Analysis on Ozone Level Detection

Yang Yifeng

yyangbf@ust.hk

Introduction

The environment issue has always been a popular topic with the vicissitude of daily life and of great concern among the public. Ozone is a common kind of toxic gas existing in the atmosphere, which also plays an irreplaceable role in preventing ultraviolet light from outer space. A variety of sophisticated methods has been employed on analyzing and forecasting ozone level in different regions of the world. In this report, we utilized the traditional logit regression model to analyze the ozone level from 1/1/1998 to 1/26/2004.

According to Texas Commission on Environmental Quality (TCEQ), the following attributes (see Table 1) are the most important factors influencing ozone level in the atmosphere.

Table 1. Factors influencing ozone level

O3	Local ozone peak prediction
Upwind	Upwind ozone background level
EmFactor	Precursor emissions related factor
Tmax	Maximum temperature in degrees F
Tb	Base temperature where net ozone production begins (50 F)
SRd	Solar radiation total for the day
Wsa	Wind speed near sunrise (using 09-12 UTC forecast mode)
WSp	Wind speed mid-day (using 15-21 UTC forecast mode)

Our target is to build a statistical model to predict whether it is an ozone day or not based on the factors listed above. The data collected are ungrouped. The response is binary. “1” means ozone day and “0” means normal day. 2536 samples were collected from 1/1/1998 to 1/26/2004. 73 attributes are considered as independent variables in the regression model, which are listed below in Table 2.

Table 2. Attributes used in the regression model

WSR0	WSR6	WSR12	WSR18	WSR_PK	T4	T10	T16	T22	U85	V70
WSR1	WSR7	WSR13	WSR19	WSR_AV	T5	T11	T17	T23	V85	HT70
WSR2	WSR8	WSR14	WSR20	T0	T6	T12	T18	T_PK	HT85	T50
WSR3	WSR9	WSR15	WSR21	T1	T7	T13	T19	T_AV	T70	RH50
WSR4	WSR10	WSR16	WSR22	T2	T8	T14	T20	T85	RH70	U50
WSR5	WSR11	WSR17	WSR23	T3	T9	T15	T21	RH85	U70	V50
HT50	KI	TT	SLP	SLP_	Precp					

At the beginning, the core advantage of our analysis lies in the affluence of factors in our regression model. Because of the sufficient factors in the model, we were able to achieve satisfactory deviance and Pearson goodness-of-fit statistics. We will show that after considering multi-collinearity and model selection, we were still able to maintain the same level of goodness-of-fit statistics while keeping less variables in the model.

Methodology

It should be pointed out at the beginning that all the attributes used are continuous. Therefore, the model does not involve any interaction terms. Since the response is binary, we use logit regression given that its assumptions are satisfied.

$$\log\left(\frac{P_i}{1-P_i}\right) = \vec{X}\vec{B}$$

First, the logit regression requires the dependent variable to be binary. In our case, the dependent variable can take either 1(ozone day) or 0(normal day).

Second, the logit regression requires the observations to be independent of each other.

Third, there should be little or no multicollinearity among the independent variables. This means that the independent variables should not be too highly correlated with each other. In our case, we eliminated multicollinearity by calculating variance inflation factor(VIF).

The first step was to test whether there is lack of fit in the logit model. The common measures of lack-of-fit are the deviance goodness-of-fit statistics and Pearson goodness-of-fit statistics. A high p-value indicates little lack-of-fit in the model. We also tested the level of agreement between the predicted probabilities and observed responses. Common measures include concordant & discordant, Somer's D, Tau-a, Gamma and ROC. They are defined as:

Somers′ D = (nc − nd)/t; ranging from -1(all pairs disagree)to 1(all pairs agree)

Gamma = (nc − nd)/(nc + nd); ranging from -1(no association) to 1(perfect association)

Tau − a = (nc − nd)/(0.5N(N − 1))

ROC = (nc + 0.5(t − nc − nd))/t; ranging from 0.5(model randomly predicts the response) to 1(model perfectly discriminates the response)

To satisfy the third assumption of logit model, we needed to eliminate the multicollinearity among the independent variables. A common measure of multicollinearity is the variance inflation factor(VIF), which can be calculated as following:

$$VIF_i = \frac{1}{1-R_i^2}$$

We adopted a common cut-off value of 5. Namely, if VIF(X_i) >5, then multicollinearity is high and the attribute with highest multicollinearity should be deleted. We repeated this procedure until all VIF were smaller than 5.

