GLAM - Final project

Marilyn Ello, Shiva Motie, Olivier Grognuz

June 7, 2021

Contents

1 Introduction

It is known that common airborne pollutants can have negative effects on the general health of citizens. In the short term, the presence of pollutants can yield to increasing risk of respiratory problems and culminate in long-term effects on the pulmonary function.

In order to assess the short-term implications of the presence of pollutants in the air, data on hospital admissions for respiratory diseases as well as daily average 24-hour concentrations of common pollutants in an asian city were gathered and compiled in a dataset.

The following data is available:

- `resp`: daily hospital admissions for respiratory diseases;
- `no2`: average daily (24h) concentration of nitrogen dioxide NO_2;
- `so2`: average daily (24h) concentration of sulphur dioxide SO_2;
- `rsp`: average daily (24h) concentration of respirable suspended particules;
- `o3`: average concentration of O_3;
- `temp`: average daily temperature (in Celsius);
- `hum`: average daily humidity rate;
- time related variables as to when the observations were taken, `day`, `month`, `year` and the day of the week `weekday`.

In this report, we describe our analysis on the effect of the different air pollutants as well as the temperature and humidity rate on the count of hospital admissions. We also assess the impact of the time (day of the month , month of the year , weekdays, and year of interest) on this admissions count. We start by exploring the data to see what we have to deal with. Then, we suggest some models and try to measure the effect of the different variables on the count data as well as possible. For each model, we check if the assumptions are satisfied and we choose the best model among all. In the last section, we interpret the results of the final, selected model.

2 Exploratory Data Analysis

Since the variable `resp` that we are trying to model is a count, the basic approach is to model it with a Poisson model with the canonical `log` link.

The average number of hospital admissions is 185.19 (var = 2186.43). Although the histogram looks normal (see Figure 1), this is no counter-indication that the Poisson model should not be used. On any given day, at least 87 were admitted in the hospital and at most 332. This indicates that we do not have to deal with zero-inflated data.

When considering the variable `year` as a factor, by looking at the box plots, we clearly see an increase in `resp` as years go by, while the other variables remain constant. It seems that it may be also the case for some days of the week (see Figure 2).

Regarding the factor `month`, the box plots seem to suggest some pattern throughout the year which could be seen as akin to seasons. Indeed, similarities between consecutive months were observed so we created another factor `month_bis` to capture this information. The latter is similar to the former but with four levels instead of twelve. Levels 1, 2 and 12 of `month` were grouped in a level called `1`, levels 3, 4 and 5 in a level called `2`, levels 6, 7, 8 in `3` and levels 9, 10, 11 in `4`.

One way to see the linear or nonlinear relationship between the continuous variables is to look at their scatter plot (see Figure 3). An interesting thing to notice is that it seems there is no clear relationship between `rsp` (respirable suspended particles) and the hospital admissions, as it would be expected. It can also be seen that there may be a linear relationship between two variables `temp` (temperature) and `hum` (humidity), but it is not so clear. It may imply that considering only one of them in the model should be enough. Furthermore, a stronger linear relationship between `no2` and `so2`, and also `no2` and `rsp` can be seen.

All these intuitions help us to better work with the dataset and the models, but the final model and variables to be used should be selected during the modeling procedure (model validation and variable selection).

