The Use of Quantile Regression to Forecast Higher Than Expected Respiratory Deaths in a Daily Time Series: A Study of New York City Data 1987-2000

Ireneous N. Soyiri[1,2], **Daniel D. Reidpath**[1,2*]

1 South East Asia Community Observatory (SEACO), School of Medicine and Health Sciences, Monash University, Kuala Lumpur, Malaysia, **2** Global Public Health, School of Medicine and Health Sciences, Monash University, Kuala Lumpur, Malaysia

Abstract

Forecasting higher than expected numbers of health events provides potentially valuable insights in its own right, and may contribute to health services management and syndromic surveillance. This study investigates the use of quantile regression to predict higher than expected respiratory deaths.

Data taken from 70,830 deaths occurring in New York were used. Temporal, weather and air quality measures were fitted using quantile regression at the 90th-percentile with half the data (in-sample). Four QR models were fitted: an unconditional model predicting the 90th-percentile of deaths (Model 1), a seasonal / temporal (Model 2), a seasonal, temporal plus lags of weather and air quality (Model 3), and a seasonal, temporal model with 7-day moving averages of weather and air quality. Models were cross-validated with the out of sample data. Performance was measured as proportionate reduction in weighted sum of absolute deviations by a conditional, over unconditional models; i.e., the coefficient of determination (R1).

The coefficient of determination showed an improvement over the unconditional model between 0.16 and 0.19. The greatest improvement in predictive and forecasting accuracy of daily mortality was associated with the inclusion of seasonal and temporal predictors (Model 2). No gains were made in the predictive models with the addition of weather and air quality predictors (Models 3 and 4). However, forecasting models that included weather and air quality predictors performed slightly better than the seasonal and temporal model alone (i.e., Model 3 > Model 4 > Model 2)

This study provided a new approach to predict higher than expected numbers of respiratory related-deaths. The approach, while promising, has limitations and should be treated at this stage as a proof of concept.

Citation: Soyiri IN, Reidpath DD (2013) The Use of Quantile Regression to Forecast Higher Than Expected Respiratory Deaths in a Daily Time Series: A Study of New York City Data 1987-2000. PLoS ONE 8(10): e78215. doi:10.1371/journal.pone.0078215

Editor: Cécile Viboud, National Institutes of Health, United States of America

Received November 19, 2012; **Accepted** September 13, 2013; **Published** October 11, 2013

Funding: These authors have no support or funding to report.

Competing interests: The authors have declared that no competing interests exist.

* E-mail: daniel.reidpath@monash.edu

Introduction

There is an increasing body of literature looking at the causal relationship between weather, air quality factors, and health outcomes [1-4]. Forecasting health outcomes has attracted less attention, but it too has a developing base in the scientific literature [5-9]. Traditionally, both the causal modelling and the forecast research have focused on the central tendencies of the distribution of data; i.e., the expected and conditional expected value. For instance, a typical generalised linear model of daily COPD events will model the expected number of COPD cases each day conditioned on a series of weather, air quality, and perhaps individual factors [10,11].

Although the expected outcomes can be important, the central portion of the conditional distribution is only one part of the story, and other parts of the conditional distribution can give quite different insights – particularly when the distributions are skewed. Studies of birth weight for example have shown quite different relationships between the explanatory variables and birth weight when modelling the conditional mean than they have when modelling low birth weight, such as birth weights in the lowest decile [12]. There is no requirement for the factors explaining low birth weight to be the same factors that explain average birth weight or for the explanation to be of the same form as for the central part of the conditional distribution.

Similarly, in modelling daily respiratory events (morbidity or mortality) and their relationship to air quality or weather, there is no strong requirement for the relationships that model average events to be the same as the relationships that model days with unusually high or unusually low numbers of events. By extension, forecasting the numbers of respiratory events on the outer arms of a conditional distribution need not rely on the same predictors that would be useful in forecasting the expected number of respiratory events.

To our knowledge, and not withstanding its potential value, there has only been one study looking at the forecasting of the number of respiratory events on the outer arm of a conditional distribution (such as the 90th percentile)[13]. There is the simple scientific interest in our capacity to make such forecasts, and what insights it might provide into the data; but there is also potential value for forecasting likely resource needs, as well as in areas such as syndromic surveillance, where the number of events exceeding a threshold is used to trigger a health systems response. Quantile regression remains a relatively unusual modelling technique in health research, which can be used to model conditional responses at any quantile of interest; and – although it has been used (rarely) for forecasting [14,15] – to our knowledge has never been used to forecast mortality.