Unusual points in the sample may influence the regression results significantly. Therefore, detecting unusual points in the model is crucial in achieving accurate estimation. We evaluated CBAR, DIFCHISQ and DIFDEV to examine whether a certain observation was unusual.

They are defined as:

$$\overline{C_J} = C_j(1 - h_{ii})$$

$$\text{DIFCHISQ} = \frac{\overline{C_J}}{h_{ii}}$$

$$\text{DIFDEV} = d_j^2 + \overline{C_J}$$

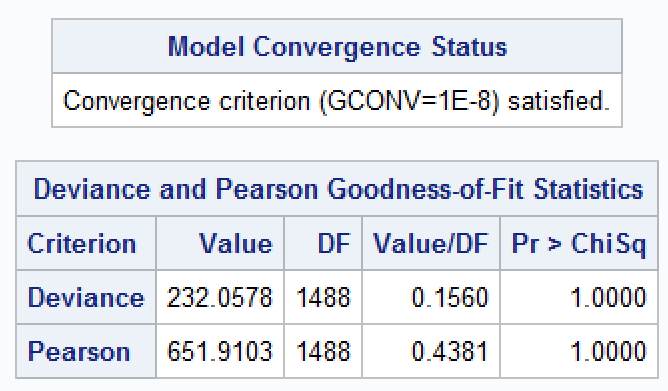
We chose a cut-off value for each measure separately based on the statistics calculated and deleted those observations exceeding the cut-off values.

Model selection was performed to achieve a succinct model. Forward selection, backward selection and stepwise selection were performed to determine the best model. Then we chose the model which most selection methods agreed on. After obtaining the best model, we needed to add back the observations deleted previously and found new unusual points again.

Results

The first step was to test whether there is lack of fit in the logit model. Deviance goodness-of-fit statistics and Pearson goodness-of-fit statistics were calculated (see Figure 1).

Figure 1. Deviance goodness-of-fit statistics and Pearson goodness-of-fit statistics



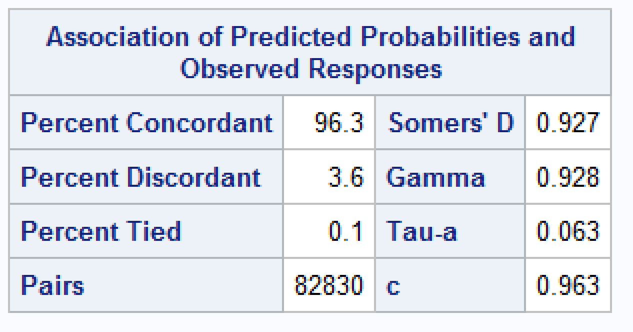
Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Deviance and Pearson Goodness-of-Fit Statistics				
Criterion	Value	DF	Value/DF	Pr > ChiSq
Deviance	232.0578	1488	0.1560	1.0000
Pearson	651.9103	1488	0.4381	1.0000

Notice that the p-value for both statistics are 1, meaning that there is almost no lack-of-fit in the model.

In order to test the level of agreement between the predicted probabilities and observed responses, we examined Somer's D, Tau-a, Gamma and ROC. See Figure 2.