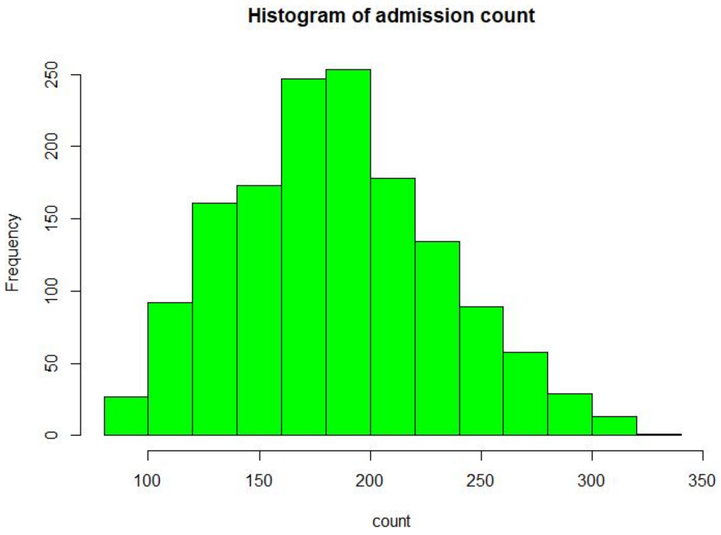

Figure 1: Histogram of Hospital admission count per day

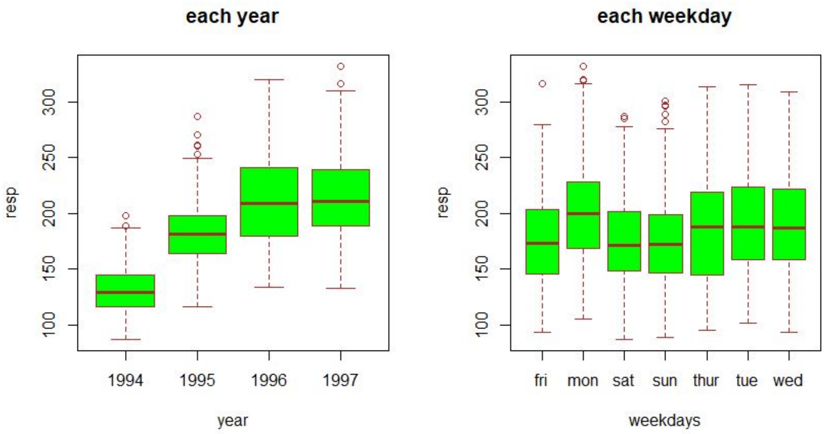

Figure 2: Box plots of years and weekdays

Before jumping to models, we should make sure that we have proper data to avoid probable problems. By looking at the data, we observe that at some days (10 days), we have some missing values corresponding to predictive variables or the response variable itself. Therefore, as these 10 days form a small proportion of all data (1455 days) we decided to remove them from our analysis.

3 Possible Models and model checks

Please note that the results mentioned in this section are not illustrated by plots and R outputs for the sake of readability. All results can be reproduced using the separate R script.

3.1 Poisson

As previously discussed, according to the response variable (which is a count data), the first model which comes to mind is a Poisson Generalized Linear Model (GLM) with a log link. To fit this model, we use R studio platform, and we define the first Poisson GLM with all possible variables as follows:

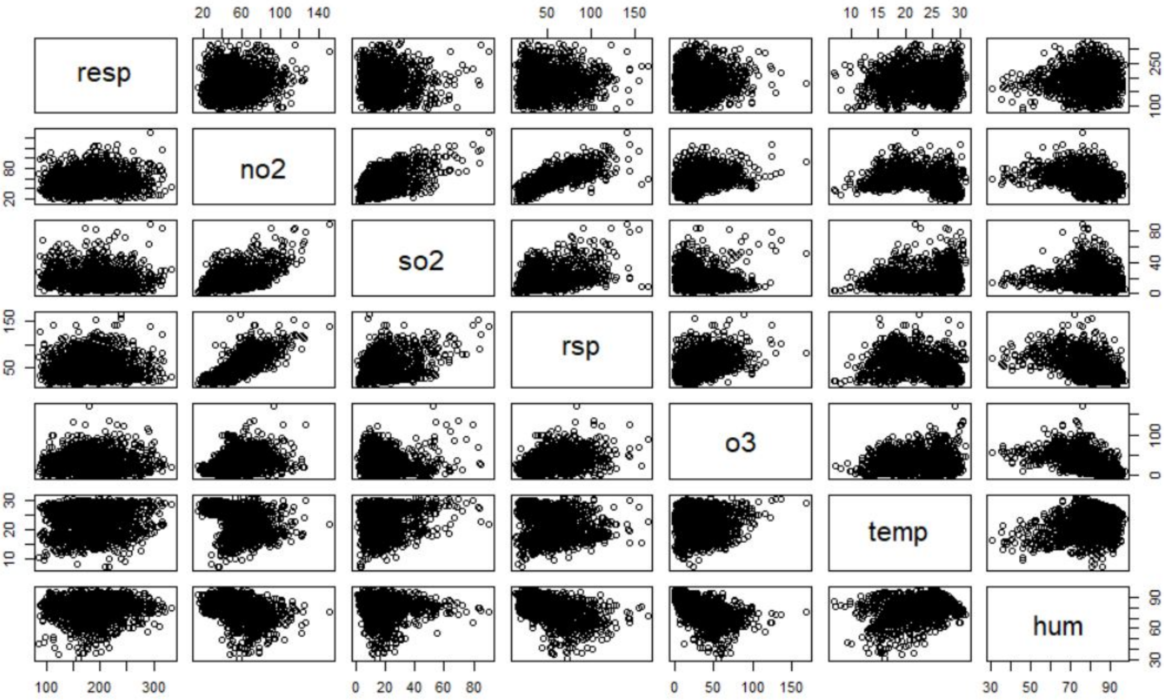

Figure 3: scatterplot of continuous variables

$$\begin{aligned}\log(\texttt{resp}) = \beta_0 + \beta_1 \times \texttt{no2} + \beta_2 \times \texttt{so2} + \beta_3 \times \texttt{rsp} + \beta_4 \times \texttt{o3} + \beta_5 \times \texttt{temp} + \beta_6 \times \texttt{hum} \\ + \beta_7 \times \texttt{month} + \beta_8 \times \texttt{weekday} + \beta_9 \times \texttt{year} + \beta_{10} \times \texttt{day}.\end{aligned} \tag{1}$$

From the results we see that some variables such as `rsp`, `temp`, `day`, and `day` are not significant at the 5% level. Thus, we should perform some variable selection method. Here, we use the `step` function, which is a grid search variable selection method based on the Akaike Information Criterion (AIC).

As a result, we decide to remove `day` and `rsp` and continue the modeling with the remaining variables:

$$\begin{aligned}\log(\texttt{resp}) = \beta_0 + \beta_1 \times \texttt{no2} + \beta_2 \times \texttt{so2} + \beta_3 \times \texttt{o3} + \beta_4 \times \texttt{temp} + \beta_5 \times \texttt{hum} + \beta_6 \times \texttt{month} \\ + \beta_7 \times \texttt{weekday} + \beta_8 \times \texttt{year}.\end{aligned} \tag{2}$$

We also considered some interactions between the factors (`weekday`, `month`, `year`) at some point. Now we observe that all the variables are significant (`p-value` < 0.05).

The next step is to check the goodness of fit and model validation for the best model up to now. To do so, we can look at the percentage of the deviance explained by the model, which is derived by this formula:

$$D_{\text{exp}} = \frac{D_{\text{null}} - D_{\text{fit}}}{D_{\text{null}}}.$$

It is also possible to look at the p-value of residual deviance. Under the null hypothesis, if the model is a good model, it should follow a Chi squared distribution. So if the H_0 is rejected, we can infer that our model is not good enough.

By calculating these two values, we see that the percentage of the deviance explained is 0.83, but the H_0 will be rejected (p-value = 0). So it seems that this poisson GLM is not good enough.

It may have happened because the model assumptions are not satisfied. As it is known that the poisson model assumes the variance and the mean of the response variable to be the same, and if it is not the case for the data in hand, it may causes some problems like overdispersion. To check that, we computed the magnitude of the variance over the mean, and we recognized that it is equal to 11.8, indicating that we have overdispersion and a negative binomial model may be better in this case.

Before that, we also consider a non-parametric approach with a Generalized Additive Model (GAM).

