Environments **2014**, *1*, 54-59; doi:10.3390/environments1010054

OPEN ACCESS

environments

ISSN 2076-3298

www.mdpi.com/journal/environments

Article

A Proportional Odds Model of Particle Pollution

Justin R. Chimka * and Ege Ozdemir

Department of Industrial Engineering, University of Arkansas, 800 W. Dickson St., Fayetteville, AR 72701, USA; E-Mail: oege@email.uark.edu

* Author to whom correspondence should be addressed; E-Mail: jchimka@uark.edu; Tel.: +1-479-575-3156; Fax: +1-479-575-8431.

Received: 25 May 2014; in revised form: 4 August 2014 / Accepted: 4 August 2014 / Published: 12 August 2014

Abstract: A linear regression model of particle pollution and an ordered logistic regression model of the relevant index were selected for observations in the US city of Los Angeles, California. Models were used to forecast Air Quality Index (AQI) from a sample, and were compared and contrasted. Methods are comparable overall but markedly different in their powers to predict certain categories. Linear regression models of AQI through particle pollution are more favored to predict moderate air quality; ordered logistic regression models of AQI directly are more favored to predict good air quality.

Keywords: air quality index; particle pollution; linear regression; ordered logistic regression

1. Introduction

The availability of air pollution statistics has led to the development of different models and techniques to forecast air quality. For example, the literature on models of ozone levels is relatively well developed [1–4]. More recently, interest in air quality indices has increased [5,6], and more diverse measures of air pollution have been subject to time series analysis [7,8]. Different kinds of regression analyses such as multiple linear regression, principal component regression, independent component regression, quantile regression, and partial least squares regression have been used for forecasting daily air quality levels [9]. Vlachogianni, *et al.* [10] compared forecasts of multiple linear regression to that of artificial neural networks to investigate air quality. Stadlober, *et al.* [11] used linear regression models to combine information of the present day with meteorological forecasts of the next to predict daily PM_{10} concentrations, and showed that PM_{10} forecasting models based on

linear regression give suitable results in three European cities. Silva, *et al.* [12] applied nonparametric procedures to describe and forecast particulate material concentrations.

More recent literature related to pollution analysis with regression focuses on asthma [13], heart attack [14], health effects in general [15], and mortality [16–20].

The research reported here was motivated by interest in regression models of Air Quality Index (AQI) for particle pollution. The focus was small particles or droplets in the air that are 2.5 micrometers in diameter or smaller, emitted directly from forest fires and dust or indirectly from automobiles, industries and power plants. Excessive exposure to small particles could cause major health effects in humans including heart stroke, cancer, problems in pregnancy and many other short and long term health effects. With nearly 4 million residents, its high population density and traffic, Los Angeles has some of the most affected air in the United States (US). Los Angeles remains the worst city in America for ozone concentration, and one of the worst in particulate matter concentrations. It is reported that $PM_{2.5}$ is responsible for more than 125,000 cancer cases in the US and 16,250 in Los Angeles alone, and causes over 5000 premature deaths per year in the Los Angeles area (Air Quality Management District).

Negative health effects caused by particulate matter have been analyzed in many studies. A large-scale general review can be found in Pope and Dockery [21]. The National Association of Clean Air Agencies reported that $PM_{2.5}$ is the worst air pollutant because of its small size making it relatively easy to inhale. These particulates also consist of heavy metals, solid and liquid chemical elements and toxic organic compounds. It is crucial to develop good prediction and modeling techniques for the concentration of these pollutants in the air.

The Clean Air Act requires the US Environmental Protection Agency (EPA) to set, "National Ambient Air Quality Standards for pollutants considered harmful to public health and the environment". These standards along with the particle pollution data gathered for this study are provided at EPA.gov. More specifically, they are small particles recorded in the US city of Los Angeles (2001–2011), monitored air quality data from the EPA Air Quality System Data Mart (www.epa.gov/ttn/airs/aqsdatamart/). Early years (2001–2005) were used to fit models of particles recorded; later years (2006–2011) were reserved for out of sample comparison and contrast.