Methods

We investigated the use of quantile regression to forecast the 90th percentile of daily, respiratory related deaths for New York City, in the period 1 January 1987 to 31 December 2000. The choice of the 90th percentile was somewhat arbitrary but in keeping with the idea of understanding the general capacity that a health system might need to maintain to meet typical demand. The data were drawn from the National Morbidity, Mortality, and Air Pollution Study (NMMAPS)[16], which are publicly available data through the Health and Air Pollution Surveillance System website (http://www.ihapss.jhsph.edu), and, in our case, accessed using the NMMAPS package in the R statistical environment [17]. The daily count of respiratory deaths was the outcome measure of interest. The data included 70,830 respiratory deaths over 5,114 days of surveillance.

The dataset also included a range of daily weather and air quality measures which were used as predictors in the modelling. The predictors included daily mean air temperature, dew point, ozone (O_3), sulphur dioxide (SO_2), nitrogen dioxide (NO_2), and carbon monoxide (CO). Measures of particulate matter were not included because of the levels of missing data within the dataset. In addition to the measures of weather and air-quality, cosinor values representing a yearly and a half yearly cycle[18,19], and dummy variables representing the days of the week were also used as predictors.

The data were sub-divided into two equal sized sets, from 1 January 1987 to 31 December 1993 for model development (*in-sample*), and from 1 January 1994 to 31 December 2000 for cross-validation (*out-sample*). We used the terms “prediction” to refer to the *in-sample* model development and then “forecasting” to refer to *out-sample* cross-validation. The size of the *in-sample* data was subsequently reduced to 2405 days (94.0% of total days) because of the use of lagged data, and a small amount of pre-existing missing data. The *out-sample* data (2548 days) were almost complete with a loss of only 8 days of data.

The details of quantile regression have been described elsewhere[20,21], as has its application to health problems[22,23]. The use of quantile regression with count data is unusual and its application to health forecasting remains novel [24,25].

A common challenge in modelling outcomes related to environmental exposures is the lagged effect between weather and air quality exposures and the health outcome of interest [26]. To identify an appropriate lag to represent the exposure to each of the weather and air quality measures, a series of quantile regression count models, using the *in-sample* data, were constructed testing the fit for each lag in turn, from a 1-day lag through to a 7-day lag; and was similar to approaches used elsewhere [9]. The fit of each lag, for each weather and air quality measure, was assessed using a function based on the asymmetric Laplace distribution commonly used in quantile regression. Best fit was determined by the lag that had the lowest value for:

$$\left[(1-p).\sum_{y_i<q}\left|y_i-q\right|\right]+\left[p.\sum_{y_i>q}\left|y_i-q\right|\right] \quad \textbf{(1)}$$

where the absolute deviations below quantile *q* are weighted by 1- *p* if the actual values lies below *q*, and *p* if the actual value lies above *q*. The lags that were identified for inclusion were: CO, NO_2, O_3, dew point (1-day), temperature and SO_2 (3-days). A 7-day moving average value for each of the weather and air quality factors was also included in one model as a point of contrast.

Four separate models were subsequently developed, three of which used quantile regression with either the selected lags or 7-day moving averages as predictors with the *in-sample* data. Model 1 was the intercept only model, an unconditional model predicting the value of the 90th percentile of daily respiratory deaths to be constant across the data. Model 2 was a conditional model in which the value of 90th percentile of daily respiratory deaths varied, conditioned on seasonal (cosinor values) and temporal (day of the week) predictors. Model 3 was a conditional model in which the value of 90th percentile of daily respiratory deaths varied conditioned on seasonal/ temporal predictors and the selected lags of weather and air quality predictors. Model 4 was the same as Model 3 except that the 7-day moving averages of weather and air quality predictors were used instead of selected lags.

The parameter estimates and standard errors from the quantile regression are not reported here, because they are essentially not required in the process for developing the forecasting model. Previous experience suggests that when they are provided, attention is inappropriately placed on that, rather than the predictive and forecasting capacity of the models.