3.2 Generalized Additive Model

The first Generalized Additive Model fitted, using the gam function, is a Poisson model with a log link:

$$\begin{aligned}\log(\texttt{resp}) = \alpha + f_1(\texttt{no2}) + f_2(\texttt{so2}) + f_3(\texttt{rsp}) + f_4(\texttt{o3}) + f_5(\texttt{temp}) + f_6(\texttt{hum}) + f_7(\texttt{day}) \\ + \text{factor}(\texttt{month}) + \text{factor}(\texttt{weekday}) + \text{factor}(\texttt{year}).\end{aligned} \tag{3}$$

This is a full model containing all variables without interaction terms. It has a deviance explained of 0.754 and an AIC equal to 14473.5.

Since the variable `rsp` is not significant we decide to remove it from the model and as result, we have a smaller AIC.

Now, we want to try to fit a model with an interaction between `year` and `month`:

$$\begin{aligned}\log(\texttt{resp}) = \alpha + f_1(\texttt{no2}) + f_2(\texttt{so2}) + f_3(\texttt{o3}) + f_4(\texttt{temp}) + f_5(\texttt{hum}) + f_6(\texttt{day}) + factor(\texttt{month}) \\ + \text{factor}(\texttt{weekday}) + \text{factor}(\texttt{year}) + \text{factor}(\texttt{year}) \times \text{factor}(\texttt{month}).\end{aligned} \tag{4}$$

This model has a deviance explained of 0.841 percent and an AIC equal to 13045.9; which is a substantial improvement compared to the model without the interaction term.

However, when replacing `month` by `month_bis`, with obtain AICs equal to 15260.8 and 14644.6 for the models with and without interaction term respectively. This makes us more inclined to use the variable `month`. As we will see further in the report, another model will have a lower AIC than the GAM with interaction term.

3.3 Negative Binomial

As we do not get suitable models from previous methods, and as there is an overdispersion problem in the data, we decide to fit a negative binomial model with the glm.nb function. Fortunately, this model fit the data very well and the resulting model also satisfies the assumptions. Therefore, we choose this one as our final model, and we are going to discuss it in detail in the next section.

4 Final model

4.1 Model introduction and formulation

Negative binomial model, is a kind of parametric Generalized linear model that can be used for count data with overdispersion problem, which is the case for our data set.

In order to implement this method in Rstudio, we need to use the MASS library. As before, for the first try we introduce all the possible variables and some interactions to the model:

$$\begin{aligned}\log(\texttt{resp}) = \beta_0 + \beta_1 \times \texttt{no2} + \beta_2 \times \texttt{so2} + \beta_3 \times \texttt{rsp} + \beta_4 \times \texttt{o3} + \beta_5 \times \texttt{temp} + \beta_6 \times \texttt{hum} \\ + \beta_7 \times \texttt{month} + \beta_8 \times \texttt{weekday} + \beta_9 \times \texttt{year} + \beta_{10} \times \texttt{day} + \beta_{11} \times \texttt{year} \times \texttt{month} \\ + \beta_{12} \times \texttt{year} \times \texttt{weekday} + \beta_{13} \times \texttt{month} \times \texttt{weekday}.\end{aligned} \tag{5}$$

As we can see in the output of the R code, there are some variables and interaction which seems to be insignificant such as `rsp`, `o3`, `hum`, `weekday`, interaction of `weekday` and `year`, etc. But we cannot remove them all, because removing one of them may make some others significant. Therefore, it should be done in a proper way with a variable selection method.

4.2 Variable selection

As before, we use a grid search variable selection, the R function `step`, which is based on the AIC. As we can see in the Rstudio results, we should first remove the interaction between `month` and `weekday`, afterwards the factor `day` itself, then we remove the interaction between `weekday` and the `year`, and at last we remove the variable `hum`. Although it is suggested by the AIC to remove the `o3` variable as well, we choose not to do so for two main reasons. First, it only decreases the AIC by two units which is relatively low, and second, we are especially interested in the effect of the different gases, even though they might appear not to be significant. Therefore, the model reduces to:

$$\begin{aligned}\log(\texttt{resp}) = \beta_0 + \beta_1 \times \texttt{no2} + \beta_2 \times \texttt{so2} + \beta_3 \times \texttt{rsp} + \beta_4 \times \texttt{o3} + \beta_5 \times \texttt{temp} \\ + \beta_6 \times \texttt{month} + \beta_7 \times \texttt{weekday} + \beta_8 \times \texttt{year} + \beta_9 \times \texttt{year} \times \texttt{month}.\end{aligned} \tag{6}$$

4.3 Goodness of Fit

Afterwards, we should assess the goodness of fit of this model. As previously discussed in section 3, we have to compute the percentage of the deviance explained by the model, and also the p-value of the residual deviance using chi squared distribution.