The AQI is a simple index for reporting daily air quality. AQI values map to an ordinal scale that is one where categories may be ordered, but assignment of numerical values would be arbitrary and so theoretically inappropriate. For ordinal data, we limit ourselves to statistical models that do not rely on numerical assignments. Linear regression models of particle pollution were used to generate predictions on a continuous scale that are mapped to predictions on the ordinal scale. Ordered logistic regression models were used to generate predictions directly onto the ordinal AQI scale: Good, Moderate, Unhealthy for Sensitive Groups (USG), Unhealthy, Very Unhealthy, and Hazardous.

As for independent variables that may be selected, we limited them to reasonable lagged observations of particle pollution observed today (PT): particle pollution observed yesterday (PD), particle pollution observed exactly one week ago (PW), and particle pollution observed exactly one year ago (PY). In other words we examined time series models as opposed to econometric ones that enjoy the benefit of external independent variables.

The rest of the paper is structured as follows: Section 2 includes in-sample results of Linear Regression. Section 3 includes in-sample results of Ordered Logistic Regression. Comparison and

Contrast are included in Section 4. Discussion of conclusions and future work is featured in Section 5. Tests of statistical significance are based on $\alpha = 0.10$.

2. Linear Regression

In this study, we assume particle pollution observed today PT has the normal distribution and constant variance. The mean, however, is assumed to be a linear function of lagged observations of the response: particle pollution observed yesterday (PD), particle pollution observed exactly one week ago (PW), and particle pollution observed exactly one year ago (PY).

$$PT \sim \text{normal}(\mu, \sigma)$$

$$\mu = \text{linear f}(PD, PW, PY)$$

The important question at first is whether or not coefficients (on the independent variables) are generally different from zero. Expectations based on the main effects model are given by the least squares fit (fit to particles recorded in years 2001–2005):

$$PT = 0.6891171(PD) + 0.0511629(PW) + 0.0335756(PY) + 4.601193$$

However, we fail to reject the hypothesis that β (PY) = 0, so we fit the full second order model to investigate interaction. No significant interaction in the full second order model includes particle pollution observed exactly one year ago (PY), so we drop it and reestimate the function of main effects:

$$PT = 0.6914947(PD) + 0.0377937(PW) + 5.483561$$

It explains $R^2 = 48.37\%$ of the variation in particles recorded in years 2001–2005.

3. Ordered Logistic Regression

In ordered logistic regression—a direct generalization of logistic regression—we estimate with maximum likelihood an underlying score as the linear function of independent variables and cut-points. The probability of observing an outcome is analogous to the probability that estimated linear function is within the outcome's cut-point range. We estimate the coefficients β together with the cut-points k where u is logistically distributed,

$$\begin{aligned} P(\text{outcome} = i) &= P(k_{i-1} < \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_{k+u} \leq k_I) \\ &= 1/[1 + \exp(-k_i + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k)] \\ &- 1/[1 + \exp(-k_{i-1} + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k)] \end{aligned}$$

Outcomes are Good, Moderate, Unhealthy for Sensitive Groups (USG), Unhealthy, Very Unhealthy, and Hazardous. Coefficients of the main effects ordered logistic regression model correspond to particle pollution observed yesterday (PD), particle pollution observed exactly one week ago (PW), and particle pollution observed exactly one year ago (PY). Only the coefficient on PD is generally different from zero considering years 2001–2005, so again we fit the full second order model to investigate interaction. No significant interaction includes PW, and none includes PY, so they are dropped from the main effects model which is re-estimated with PD as the lone independent variable.

$$P(\text{Good}) = P[0.138488(PD) < 2.134985]$$

$$P(\text{Moderate}) = P[2.134985 < 0.138488(PD) < 6.59586]$$

$$P(\text{USG}) = P[6.59586 < 0.138488(PD) < 9.838576]$$

$$P(\text{Unhealthy}) = P\,[9.838576 < 0.138488(PD)]$$

For a "pseudo" R^2 = 23.75% we use the formula $1 - L_1/L_0$ where L_0 is the constant-only log- likelihood, and $L_1 = -1063.8966$ is that of the model under consideration.