The measure of fit used to establish the predictive validity (*in-sample* fit) and the forecasting accuracy (*out-sample* fit) of the quantile regression models was the coefficient of determination (R1)[20]. R1 measures the proportionate reduction in the weighted sum of the absolute deviations

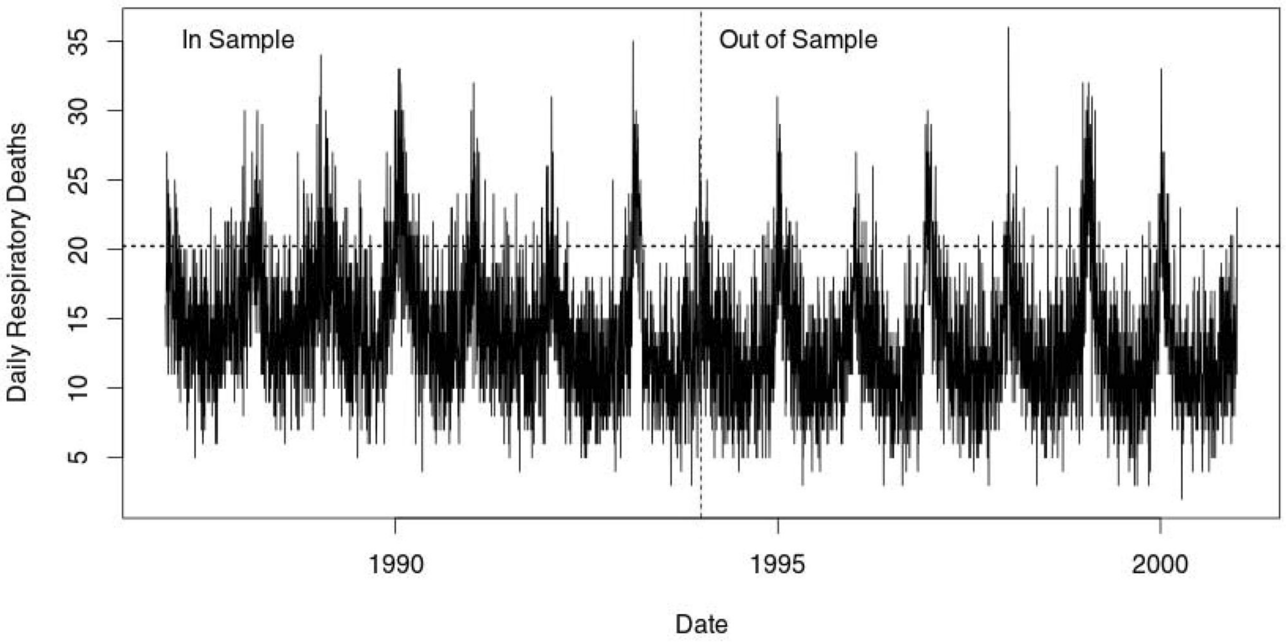

Figure 1. Time series of respiratory related deaths 1987—2000. The vertical dashed line indicates the separation between the in-sample and out-of sample data.

doi: 10.1371/journal.pone.0078215.g001

(WSADs) achieved by a conditional model over the unconditional model; where the weighted sum of the absolute deviations is given by equation 1. In this context, R1 is analogous to the mean absolute scaled error suggested by Hyndman and Koehler [27]. R1 was estimated for Models 2, 3, and 4, using the weighted sum of absolute deviations from Model 1 as the denominator. A supplementary approach which has been suggested for the application of count data in Quantile regression is for the data to be "jittered" prior to modelling [24]. In our preliminary work we did not observe any significant differences in model performance between the models of jittered data (of up to ±0.1 random numbers) and the time indicator, and those we reported.

Results

The time-series graph of the daily, respiratory deaths shows the familiar annual cycle with the peak deaths occurring in the winter months and the valleys occurring in the summer (Figure 1). The dashed vertical line in the middle of the figure shows the separation between the *in-sample* used to develop the quantile regression models and the *out-sample*, used to cross-validate the forecast models. The dotted horizontal line shows the *in-sample*, unconditional, 90th percentile number of daily deaths (20.2 per day). It is clear that a conditional distribution at the 90th percentile which included seasonal/temporal predictors would be quite different from the straight line of the unconditional quantile.

