	3	0432	1930		0	M	0.0	0.00	29.75	29.89	1.0	35	5.0	17	300	14	060	08
09	73	61	67*	-10	56	60	0	2	0433	1929		0	M	0.0	0.00	30.03	30.19	6.4	07	7.8	17	060	14	060	09
10	78	60	69	-8	55	60	0	4	0434	1929		0	M	0.0	0.00	30.14	30.28	0.4	05	4.4	17	130	12	140	10
11	78	63	71	-6	58	63	0	6	0434	1929	RA BR	0	M	0.0	0.33	29.95	30.12	2.1	19	5.5	21	240	13	210	11
12	81	64	73	-4	54	62	0	8	0435	1928	BR	0	M	0.0	0.02	29.78	29.91	1.9	28	5.0	21	310	13	290	12
13	79	61	70	-7	51	59	0	5	0436	1928		0	M	0.0	0.00	29.77	29.91	0.1	05	3.1	18	150	12	160	13
14	79	61	70	-7	48	58	0	5	0436	1927		0	M	0.0	0.00	29.88	30.03	2.2	27	5.0	22	270	16	280	14
15	82	63	73	-4	55	62	0	8	0437	1927		0	M	0.0	0.00	29.93	30.07	0.9	25	3.5	21	240	14	260	15
16	84	70	77	0	65	69	0	12	0438	1926		0	M	0.0	0.00	29.67	29.82	1.8	23	4.8	23	230	14	210	16
17	86*	68	77	0	68	71	0	12	0439	1925	RA BR HZ	0	M	0.0	0.21	29.64	29.77	1.5	12	4.2	16	160	12	160	17
18	83	68	76	-1	61	66	0	11	0440	1925	BR	0	M	0.0	0.02	29.61	29.80	3.5	26	6.4	20	260	14	250	18
19	81	64	73	-4	56	62	0	8	0440	1924		0	M	0.0	0.00	29.92	30.07	1.3	26	4.7	17	210	13	230	19
20	81	67	74	-3	60	65	0	9	0441	1923		0	M	0.0	0.00	30.01	30.15	1.6	10	4.4	15	170	10	170	20
21	71	64	68	-9	65	65	0	3	0442	1923	RA BR	0	M	0.0	1.14	29.97	30.11	4.7	05	6.2	22	060	15	060	21
22	82	64	73	-4	66	68	0	8	0443	1922	BR	0	M	0.0	0.00	29.98	30.13	0.1	24	2.9	15	190	10	170	22
23	77	64	71	-6	66	67	0	6	0444	1921	RA BR	0	M	0.0	0.39	29.84	29.98	5.5	07	7.7	30	080	21	060	23
24	78	63	71	-6	64	67	0	6	0445	1920		0	M	0.0	0.02	29.72	29.87	0.9	25	4.6	15	160	9	160	24
25	83	66	75	-2	67	69	0	10	0446	1919		0	M	0.0	0.00	29.78	29.92	2.2	15	4.8	20	170	13	150	25
26	84	67	76	-1	69	71	0	11	0446	1919	RA BR	0	M	0.0	1.42	29.75	29.89	1.5	16	4.7	23	230	14	230	26
27	83	68	76	-1	70	71	0	11	0447	1918	RA BR	0	M	0.0	0.60	29.78	29.92	1.1	23	3.8	18	160	12	200	27
28	85	71	78	1	69	72	0	13	0448	1917		0	M	0.0	0.00	29.79	29.93	1.3	17	2.9	20	180	12	170	28
29	83	70	77	0	72	73	0	12	0449	1916	RA BR	0	M	0.0	1.52	29.69	29.84	2.0	20	4.8	30	220	18	210	29
30	85	72	79*	2	67	71	0	14	0450	1915		0	M	0.0	0.00	29.69	29.84	0.9	25	5.4	16	250	10	250	30
31	85	67	76	-1	70	71	0	11	0451	1914	RA BR	0	M	0.0	0.71	29.72	29.87	0.8	25	4.6	24	330	15	240	31
	80.4	64.9	72.7		61.3	65.5	0.0	7.9	<-----Monthly Averages | Totals----->				M	0.0	7.11	29.78	29.93	0.1	12	4.9	<Monthly Average				
	-3.7	-3.9	-3.8	<-----------Departure From Normal----------->											2.51										