Firgure 2. Different measures of agreement

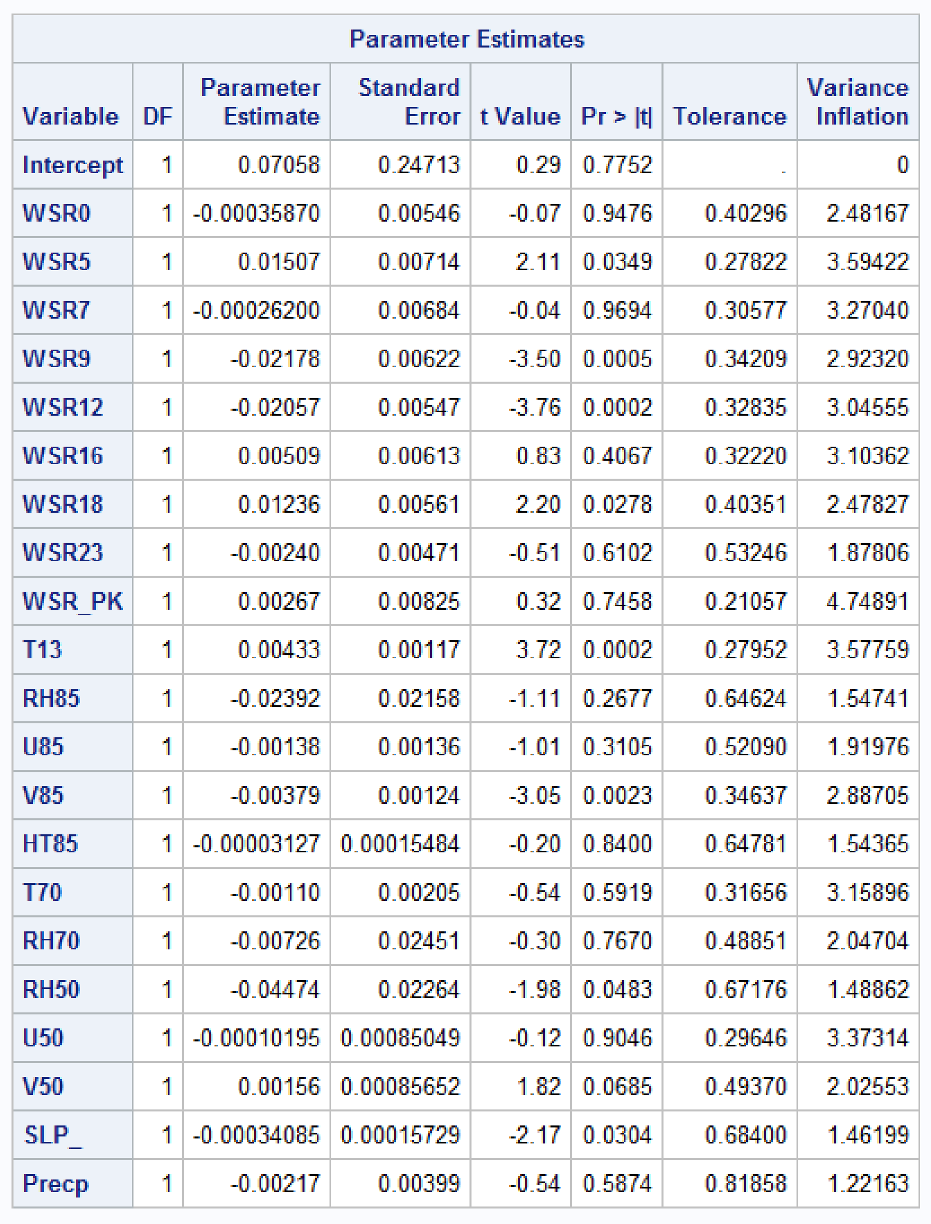


Association of Predicted Probabilities and Observed Responses			
Percent Concordant	96.3	Somers' D	0.927
Percent Discordant	3.6	Gamma	0.928
Percent Tied	0.1	Tau-a	0.063
Pairs	82830	c	0.963

Notice that Somer'D, Gamma and ROC are very close to one, indicating a high level of agreement between the predicted probabilities and observed responses. Moreover, we also notice that the percent concordant is significantly higher than percent discordant, which manifests the excellent performance of the model.

To satisfy the third assumption of logit model, we then eliminated the multicollinearity among the independent variables by calculating variance inflation factor(VIF). We deleted the attribute with highest multicollinearity until all VIF were smaller than 5. See Figure 3.