The top half of Table 1 shows the predictive capacity of the three conditional models (Models 2, 3, and 4) relative to the unconditional model (Model 1). The coefficient of determination (R1) showed improvements between .165 (i.e., a 16.5% improvement in the fit) and .191 over the unconditional model. The seasonal model (Model 2) slightly outperformed both the selected lags model and the 7-day moving average model.

Table 1. A comparison of the predictive and forecasting capacity of the models.

	Parameter	Model 1	Model 2	Model 3	Model 4
Prediction					
In sample	WSAD	2328.2	1882.5	1943.3	1927.5
	R1	0	.191	.165	.172
Forecasting					
Out sample	WSAD	2529.5	2121.3	2039.7	2055.3
	R1	0	.161	.194	.187

Perdition was based on the In Sample (days=2445) and forecasting was based on cross-validation of the Out Sample (days=2548): Model 1, intercept only; Model 2, temporal/seasonal model, Model 3, temporal/seasonal model with selected lags of weather and air quality; and Model 4 temporal/seasonal model with a 7-day moving average of weather and air quality. The weighted sum of the absolute deviation (WSAD) and the coefficient of determination (R1) are used to compare the models.

doi: 10.1371/journal.pone.0078215.t001

The models developed using the *in-sample* data were cross-validated using the *out-sample* data. The lower half of Table 1 shows the forecasting capacity of the three conditional models (Models 2, 3, and 4) relative to the unconditional model (Model 1). As anticipated, the forecasting performance of Model 2

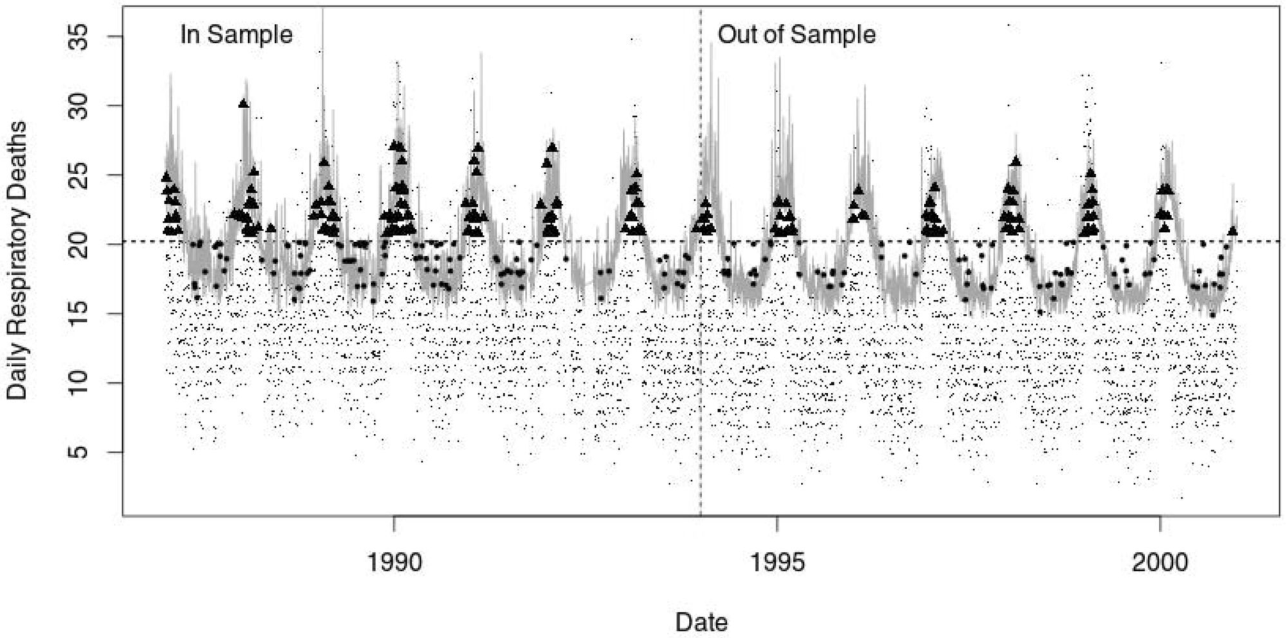

Figure 2. The quantile regression (90th percentile) model of respiratory related deaths. The small dots indicate daily deaths. The dotted horizontal line shows the unconditional 90th percentile.