Degree Days Monthly Season to Date

	Monthly Total	Monthly Departure	Season to Date Total	Season to Date Departure
Heating:	0	0	0	0
Cooling:	246	-109	447	-189

Greatest 24-hr Precipitation: 1.90 Date: 26-27
Greatest 24-hr Snowfall: 0.0 Date: M
Greatest Snow Depth: 0 Date: M

Sea Level Pressure		Date	Time (LST)
Maximum	30.32	10	0851
Minimum	29.65	01	0307

Number of Days with -------->

Max Temp >=90: 0
Max Temp <=32: 0
Thunderstorms : 0

Min Temp <=32: 0
Min Temp <=0 : 0
Heavy Fog : 0

Precipitation >=.01 inch: 13
Precipitation >=.10 inch: 10
Snowfall >=1.0 inch : 0

* EXTREME FOR THE MONTH - LAST OCCURRENCE IF MORE THAN ONE.

Data Version: VER3

http://cdo.ncdc.noaa.gov/qclcd/QCLCD[5/15/2013 8:02:53 PM]

APPENDIX B: SAS codes

NOTE: The code numbers corresponds to the output numbers
Codes that omit some part of lines due to their lengths are indicated with +

CODE 1+

```
data final;
input X1     X2    X3    X4    X5    X6    X7    X8    X9    X10   X11   X12
 X13   X14   X15   X16   X17   X18   X19   X20   X21   X22   X23   X24   X25
 Y;
datalines;
47     37    42    9     30    37    23    0     29.73 29.83 2.2   20    6.4
 25    240   16    150   0     0     1     0     0     1     10    1     1
38     17    28    -5    11    23    37    0     29.7  29.92 7     31    9.2
 32    290   18    300   0     0     0     0     0     1     21    1     0
20     12    16    -17   -2    13    49    0     30.37 30.56 3.7   30    6.4
 29    310   15    300   0     0     0     0     0     0     8     1     0

…

55     40    48    2     32    40    17    0     29.96 30.06 6.7   28    5.6
 24    340   15    310   0     0     0     0     0     0     15    1     0
59     40    50    3     32    40    15    0     29.97 30.08 4.5   28    4.2
 17    280   12    300   0     0     0     0     0     0     19    1     1
;
run;
```

CODE 2

```
proc logistic data=final;
model y=X1   X2    X3    X4    X5    X6    X7    X8    X9    X10   X11   X12
 X13   X14   X15   X16   X17   X18   X19   X20   X21   X22   X23   X24   X25;
run;
```

CODE 3

```
proc logistic data=final;
model y=X3   X4    X5    X6    X9    X10   X11   X12   X13   X14   X15   X16
 X17   X18   X19   X20   X21   X22   X23   X24   X25;
run;
```

CODE 4

```
ods graphics on;
proc logistic data=final plots (only label)=(influence);
model y=X3 X4 X5 X6 X10 X16 X19 X20 X21 X22 X24 X25/influence iplots;
run;
ods graphics off;
```

CODE 5

```
proc iml;
edit final;
delete point 452;
delete point 1491;
```

CODE 6

```
proc corr data=final;
var X9 X10   X11 X13 X14 X16 X12 X15 X17;
run;
```

CODE 7

```
proc reg data=final;
model y=X3   X4    X5    X6    X10   X16   X19   X20   X21   X22   X24
 X25/vif;
run;
```

CODE 8

```
proc reg data=final;
model y=X4   X5    X10   X16   X19   X20   X21   X22   X23   X24   X25/vif;
run;
```

CODE 9

```
proc logistic data=final;
model y=X4 X5 X10 X16 X19 X20 X21 X22 X24 X25/selection=stepwise sle=0.99
sls=0.99;
ods output FitStatistics=FIT;
run;
```