Listing 1: R output for the goodness of fit

```
> 1-pchisq(summary(NB_2)$deviance,summary(NB_2)$df[2])
[1] 0.1339717
> D_exp = (summary(NB_2)$null.deviance - summary(NB_2)$deviance)/summary(NB_2)$null.deviance
> D_exp
[1] 0.8310889
```

By calculating these two values, we see that the deviance explained is 0.8311, which is very good, and the p-value of the residual deviance under H_0 is equal to 0.134 (larger than 0.05) see listing 1. Therefore, the null hypothesis will not be rejected, and it tells us that this Negative Binomial model is good enough. According to the listing 2 we also get an AIC equal to 12716 which is lower than the other models.

4.4 Model Validation

The next step is to check whether our model is valid or not. It means that we should check that the model assumptions are satisfied. These assumptions are normality of the residuals, homoscedastic residuals, lack of structure in residuals versus fitted values, mean zero in residuals, and so on. This model check can be done and viewed through some plots.

As it can be seen in Figure 4 the residuals are scattered around zero and show no strong structure while plotting versus fitted value. Furthermore, looking at the residuals versus normal Q-Q plot, demonstrate that they follow a normal distribution.

To be sure that the latter result was not due by chance, we also check the Randomized Quantile residuals in Figure 5, and it also confirms the normality of residuals.

Therefore, we can say that our model is valid and it can be used to assess the impact of different pollutants on the admission count. This assessment and the relevant interpretations are going to be done in the next section.

5 Interpretation

Before analyzing the coefficients themselves in Listing 2, we notice that the estimated dispersion parameter θ is 183.1, thus confirming the overdispersion issue. From the estimated coefficients, we see that the daily average concentration of `no2` and `so2` have a significant impact on the number of admissions, whereas the average concentration of `o3` and respirable suspended particles `rsp` do not seem to have any significant impact. The latter result is counter-intuitive to say the least, but since no details as to what those "respirable particules" refer to are available, further investigation would need to be conducted.

In details, we see that the significant coefficient `no2` increases the log count of hospital admissions by approximately 0.0015 for each added unit, while `so2` decreases the log count of admissions by 0.0009 while the other variables remain constant. Similarly, the temperature also has a significant increasing effect, as a unit-increase positively impacts the log of number of admissions by 0.0073.