Figure 3. VIF

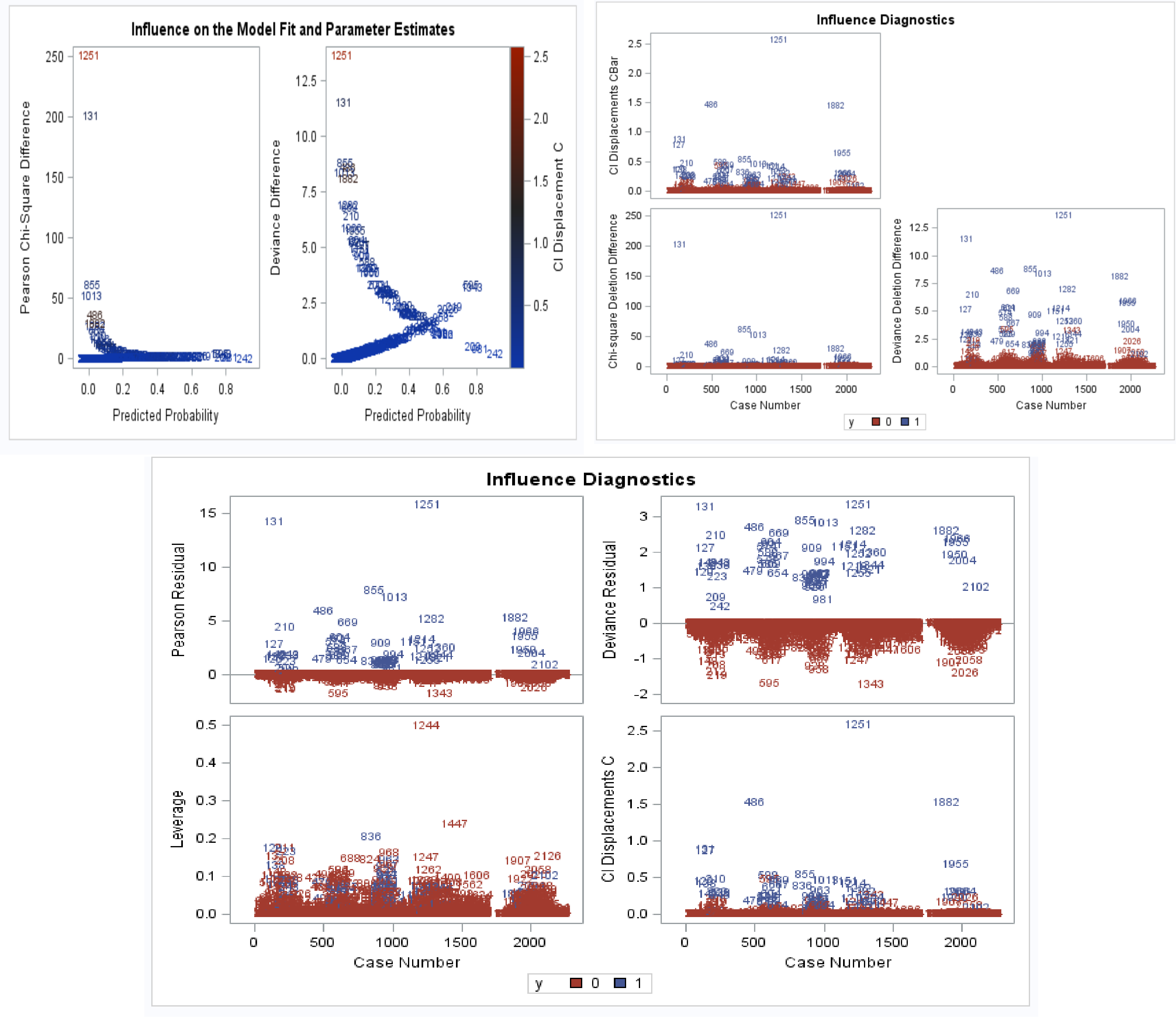


Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Tolerance	Variance Inflation
Intercept	1	0.07058	0.24713	0.29	0.7752	.	0
WSR0	1	-0.00035870	0.00546	-0.07	0.9476	0.40296	2.48167
WSR5	1	0.01507	0.00714	2.11	0.0349	0.27822	3.59422
WSR7	1	-0.00026200	0.00684	-0.04	0.9694	0.30577	3.27040
WSR9	1	-0.02178	0.00622	-3.50	0.0005	0.34209	2.92320
WSR12	1	-0.02057	0.00547	-3.76	0.0002	0.32835	3.04555
WSR16	1	0.00509	0.00613	0.83	0.4067	0.32220	3.10362
WSR18	1	0.01236	0.00561	2.20	0.0278	0.40351	2.47827
WSR23	1	-0.00240	0.00471	-0.51	0.6102	0.53246	1.87806
WSR_PK	1	0.00267	0.00825	0.32	0.7458	0.21057	4.74891
T13	1	0.00433	0.00117	3.72	0.0002	0.27952	3.57759
RH85	1	-0.02392	0.02158	-1.11	0.2677	0.64624	1.54741
U85	1	-0.00138	0.00136	-1.01	0.3105	0.52090	1.91976
V85	1	-0.00379	0.00124	-3.05	0.0023	0.34637	2.88705
HT85	1	-0.00003127	0.00015484	-0.20	0.8400	0.64781	1.54365
T70	1	-0.00110	0.00205	-0.54	0.5919	0.31656	3.15896
RH70	1	-0.00726	0.02451	-0.30	0.7670	0.48851	2.04704
RH50	1	-0.04474	0.02264	-1.98	0.0483	0.67176	1.48862
U50	1	-0.00010195	0.00085049	-0.12	0.9046	0.29646	3.37314
V50	1	0.00156	0.00085652	1.82	0.0685	0.49370	2.02553
SLP_	1	-0.00034085	0.00015729	-2.17	0.0304	0.68400	1.46199
Precp	1	-0.00217	0.00399	-0.54	0.5874	0.81858	1.22163

We then calculated CBAR, DIFCHISQ and DIFDEV to detect unusual points. See Figure 4 for the diagnostic plots.