CODE 10

```
proc print data=FIT;
run;
```

CODE 11

```
proc logistic data=final;
model y=X4 X5 X10 X16 X19 X20 X21 X22 X24 X25/selection=f sle=0.05;
run;
```

CODE 12

```
proc logistic data=final;
model y=X4 X5 X10 X16 X19 X20 X21 X22 X24 X25/selection=b sls=0.05;
run;
```

CODE 13

```
proc logistic data=final;
model y=X4 X5 X10 X16 X19 X20 X21 X22 X24 X25/selection=stepwise sle=0.05
sls=0.05;
run;
```

CODE 14

```
proc logistic data=final;
model y=X4 X5 X10 X16 X19 X21 X22 X24 X25/selection=stepwise sle=0.99
sls=0.99;
ods output FitStatistics=FIT;
run;

proc print data=FIT;
run;


proc logistic data=final;
model y=X4 X5 X10 X16 X19 X21 X22 X24 X25/selection=f sle=0.05;
run;

proc logistic data=final;
model y=X4 X5 X10 X16 X19 X21 X22 X24 X25/selection=b sls=0.05;
run;

proc logistic data=final;
model y=X4 X5 X10 X16 X19 X21 X22 X24 X25/selection=stepwise sle=0.05
sls=0.05;
run;


proc logistic data=final;
model y=X4 X5 X10 X16 X19 X20 X21 X22 X25/selection=stepwise sle=0.99
sls=0.99;
ods output FitStatistics=FIT;
run;

proc print data=FIT;
run;


proc logistic data=final;
model y=X4 X5 X10 X16 X19 X20 X21 X22 X25/selection=f sle=0.05;
run;

proc logistic data=final;
model y=X4 X5 X10 X16 X19 X20 X21 X22 X25/selection=b sls=0.05;
run;
```

```
proc logistic data=final;
model y=X4 X5 X10 X16 X19 X20 X21 X22 X25/selection=stepwise sle=0.05
sls=0.05;
run;

proc logistic data=final;
model y=X4 X5 X10 X16 X19 X21 X22 X25/selection=stepwise sle=0.99 sls=0.99;
ods output FitStatistics=FIT;
run;

proc print data=FIT;
run;


proc logistic data=final;
model y=X4 X5 X10 X16 X19 X21 X22 X25/selection=f sle=0.05;
run;


proc logistic data=final;
model y=X4 X5 X10 X16 X19 X21 X22 X25/selection=b sls=0.05;
run;

proc logistic data=final;
model y=X4 X5 X10 X16 X19 X21 X22 X25/selection=stepwise sle=0.05 sls=0.05;
run;
```

CODE 15

```
proc logistic data=final;
model y=X4 X5 X24 X25;
run;
```

CODE 16

```
proc logistic data=final;
model y=X4 X5 X16 X22 X24 X25;
run;
```

CODE 17

```
proc logistic data=final;
model y=X4 X5 X22 X24 X25;
run;
```

CODE 18

```
proc means data=final;
run;

data final;
set final;
X4C=X4-1.4869383;
X4CSQ=X4C**2;
```

```
X5C=X5-41.2021944;
X5CSQ=X5C**2;
X24C=X24-14.2001045;
X24CSQ=X24C**2;
run;

proc logistic data=final;
model y=X4 X5 X22 X24 X25 X4csq X5csq X24csq;
run;
```

CODE 19

```
data final;
set final;
X4X5=X4*X5;
X4X24=X4*X24;
X5X24=X5*X24;
run;

proc insight data=final;
run;
```

CODE 20

```
proc logistic data=final;
model y=X4 X5 X22 X24 X25/scale=none aggregate;
run;
```

CODE 21

```
proc logistic data=final;
model y=X4 X5 X22 X24 X25/lackfit;
run;
```

CODE 22+

```
data VLD;
input X4 X5 X22 X24 X25 y;
datalines;
17      36    0     20    1     0
11      32    0     13    1     1
21      46    0     9     1     1

…

11      31    0     9     0     1
13      37    0     9     0     1
528     0     8     0     1
;
run;
```

CODE 23

```
proc logistic data=VLD;
model y=X4 X5 X22 X24 X25;
run;
```