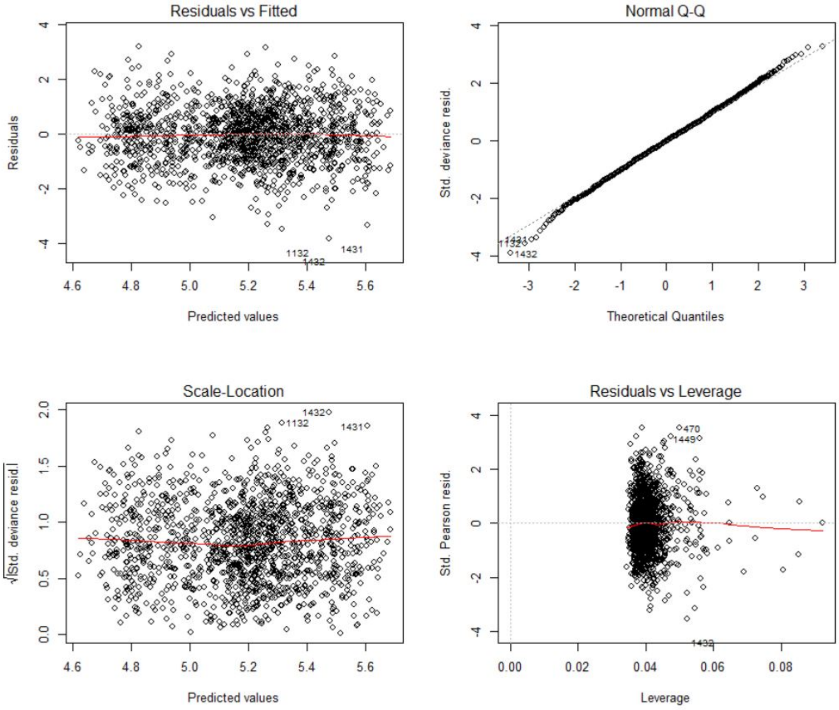

Figure 4: Residual Plots

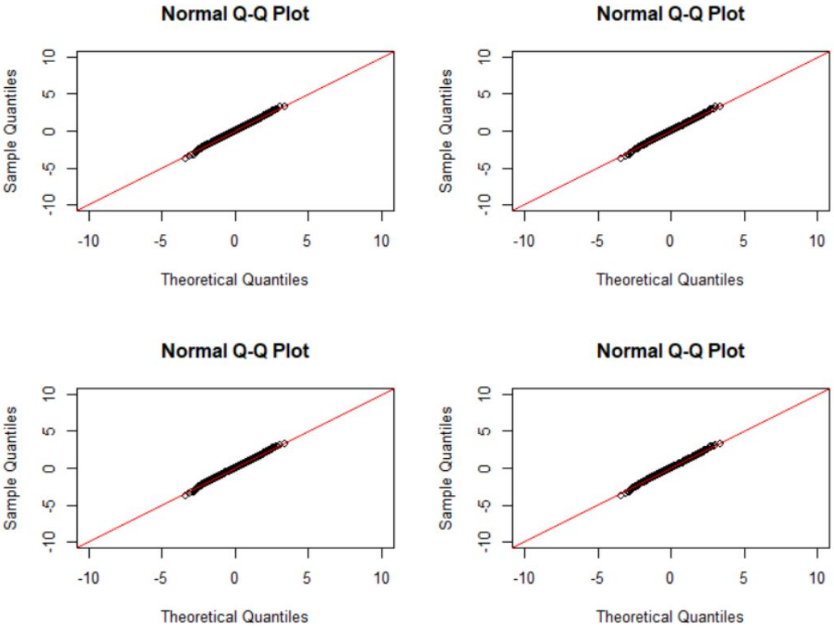

Figure 5: Randomized Quantile Residuals

With respect to the first month (January, which is also the coldest on average), all the other warmer months see an increase in the hospital admissions. Although a direct link with the temperature can surely be done, the months could also be a proxy for other non-observed variables like holiday periods. With that in mind, we see that the months `4, 5, 8` and `9` are not significant and that despite being also cold, December has an even greater effect (0.204) than some warmer months like April (0.082), again suggesting that the `month` variable is not only capturing the temperature change. Comparing to 1994, all the following years in the data see a significant increase in the log count of admissions, as it was hypothesized in the exploratory data analysis (Figure 2 left). Bear in mind that this interpretation about `month` and `year` has to be taken with a grain of salt, since we are also considering the interaction effect between these two variables.

When observing the `weekday` coefficients, it is interesting to see that all working days see an increase in the log count when comparing to Friday as a base, whereas the weekend see a decrease, although the effect is not significant for the weekend. This could be a hint that less air pollutants are present in the air during the weekends or that the eventual patients prefer to visit the hospital during the week.

After that, we see the effect of the interaction between `month` and `year` when compared to January of 1994 as the base. For 1995, months `6, 8` and `9` are not significant. The first months of the year all see a significant increase in the log count of admissions of respectively 0.117, 0.089, 0.153 and 0.198, while the later months all see a decrease in the log count. This behavior is similar in 1996, where the first half of the year see an increase in the admissions and the second half a decrease, with August (`month 8`) being not significant. This is no more the case in 1997, when more months become not significant (`2, 4, 7, 12`). This specific year, the first months see a decrease in the log count whereas an increase is observed for May and June.