Figure 4. Diagnostic plots

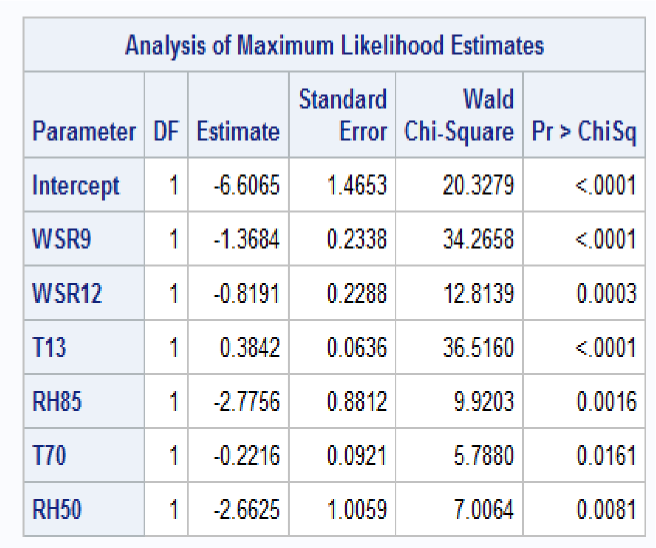


Based on the diagnostic plots, we chose our cut-off values for each measure. We deleted an observation if its |DIFDEV |>7.5 or |DIFCHISQ |>50 or |CBAR|> 0.5. Only 8 observations were deleted.

Forward selection, backward selection and stepwise selection were performed to determine the best model. We set both the entry significance level and staying significance level to be 0.05. The results are shown in Figure 5.

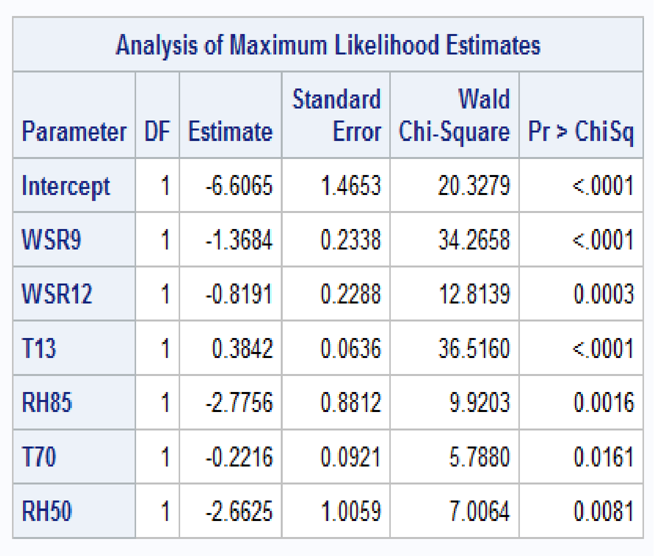
Figure 5. Model selection results

(a)forward selection



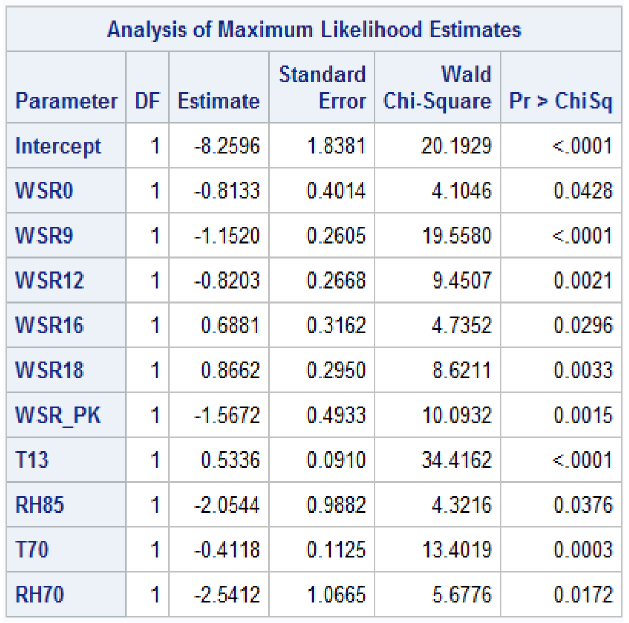
Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-6.6065	1.4653	20.3279	<.0001
WSR9	1	-1.3684	0.2338	34.2658	<.0001
WSR12	1	-0.8191	0.2288	12.8139	0.0003
T13	1	0.3842	0.0636	36.5160	<.0001
RH85	1	-2.7756	0.8812	9.9203	0.0016
T70	1	-0.2216	0.0921	5.7880	0.0161
RH50	1	-2.6625	1.0059	7.0064	0.0081