4. Comparison and Contrast

In order to gain some relative insight into the power of linear and ordered logistic regression for particle pollution, we evaluated models out of sample (2006–2011). Expected values based on linear regression were mapped to the ordinal scale; those based on ordered logistic regression were most likely according to expected probabilities. Results of linear regression are in Table 1; those of ordered logistic regression are in Table 2.

Table 1. Observed *versus* expected outcomes based on linear regression (out of sample).

Outcomes	Unhealthy	USG	Moderate	Good	Total
Good	0	0	315	645	960
Moderate	0	7	624	102	733
USG	0	9	26	1	36
Unhealthy	0	0	2	0	2
Total	0	16	967	748	1731

Table 2. Observed *versus* expected outcomes based on ordered logistic regression (out of sample).

Outcomes	Unhealthy	USG	Moderate	Good	Total
Good	0	0	218	819	1037
Moderate	1	7	565	219	792
USG	0	9	29	2	40
Unhealthy	0	0	2	0	2
Total	1	16	814	1040	1871

Total observations for ordered logistic regression are greater because fewer independent variables meant fewer missing data.

To summarize, we provide power to predict the outcomes for linear (REGRESS) and ordered logistic (OLOGIT) regression in Table 3.

Table 3. The power to predict.

Outcomes	Regress	OLOGIT
Good	645/960 = 67.2%	819/1037 = 79.0%
Moderate	624/733 = 85.2%	565/792 = 71.3%
USG	9/36 = 25%	9/40 = 22.5%
Unhealthy	0/2 = 0%	0/2 = 0%
Total	1278/1731 = 73.83%	1393/1871 = 74.45%

5. Conclusions

While linear (REGRESS) and ordered logistic (OLOGIT) regression performed similarly overall, their greatest powers clearly lie in the prediction of different outcomes. REGRESS is 1.19 times more powerful than OLOGIT to predict the Moderate outcome. OLOGIT is 1.88 times more powerful than REGRESS to predict the Good outcome. Future research into models of air quality index for particle pollution should determine the relative usefulness of those with approximately 74% total power to predict. Other future work should address the matter of extrapolation, which is for better or worse invited by REGRESS but prohibited by OLOGIT. Finally, these results are based entirely on arbitrary decisions including the choice of Los Angeles (2001–2011) and the cut-point to define out of sample years (2006–2011).

Author Contributions

Ege Ozdemir had the original idea for the study, was responsible for data cleaning and carried out the analysis. Justin R. Chimka drafted the manuscript which was revised by all authors. All authors read and approved the final the manuscript.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Robeson, S.M.; Steyn, D.G. Evaluation and comparison of statistical forecast models for daily maximum ozone concentrations. *Atmos. Environ. B Urban Atmos.* **1990**, *24*, 303–312.
2. Rao, S.T.; Zurbenko, I.G. Detecting and tracking changes in ozone air quality. *J. Air Waste Manag. Assoc.* **1994**, *44*, 1089–1092.
3. Yi, J.; Prybutok, V.R. A neural network model forecasting for prediction of daily maximum ozone concentration in an industrialized urban area. *Environ. Pollut.* **1996**, *92*, 349–357.
4. Kim, S.E.; Kumar, A. Accounting seasonality non-stationarity in time series models for short-term ozone level forecast. *Stoch. Environ. Res. Risk Assess.* **2005**, *19*, 241–248.
5. Bruno, F.; Cocci, D. A unified strategy for building simple air quality indices. *Environmetrics* **2002**, *13*, 243–261.
6. Bishoi, B.; Prakash, A.; Jain, V.K. A comparative study of air quality index based on factor analysis and US-EPA methods for an urban environment. *Aerosol Air Qual. Res.* **2009**, *9*, 1–17.
7. Modarres, R.; Dehkordi, A.K. Daily air pollution time series analysis of Isfahan City. *Int. J. Environ. Sci. Technol.* **2005**, *2*, 259–267.
8. Diaz-Robles, L.A.; Ortega, J.C.; Fu, J.S.; Reed, G.D.; Chow, J.C.; Watson, J.G.; Moncada-Herrera, J.A. A hybrid ARIMA and artificial neural networks model to forecast particulate matter in urban areas: The case of Temuco, Chile. *Atmos. Environ.* **2008**, *42*, 8331–8340.
9. Pires, J.C.M.; Martins, F.G.; Sousa, S.I.V.; Alvim-Ferraz, M.C.M.; Pereira, M.C. Prediction of the daily PM_{10} concentrations using linear models. *Am. J. Environ. Sci.* **2008**, *4*, 445–453.

