Detection and correction of violations of the basic linear assumptions: case study with age-adjusted mortality data

Sandro Petrillo (Gibo) and Roberto Stoppa (Bob)
University of Neuchâtel - Master Degree in Statistics
Project for the course *Robust Regression Diagnostics*, Prof. A.S. Hadi

February 2005

Contents

1 Statement of the problem **3**

2 Data description **3**
2.1 Source of the data 3
2.2 Detailed description of the variables 3
2.3 Intuition hypothesis of response and predictors variables 4
2.4 Initial model and assumptions 6

3 Data analysis **7**
3.1 Graphs before fitting a model 7
3.1.1 One-dimensional graphs 7
3.1.2 Two-dimensional graphs 8
3.2 Initial model fit to the data 9
3.2.1 Estimation of the full model 9
3.2.2 Multicollinearity and variable selection 10
3.2.3 Backward elimination procedure 11
3.3 Graphs after fitting the model to the data 12
3.3.1 Graphs for checking the linearity and normality assumptions 12
3.3.2 Graphs for the detection of outliers and influential observations 13
3.3.3 Graphs for the effect of variables 14

4 Summary and conclusions **14**
4.1 Final model 15
4.2 Conclusions 15

A List of the data set **18**

B Tables **20**

C Graphics **22**

List of Tables

1 Summary statistics of the variables 8
2 Estimate of the initial, full model 10
3 VIF's after elimination of log(HCPot) 11
4 Steps of the backward elimination procedure 11
5 Estimation of the final model without outliers 15
6 Correlation matrix of all the variables 20
7 Results of omitting influential points in the reduced model 21
8 Results of omitting outliers in the Y-direction in the reduced model 21

List of Figures

1 Histograms of all the variables 7
2 Histograms of logarithmic transformation of HCPot, NOxPot and S02Pot 8
3 Pairwise scatterplots of Mortality, log(HCPot), log(NOxPot) and log(S02Pot) 9
4 Boxplots of all the variables 22
5 Normal probability plot of reduced model's residuals 22
6 Added variable plots before the backward elimination procedure 23
7 Pairwise scatter plots of all variables 24
8 Scatter plots of the standardized residuals vs. each of the predictor 25
9 Scatter plot of the standardized residuals vs. the fitted values . . 26
10 Index plot of the the leverage values p_{ii} (reduced model) 26
11 Index plot of the Cook's distance and the Hadi's Influence Measure (reduced model) 27
12 Potential-Residual plot (reduced model) 27
13 Added variable plot for the reduced model 28
14 Normal probability plot for the final model 29
15 Scatter plot of the standardized residuals vs. the fitted values (final model) 29
16 Scatter plots of the standardized residuals vs. each predictor (final model) 30

1 Statement of the problem

Air pollution in cities has been linked to increased rates of mortality in developed and developing countries. Although these findings have helped lead to a tightening of air-quality standards, their validity with respect to public health has been questioned.

In order to analyze the relation between air pollution and mortality, this work takes data collected on U.S. Standard Metropolitan Statistical Areas in 1980. The aim is to analyze mortality rates in U.S. Standard Metropolitan Statistical Areas (cross-section analysis) using some components of air pollution (3 different air pollutants), climate and socio-economic characteristics of the cities' population.

The aim is to find out which variable is likely to influence the mortality and in which way. One side of the analysis is to fit a multiple linear regression model that describes the relationship among variables using a regression diagnostics procedure. The distribution theory , confidence intervals, and tests of hypotheses in order to fit a regression model are valid and have meaning only if the standard regression assumptions are satisfied. Moreover we would like to ensure that the fit is not overly determined by one or few observations. When these assumptions are violated the results do not hold and an application of them may lead to serious errors (S. Chatterjee, A. S. Hadi, B. Price 2000, p. 85 [2]). The focus of this article is to detect and correct violations (if any) of the basic linear assumptions.

2 Data description

2.1 Source of the data

A cross-section data set of 60 U.S. Standard Metropolitan Statistical Areas (SMSA's) was collected by researchers at General Motors in a study of whether air pollution contributes to mortality. The dependent variable for analysis is age adjusted mortality (called "Mortality"). The data include variables measuring demographic characteristics of the cities, variables measuring climate characteristics, and variables recording the pollution potential of three different air pollutants. Data refers to year 1980.
The dataset has been downloaded from url http://lib.stat.cmu.edu/DASL/Datafiles/SMSA.html.
One of the observations had to be deleted because there were missing values for two variables[1]. Final dataset represents therefore observations of 59 SMSA's.

2.2 Detailed description of the variables

The dependent variable (response variable) express the total age adjusted mortality, that is, the number of deaths from all causes per 100'000 people per year in each of the 59 cities (SMSA's).

[1] Deleted observation is referred to the city of Fort Worth, TX.