(b) stepwise selection



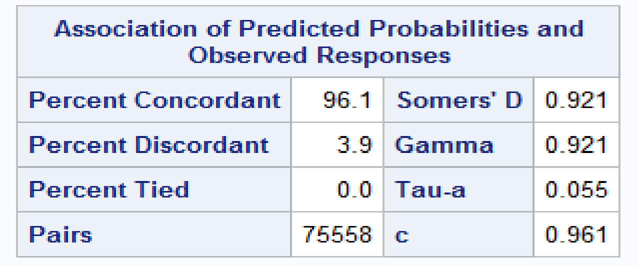
Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-6.6065	1.4653	20.3279	<.0001
WSR9	1	-1.3684	0.2338	34.2658	<.0001
WSR12	1	-0.8191	0.2288	12.8139	0.0003
T13	1	0.3842	0.0636	36.5160	<.0001
RH85	1	-2.7756	0.8812	9.9203	0.0016
T70	1	-0.2216	0.0921	5.7880	0.0161
RH50	1	-2.6625	1.0059	7.0064	0.0081

(c) backward selection



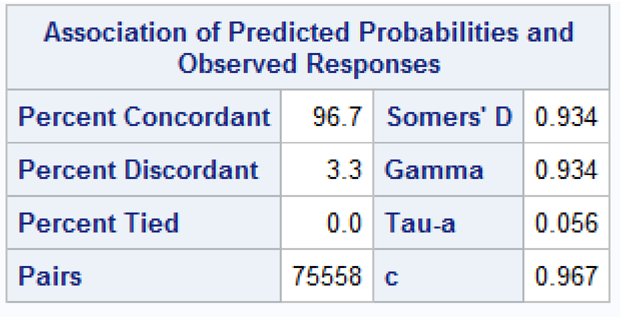
Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-8.2596	1.8381	20.1929	<.0001
WSR0	1	-0.8133	0.4014	4.1046	0.0428
WSR9	1	-1.1520	0.2605	19.5580	<.0001
WSR12	1	-0.8203	0.2668	9.4507	0.0021
WSR16	1	0.6881	0.3162	4.7352	0.0296
WSR18	1	0.8662	0.2950	8.6211	0.0033
WSR_PK	1	-1.5672	0.4933	10.0932	0.0015
T13	1	0.5336	0.0910	34.4162	<.0001
RH85	1	-2.0544	0.9882	4.3216	0.0376
T70	1	-0.4118	0.1125	13.4019	0.0003
RH70	1	-2.5412	1.0665	5.6776	0.0172

(d)forward &stepwise selection



Association of Predicted Probabilities and Observed Responses			
Percent Concordant	96.1	Somers' D	0.921
Percent Discordant	3.9	Gamma	0.921
Percent Tied	0.0	Tau-a	0.055
Pairs	75558	c	0.961