doi: 10.1371/journal.pone.0078215.g002

(seasonal/temporal predictors) was slightly worse than the predictive performance (R1=.161); i.e., cross-validation against the *out-sample* was slightly weaker than predictions based on the model development data (*in-sample*). Model 3 (seasonal/ temporal predictors and selected lags for weather and air quality) performed slightly better in forecasting than it had in prediction (R1=.194). The performance of the 7-day moving average model (Model 4) also improved slightly (R1=.187).

As an illustration, Figure 2 shows the predictions (*in-sample*) and forecasts (*out-sample*) based on Model 3 (seasonal, temporal and selected lags as predictors) against the unconditional Model 1. Each point (small dot, larger dot, and solid triangle) represents the number of respiratory deaths recorded for a particular day. The conditional quantile regression model for the 90th percentile (shown in grey) lies, as expected, well above the central portion of the data – where a model of the conditional mean or median would lie – and shows clear seasonal variation. The dotted horizontal line shows the unconditional 90th percentile (Model 1). A few matters are worthy of note. All the points that lie above the dashed horizontal line would be regarded, under the unconditional model, as reflecting more unusual numbers of respiratory deaths. The points shown as black triangles reflect those days with numbers of deaths identified as more unusual under the unconditional model, but more typical under the more complex, conditional model. Conversely, all the points that lie below the dashed line would be regarded as more typical under the unconditional. The points shown as larger dots reflect those days with numbers of deaths identified as more typical under the unconditional model, but more unusual under the more complex, conditional model. Visually, the predictive and forecast performance of the model appears to be reasonably consistent (Figure 2).

Discussion

The notion of modelling and forecasting the expected number of daily deaths is well described in the literature [2,28-30]. Forecasting any health outcome on the outer arms of a conditional distribution, however, is unusual [15], and appears not to have been done in the analysis of daily time series data related to mortality. This is unfortunate, because there are things to be learned from forecasts made at, for instance, the 90th percentile that could not be learned from forecasting the expected number of daily deaths.

For example, forecasts of the expected number of deaths will underestimate the kinds of resources that need to be available much of the time, particularly in an environment with the kind of variability shown in the daily respiratory deaths data. There is cyclical variation in the data, but even within the data at any one part of the annual cycle, there is substantial daily variation. Forecasts of likely numbers of deaths (i.e., occurring more than 90% of the time), can also feed into a mechanism for identifying when there is a concerning deviation in the number of deaths. Sustained numbers of days with deaths above the forecast can inform a health system about the occurrence of a likely environmental exposure or emerging disease.

Furthermore, because the forecasts are conditional, relatively low absolute numbers of deaths occurring in the summer, can still trigger a response when those numbers (although low) fall consistently above the conditional 90th percentile. They also forecast when resources and capacity may be reduced.

The analysis presented here showed some forecasting benefit associated with the inclusion of selected lags of daily weather or air quality data (i.e., a difference in R1 between .161 and .194 – a 3% improvement over the unconditional model). A trade-off arises, however, between developing more complex conditional models over models including only temporal and seasonal predictors.

There are important limitations with the approach taken here, and these can be used to highlight future pathways for analysis. The first limitation is with the use of the 90th percentile. One can potentially analyse the data at any conditional quantile, and for different purposes (such as surveillance or resource allocation) analyses at different quantiles – or multiple quantiles – may be more useful. The utility is driven by the application, and as we were seeking a proof of concept, the 90th percentile seemed to be appropriate level. Using cosinor values of yearly and half yearly cycles may not capture important seasonal information that could be built into the forecasts, and is certainly worthy of future investigation. There is a balance to be made in forecasting between the gain in accuracy and the cost of implementation. Sinusoidal functions capturing seasonal and temporal variation are trivial to develop and implement, and provide around a 15-20% improvement in accuracy over using an annual figure for the 90th percentile. More complicated conditional models appear to add a 2-3% improvement. The utility of the gain for the effort is uncertain. The final limitation we consider here is a theoretical one. There is often concern expressed with forecasting models that do not take a more traditional causal modelling approach [31]. We would take two distinct lines of argument in response. The first line of response is that the purpose of forecasting is not about determining cause and effect, and therefore forecasting models should be judged according to their forecasting accuracy, not for their inadequacy at providing causal explanations. The second line of response is that if a causal model out-performs non-causal model in forecast accuracy, then the causal model should absolutely replace the non-causal model. The causal model was not developed here, but there is some reason to believe that it may not perform as well as a “dust bowl” empirical approach [31] that has no interest in explaining the relationships between factors involved but rather for forecasting the outcome.