6 Conclusion

Let us first recall that our goal was to find the effect of different air pollutants (`no2`, `so2`, `rsp` and `o3`) as well as the meteorological conditions as measured by the temperature, the humidity and generally the time (month, year and day), on the hospital admission count of an Asian city. In the exploratory phase, we could confirm that we were not dealing with zero-inflated data but that we could have overdispersion. This claim was further supported when fitting a first Poisson model which was not a good fit. Before fitting a Negative binomial model, we also tried a highly flexible non-parametric GAM, which ended up yielding results that were not satisfactory. All this convinced us to keep the more simple Negative Binomial model, which would simply account for the overdispersion issue.

After validating the latter, we were able to outline the following interpretation. First, only NO_2 and SO_2 seem to have a direct, significant impact on the admission count, while the average concentration of respirable particles and O_3 do not have a significant effect. Furthermore, NO_2 has a positive effect, whereas SO_2 has a negative effect on the admission count. The meteorological conditions, translated as the temperature and the time data in our model, also has an impact. The higher the temperature, the higher the count. When looking at the time data, we see that the result is more nuanced, but the general trend is that admission count goes up as years go by, with peculiarities for some specific months.

It appears the data that was made available was not sufficient to be able to confidently measure the effect of respirable particles present in the air, which is unfortunate because the existing literature seems to agree on its effect. Also, further research should include more precise meteorological features like the wind direction, and speed. More context-specific information such as the name and location of the city should also be included to allow for a more accurate analysis.