(e) backward selection

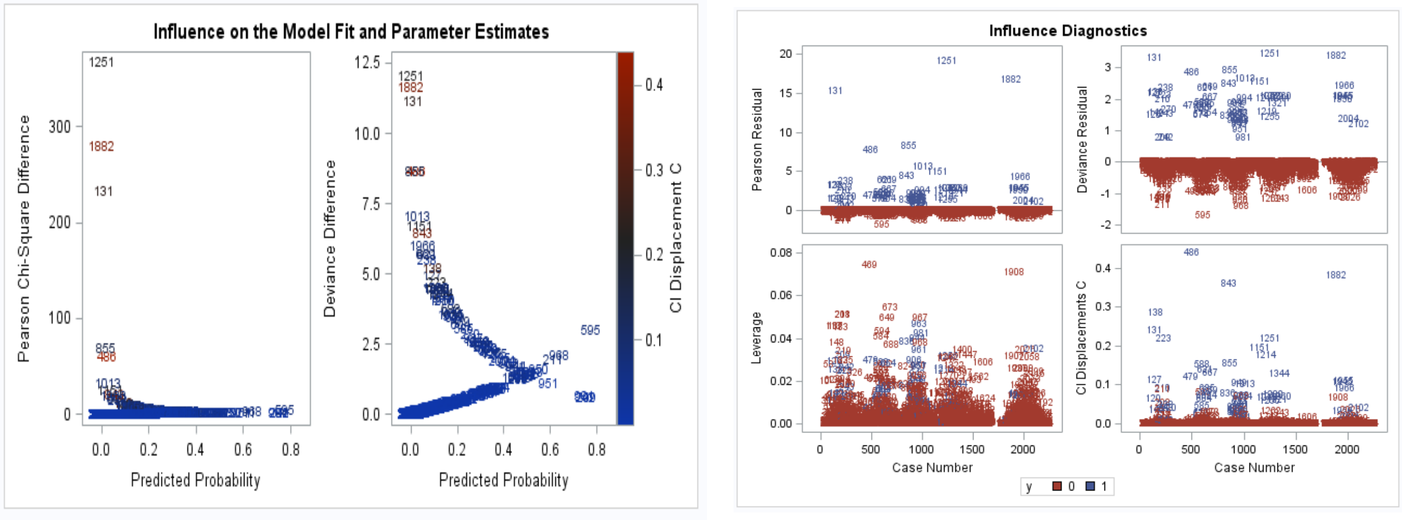


Association of Predicted Probabilities and Observed Responses			
Percent Concordant	96.7	Somers' D	0.934
Percent Discordant	3.3	Gamma	0.934
Percent Tied	0.0	Tau-a	0.056
Pairs	75558	c	0.967

Notice that forward selection and stepwise selection gave the same model, which had less independent variables than that given by backward selection. In addition, the model given by forward/stepwise selection achieved almost the same level of association between predicted probabilities and observed responses as the model given by backward selection. Therefore, we chose the model given by forward/stepwise selection as our final model. Notice that there were only 6 attributes in the model, which was a dramatic simplification of our original model containing 73 attributes.

Since our model was reduced, we needed to add back the observations deleted previously. Moreover, we needed to delete unusual points again. See Figure 6 for diagnostic plots.

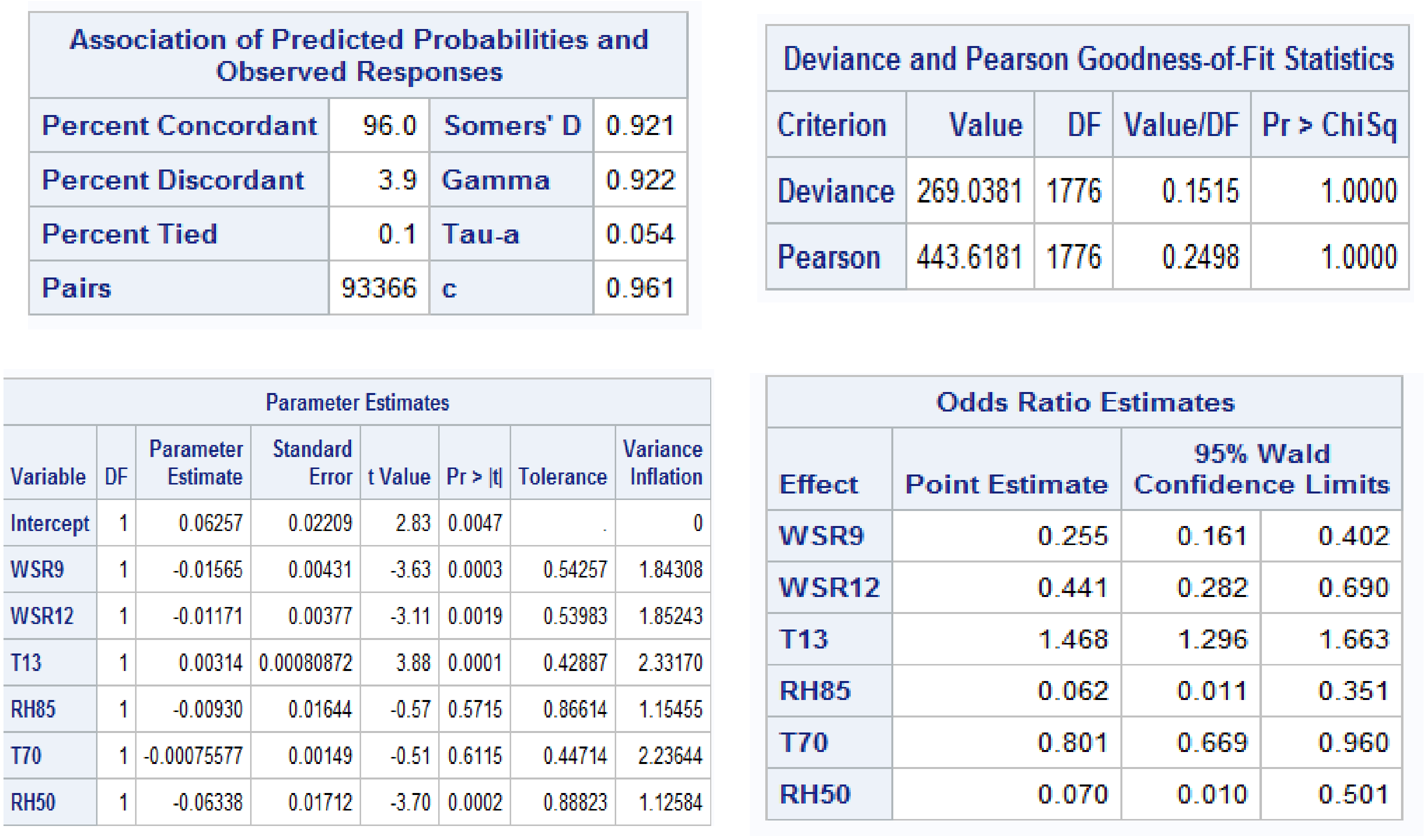
Figure 6. Diagnostic plots of final model