10. Vlachogianni, A.; Kassomenos, P.; Karppinen, A.; Karakitsios, S.; Kukkonen, J. Evaluation of a multiple regression model for the forecasting of the concentrations of NOX and PM_{10} in Athen and Helsinki. *Sci. Total Environ.* **2011**, *409*, 1559–1571.
11. Stadlober, E.; Hormann, S.; Pfeiler, B. Quality and performance of a PM_{10} daily forecasting model. *Atmos. Environ.* **2008**, *42*, 1098–1109.
12. Silva, C.; Perez, P.; Trier, A. Statistical modeling and prediction of atmospheric pollution by particulate material: Two nonparametric approaches. *Environmentrics* **2001**, *12*, 147–159.
13. Li, S.; Batterman, S.; Wasilevich, E.; Wahl, R.; Wirth, J.; Su, F.; Mukherjee, B. Association of daily asthma emergency department visits and hospital admissions with ambient air pollutants among the perdiatric Medicaid population in Detroit: Time-series and time-stratified case-crossover anayses with threshold effects. *Environ. Res.* **2011**, *111*, 1137–1147.
14. Bhaskaran, K.; Hajat, S.; Armstrong, B.; Haines, A.; Herrett, E.; Wilkinson, P.; Smeeth, L. The effects of hourly differences in air pollution on the risk of myocardial infarction: Case crossover analysis of the MINAP database. *BMJ* **2011**, *343*, doi:10.1136/bmj.d5531.
15. Butland, B.K.; Armstrong, B.; Atkinson, R.W.; Wilkinson, P.; Heal, R.M.; Doherty, R.M.; Vieno, M. Measurement error in time-series analysis: A simulation study comparing modelled and monitored data. *BMC Med. Res. Methodol.* **2013**, *13*, 136, doi:10.1186/1471-2288-13-136.
16. Guo, Y.; Barnett, A.G.; Pan, X.; Yu, W.; Tong, S. The impact of temperature on mortality in Tianjin, China: A case-crossover design with a distributed lag non-linear model. *Environ. Health Perspect.* **2011**, *119*, 1719–1725.
17. Johnston, F.; Hanigan, I.; Henderson, S.; Morgan, G.; Bowman, D. Extreme air pollution events from bushfires and dust storm and their association with mortality in Sidney, Australia 1994–2007. *Environ. Res.* **2011**, *111*, 811–816.
18. Beverland, I.J.; Cohen, G.R.; Heal, M.R.; Carder, M.; Yap, C.; Robertson, C.; Hart, C.L.; Agius, R.M. A comparison of short-term and long-term air pollution exposure associations with mortality in two cohorts in Scotland. *Environ. Health Perspect.* **2012**, *120*, 1280–1285.
19. Hales, S.; Blakely, T.; Woodward, A. Air pollution and mortality in New Zealand: Cohort study. *J. Epidemiol. Community Health* **2012**, *66*, 468–473.
20. Zanobetti, A.; Dominico, F.; Wang, Y.; Schwartz, J.D. A national case-crossover analysis of the short-term effect of $PM_{2.5}$ of hospitalizations and mortality in subjects with diabetes and neurological disorders. *Environ. Health* **2014**, *13*, 38, doi:10.1186/1476-069X-13-38.
21. Pope, C.A.; Dockery, D.W. Health effects of fintie particulate air pollution: Lineas that connect. *J. Air Waster Manag. Assoc.* **2006**, *56,* 709–7042.