Listing 2: Negative Binomial output

```
Call:
glm.nb(formula = resp ~ no2 + so2 + rsp + o3 + temp + factor(month) +
    factor(weekday) + factor(year) + factor(year):factor(month),
    data = data, init.theta = 183.1460895, link = log)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.8113  -0.6562  -0.0078   0.6198   3.1899

Coefficients:
                                   Estimate Std. Error z value Pr(>|z|)
(Intercept)                       4.470e+00  3.617e-02 123.564  < 2e-16 ***
no2                               1.481e-03  2.990e-04   4.953 7.31e-07 ***
so2                              -9.850e-04  3.652e-04  -2.697 0.006992 **
rsp                               3.362e-04  2.145e-04   1.568 0.116895
o3                                7.474e-05  1.642e-04   0.455 0.648914
temp                              7.250e-03  1.457e-03   4.975 6.51e-07 ***
factor(month)2                    6.485e-02  3.108e-02   2.086 0.036946 *
factor(month)3                    2.105e-01  2.976e-02   7.072 1.53e-12 ***
factor(month)4                    8.213e-02  3.292e-02   2.495 0.012603 *
factor(month)5                    1.735e-02  3.383e-02   0.513 0.608132
factor(month)6                    4.192e-02  3.488e-02   1.202 0.229508
factor(month)7                    2.298e-01  3.397e-02   6.764 1.34e-11 ***
factor(month)8                    3.309e-02  3.417e-02   0.968 0.332902
factor(month)9                    4.311e-02  3.380e-02   1.275 0.202194
factor(month)10                   1.980e-01  3.209e-02   6.170 6.82e-10 ***
factor(month)11                   1.815e-01  3.127e-02   5.804 6.49e-09 ***
factor(month)12                   2.050e-01  3.023e-02   6.783 1.18e-11 ***
factor(weekday)mon                1.344e-01  1.029e-02  13.059  < 2e-16 ***
factor(weekday)sat               -6.366e-03  1.045e-02  -0.609 0.542399
factor(weekday)sun               -2.270e-03  1.060e-02  -0.214 0.830367
factor(weekday)thur               5.044e-02  1.044e-02   4.831 1.36e-06 ***
factor(weekday)tue                7.317e-02  1.036e-02   7.066 1.59e-12 ***
factor(weekday)wed                7.504e-02  1.040e-02   7.214 5.43e-13 ***
factor(year)1995                  3.461e-01  2.908e-02  11.902  < 2e-16 ***
factor(year)1996                  4.330e-01  2.869e-02  15.091  < 2e-16 ***
factor(year)1997                  5.176e-01  2.845e-02  18.194  < 2e-16 ***
factor(month)2:factor(year)1995   1.166e-01  4.161e-02   2.802 0.005085 **
factor(month)3:factor(year)1995   8.994e-02  3.995e-02   2.251 0.024370 *
factor(month)4:factor(year)1995   1.537e-01  4.091e-02   3.757 0.000172 ***
factor(month)5:factor(year)1995   1.198e-01  4.064e-02   2.949 0.003193 **
factor(month)6:factor(year)1995  -2.101e-02  4.106e-02  -0.512 0.608912
factor(month)7:factor(year)1995  -1.600e-01  3.999e-02  -4.001 6.30e-05 ***
factor(month)8:factor(year)1995   1.971e-02  4.089e-02   0.482 0.629792
factor(month)9:factor(year)1995  -4.851e-02  4.123e-02  -1.177 0.239342
factor(month)10:factor(year)1995 -2.575e-01  4.054e-02  -6.351 2.14e-10 ***
factor(month)11:factor(year)1995 -1.232e-01  4.071e-02  -3.026 0.002475 **
factor(month)12:factor(year)1995 -9.372e-02  4.090e-02  -2.291 0.021944 *
factor(month)2:factor(year)1996   8.385e-02  4.142e-02   2.024 0.042922 *
factor(month)3:factor(year)1996   1.818e-01  3.903e-02   4.658 3.20e-06 ***
factor(month)4:factor(year)1996   2.880e-01  4.042e-02   7.124 1.05e-12 ***
factor(month)5:factor(year)1996   1.834e-01  4.004e-02   4.581 4.63e-06 ***
factor(month)6:factor(year)1996   1.525e-01  4.016e-02   3.798 0.000146 ***
factor(month)7:factor(year)1996   1.209e-01  3.896e-02   3.104 0.001909 **
factor(month)8:factor(year)1996  -3.211e-02  4.012e-02  -0.801 0.423417
factor(month)9:factor(year)1996  -1.016e-01  4.045e-02  -2.511 0.012038 *
factor(month)10:factor(year)1996 -2.239e-01  3.961e-02  -5.652 1.58e-08 ***
factor(month)11:factor(year)1996 -1.602e-01  4.003e-02  -4.003 6.26e-05 ***
factor(month)12:factor(year)1996 -6.502e-02  4.014e-02  -1.620 0.105253
factor(month)2:factor(year)1997  -2.274e-03  4.093e-02  -0.056 0.955693
factor(month)3:factor(year)1997  -1.220e-01  3.933e-02  -3.101 0.001926 **
factor(month)4:factor(year)1997   5.416e-02  4.018e-02   1.348 0.177625
factor(month)5:factor(year)1997   2.281e-01  3.940e-02   5.788 7.13e-09 ***
factor(month)6:factor(year)1997   2.063e-01  3.979e-02   5.185 2.16e-07 ***
factor(month)7:factor(year)1997  -5.928e-02  3.920e-02  -1.512 0.130506
factor(month)8:factor(year)1997  -1.261e-01  3.997e-02  -3.154 0.001610 **
factor(month)9:factor(year)1997  -1.400e-01  4.022e-02  -3.481 0.000499 ***
factor(month)10:factor(year)1997 -2.958e-01  3.952e-02  -7.484 7.22e-14 ***
factor(month)11:factor(year)1997 -1.928e-01  3.980e-02  -4.846 1.26e-06 ***
factor(month)12:factor(year)1997  1.947e-02  4.078e-02   0.477 0.633028
---
Signif. codes:  0 ‘’*** 0.001 ‘’** 0.01 ‘’* 0.05 ‘’. 0.1 ‘’  1

(Dispersion parameter for Negative Binomial(183.1461) family taken to be 1)

    Null deviance: 8551.6  on 1444  degrees of freedom
Residual deviance: 1444.5  on 1386  degrees of freedom
AIC: 12716

Number of Fisher Scoring iterations: 1

              Theta:  183.1
          Std. Err.:  13.6

 2 x log-likelihood:  -12595.57
```