Conclusion

This study reports for the first time, a statistical approach for forecasting respiratory related deaths at the 90th percentile using quantile regressions. The results suggest there is potential value in this, even when the model is no more sophisticated than a seasonal/temporal model. The study should, however, be treated as a proof of concept, rather than definitive.

Author Contributions

Conceived and designed the experiments: DDR INS. Performed the experiments: INS DDR. Analyzed the data: DDR INS. Contributed reagents/materials/analysis tools: INS DDR. Wrote the manuscript: DDR INS.

References

1. Cheng Y, Kan H (2012) Effect of the Interaction Between Outdoor Air Pollution and Extreme Temperature on Daily Mortality in Shanghai, China. J Epidemiol 22: 28-36. doi:10.2188/jea.JE20110049. PubMed: 22041530
2. Metzger KB, Ito K, Matte TD (2009) Summer heat and mortality in New York city: How hot is too hot? Environ Health Perspect 118: 80–6. PubMed: 20056571.
3. Rajarathnam U, Sehgal M, Nairy S, Patnayak RC, Chhabra SK et al. (2011) Time-series study on air pollution and mortality in Delhi Res Rep Health Eff Inst 157: 47-74.
4. Zibners LM, Bonsu BK, Hayes JR, Cohen DM (2006) Local weather effects on emergency department visits: a time series and regression analysis. Pediatr Emerg Care 22: 104-106. doi:10.1097/01.pec.0000199561.34475.29. PubMed: 16481926.
5. Chatterjee C, Sarkar RR (2009) Multi-step polynomial regression method to model and forecast malaria incidence. PLOS ONE 4: 6. PubMed: 19266093.
6. Gharbi M, Quenel P, Gustave J, Cassadou S, Ruche G et al. (2011) Time series analysis of dengue incidence in Guadeloupe, French West Indies: Forecasting models using climate variables as predictors. BMC Infect Dis 11: 166. doi:10.1186/1471-2334-11-166. PubMed: 21658238.
7. Kam HJ, Sung JO, Park RW (2010) Prediction of Daily Patient Numbers for a Regional Emergency Medical Center using Time Series Analysis. Healthc. Inf Res 16: 158-165.
8. Soyiri IN, Reidpath DD (2012) Semi-structured black-box prediction: proposed approach for asthma admissions in London. Int J Gen Med 5: 693-705. PubMed: 22973117.
9. Soyiri IN, Reidpath DD (2012) Humans as Animal Sentinels for Forecasting Asthma Events: Helping Health Services Become More Responsive. PLOS ONE 7: e47823. doi:10.1371/journal.pone.0047823. PubMed: 23118897.
10. Dominici F, Peng RD, Bell ML, Pham L, McDermott A et al. (2006) Fine Particulate Air Pollution and Hospital Admission for Cardiovascular and Respiratory Diseases. J Am Med Assoc 295: 1127-1134. doi:10.1001/jama.295.10.1127. PubMed: 16522832.
11. Stieb DM, Szyszkowicz M, Rowe BH, Leech JA (2009) Air pollution and emergency department visits for cardiac and respiratory conditions: a multi-city time-series analysis. Environ Health 8: 25. doi:10.1186/1476-069X-8-25. PubMed: 19515235.
12. Wehby GL, Murray JC, Castilla EE, Lopez-Camelo JS, Ohsfeldt RL (2009) Prenatal care effectiveness and utilization in Brazil. Health Policy Plan 24: 175–188. doi:10.1093/heapol/czp005. PubMed: 19282483.
13. Soyiri IN, Reidpath DD, Sarran C (2012) Forecasting peak asthma admissions in London: an application of quantile regression models. Int J Biometeorol, 57: 1-10. PubMed: 22886344.
14. Burgette LF, Reiter JP, Miranda ML (2011) Exploratory quantile regression with many covariates: an application to adverse birth outcomes. Epidemiology 22: 859-866. doi:10.1097/EDE.0b013e31822908b3. PubMed: 21968775.
15. Sun Y, Teow KL, Heng BH, Ooi CK, Tay SY (2012) Real-Time Prediction of Waiting Time in the Emergency Department, Using Quantile Regression. Ann Emerg Med 60: 299-308. doi:10.1016/j.annemergmed.2012.03.011. PubMed: 22579492.
16. Dominici F, McDermott A, Daniels M, Zeger SL, Samet JM (2005) Revised analyses of the National Morbidity, Mortality, and Air Pollution Study: mortality among residents of 90 cities. J Toxicol Environ Health A 68: 1071-1092. doi:10.1080/15287390590935932. PubMed: 16024489.
17. Peng RD, Welty LJ (2004) The "NMMAPSdata" Package. R NEWS 4: 10-14.
18. Barnett AG, Dobson AJ (2010) Analysing Seasonal Health Data. Berlin: Springer-Verlag. 168(139-140) p