Same as before, we chose our cut-off values for each measure based on the diagnostic plots. We deleted an observation if its |DIFDEV |>7.5 or |DIFCHISQ |>50 or |CBAR|> 0.2. It should be pointed out that the same 8 observations were deleted again.

Finally, we fit the logit model again and calculated goodness-of-fit statistics, Somer's D, Tau-a, Gamma, ROC. We also present the VIF for the remaining 6 attributes. See Figure 7.

Figure 7. Various statistics



Association of Predicted Probabilities and Observed Responses

Percent Concordant	96.0	Somers' D	0.921
Percent Discordant	3.9	Gamma	0.922
Percent Tied	0.1	Tau-a	0.054
Pairs	93366	c	0.961

Deviance and Pearson Goodness-of-Fit Statistics

Criterion	Value	DF	Value/DF	Pr > ChiSq
Deviance	269.0381	1776	0.1515	1.0000
Pearson	443.6181	1776	0.2498	1.0000

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Tolerance	Variance Inflation
Intercept	1	0.06257	0.02209	2.83	0.0047	.	0
WSR9	1	-0.01565	0.00431	-3.63	0.0003	0.54257	1.84308
WSR12	1	-0.01171	0.00377	-3.11	0.0019	0.53983	1.85243
T13	1	0.00314	0.00080872	3.88	0.0001	0.42887	2.33170
RH85	1	-0.00930	0.01644	-0.57	0.5715	0.86614	1.15455
T70	1	-0.00075577	0.00149	-0.51	0.6115	0.44714	2.23644
RH50	1	-0.06338	0.01712	-3.70	0.0002	0.88823	1.12584

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
WSR9	0.255	0.161	0.402
WSR12	0.441	0.282	0.690
T13	1.468	1.296	1.663
RH85	0.062	0.011	0.351
T70	0.801	0.669	0.960
RH50	0.070	0.010	0.501

Notice that the reduced final model still achieved high p-value in goodness-of-fit test. The performance, in terms of Somer's D, Tau-a, Gamma, ROC, did not drop dramatically after deleting most attributes.

We see that all attributes in the final model has odds ratio smaller than 1 except for T13. This means only T13 has a positive relationship with the probability of ozone day. If T13 increases, the probability of ozone day will also increase. Therefore, T13 can serve as an indicator of ozone day. For the other attributes, if the value of attribute increases, the probability of ozone day will decrease. We may also compare the influence of each factor having odds ratio smaller than 1. According to the odds ratio, we can see that RH85 and RH50 have the most significant negative impact on the probability of ozone day while T70 has less influence.

Conclusion

In this report, we utilized the logit regression model to analyze the ozone level from 1/1/1998 to 1/26/2004. Goodness-of-fit statistics, Somer's D, Tau-a, Gamma, ROC were calculated to evaluate the level of lack of fit, association between predicted probabilities and observed responses. VIF were used to examine multicollinearity among the independent variables. Unusual points were deleted by considering CBAR, DIFCHISQ and DIFDEV. Forward selection, backward selection and stepwise selection were performed to determine the best model.

We conclude that T13 has a positive relationship with the probability of ozone day while RH85 and RH50 have the most significant negative impact.

My contribution: I tested the original model. I wrote codes and determined how to delete unusual points, how to set the cut-off values. My partner calculated VIF and deleted attributes one by one, which was labor-intensive.

After that, we did the project on our own. I just gave him my codes on deleting unusual points.

Up to this point, we share the same information.

I then used the remaining attributes to do the model selection and deleted unusual points again.

How to interpret the odds ratio is my own work.