19. Stolwijk AM, Straatman H, Zielhuis GA (1999) Studying seasonality by using sine and cosine functions in regression analysis. J Epidemiol Community Health 53: 235-238. doi:10.1136/jech.53.4.235. PubMed: 10396550.
20. Hao L, Naiman DQ, editors (2007) Quantile Regression. Thousand Oaks, California 91320. Sage Publications, Inc.. p. 139.
21. Koenker R, Xiao Z (2006) Quantile Autoregression. J Am Stat Assoc 101: 980-990. doi:10.1198/016214506000000672.
22. Beyerlein A, Fahrmeir L, Mansmann U, Toschke AM (2008) Alternative regression models to assess increase in childhood BMI. BMC Med Res Methodol 8: 59. doi:10.1186/1471-2288-8-59. PubMed: 18778466.
23. Beyerlein A, Nehring I, Rzehak P, Heinrich J, Müller MJ et al. (2012) Gestational Weight Gain and Body Mass Index in Children: Results from Three German Cohort Studies. PLOS ONE 7: e33205. doi: 10.1371/journal.pone.0033205. PubMed: 22457745.
24. Machado JAF, Silva JMCS (2005) Quantiles for Counts. J Am Stat Assoc 100: 1226-1237. doi:10.1198/016214505000000330.
25. Winkelmann R (2006) Reforming health care: evidence from quantile regressions for counts. J Health Econ 25: 131-145. doi:10.1016/j.jhealeco.2005.03.005. PubMed: 15978687.
26. Zeger SL, Irizarry R, Peng RD (2006) On Time Series Analysis of Public Health and Biomedical Data. Annu Rev Public Health 27: 57-79. doi:10.1146/annurev.publhealth.26.021304.144517. PubMed: 16533109.
27. Hyndman RJ, Koehler AB (2006) Another look at measures of forecast accuracy. Int J Forecasting 22: 679-688. doi:10.1016/j.ijforecast.2006.03.001.
28. Peng RD, Dominici F, Pastor-Barriuso R, Zeger SL, Samet JM (2005) Seasonal Analyses of Air Pollution and Mortality in 100 US Cities. Am J Epidemiol 161: 585-594. doi:10.1093/aje/kwi075. PubMed: 15746475.
29. Tao Y, Huang W, Huang X, Zhong L, Lu SE et al. (2012) Estimated acute effects of ambient ozone and nitrogen dioxide on mortality in the Pearl River Delta of southern China. Environ Health Perspect 120: 393-398. PubMed: 22157208.
30. Yang C, Yang H, Guo S, Wang Z, Xu X et al. (2012) Alternative ozone metrics and daily mortality in Suzhou: The China Air Pollution and Health Effects Study (CAPES). Sci Total Environ 426: 83-89. doi: 10.1016/j.scitotenv.2012.03.036. PubMed: 22521098.
31. Breiman L (2001) Statistical Modeling: The Two Cultures. Stat Sci 16: 199-215. doi:10.1214/ss/1009213726.