The independent variables (explanatory variables or predictors) are characterized by three groups of features: socio-economics, climate and air pollution variables.

- Socio-economics variables:
  - Education: Median education (in years per person)
  - PopDensity: Population density (people per square mile)
  - Income: Median income (in US dollars per year)
- Climate variables:
  - JanTemp: Mean January temperature (degrees Farenheit)
  - JulyTemp: Mean July temperature (degrees Farenheit)
  - RelHum: Relative Humidity
  - Rain: Annual rainfall (inches)
- Air pollution variables:
  - HCPot: Hydrocarbon pollution potential
  - NOxPot: Nitrous Oxide pollution potential
  - SO2Pot: Sulfur Dioxide pollution potential

2.3 Intuition hypothesis of response and predictors variables

Before performing any analysis, we develop a reasoning about the effects of each explanatory variable on the response variable.

- Socio-economics variables:
  - **Education**: for this variable we expect a negative relationship with mortality because we think that people with more education may be more informated about health's risks;
  - **PopDensity**: we think that in areas with high density people may be more stressed and this may have a negative influence on the quality of the life. Therefore we expect a positive relationship between population density and mortality;
  - **Income**: people with high income have more possibilities to take care on health. For example rich people could afford more wellness care, etc. For this reasons, we expect a negative relationship between income and mortality.
- Climate variables: all climate and weather variables have some influence on human health. The effect may be either directly on the human body or indirectly through effects on disease-causing organisms or their vectors. Although the effects of variation of only one weather element may be examined in a particular study, that element does not act independently of other elements, for example, changes in humidity modify the effects of temperature.

- **JanTemp**: the proportion of annual deaths occurring in January would be expected to decline in a warmer climate and increase in a cold climate that's means a negative relationship with the response variable;
- **JulyTemp**: heat waves[2] are supposed to have a positive influence on mortality. High temperatures can cause great discomfort to people like brain dysfunction (e.g. a stroke or convulsions), hot and dry skin, and a body-core temperature above 40^o C (104 degrees Farenheit);
- **RelHum**: humidity has an important impact on mortality since it contributes to the body's ability to cool itself by evaporation of perspiration. It also has an important influence on morbidity in the winter because cold, dry air leads to excessive dehydratation of nasal passages and the upper respiratory tract and increased chance of microbial and viral infection (see for example Kalkstein L. S. and Valimont K. M. (1987) [7]);
- **Rain**: rainfalls can have positive or negative influence on mortality. The precipitation event itself might have an indirect impact, as the cooler temperatures coinciding with a summer rainfall provide relief from excessively warm weather. However, in certain specific cases, rainfall might induce increases in mortality. Some researchers found that fatal automobile accidents increased in frequency during very light rain episodes (less than .01 inch) and heavy rainfalls (greater than 0.1 inch per hour).

- Air pollution variables:
  - **HCPot**: eye and respiratory tract irritation, headaches, dizziness, visual disorders, and memory impairment are among the immediate symptoms that some people have experienced soon after exposure to some organics. Hydrocarbons come from motor vehicles (evaporation from gas tanks), industry, and from various household products. We expect HCPot to have a positive influence on mortality;
  - **NOxPot**: it comes from motor vehicles and industry (burning fossil fuels), can react with other gases in atmosphere to form nitric acid (HNO_3) ("acid rain"). We expect a positive relation between NOxPot and mortality;
  - **SO2Pot**: exposure to sulfur dioxide can cause impairment of respiratory function, aggravation of existing respiratory disease (especially bronchitis), and a decrease in the ability of the lungs to clear foreign particles. Groups that appear most sensitive to the effects of sulfur dioxide include asthmatics and other individuals with hyperactive airways, and individuals with chronic obstructive lung or cardiovascular disease. Elderly people and children are also likely to be more sensitive to sulfur dioxide. We expect a positive influence of SO2Pot on mortality.

[2] A common definition of heat waves is a series of days with maxima over 35^o C (95 degrees Farenheit)

2.4 Initial model and assumptions

The initial model (multiple linear regression model) constists of a dependent variable (Y) and some independent variables ($X_1, X_2, \ldots X_p$). One goal of the regression is to build a linear equation that allows us to describe, predict and control a dependent variable on the basis of one ore more predictor variables. Equation (1) describes the initial model:

$$\begin{aligned} \text{Mortality} &= \beta_0 + \beta_1 \text{Education} + \beta_2 \text{PopDensity} + \beta_3 \text{income} + \\ &+ \beta_4 \text{JanTemp} + \beta_5 \text{JulyTemp} + \beta_6 \text{RelHum} + \beta_7 \text{Rain} + \\ &+ \beta_8 \text{HCPot} + \beta_9 \text{NOxPot} + \beta_{10} \text{S0xPot} + \varepsilon \end{aligned} \tag{1}$$

where $\beta_0, \beta_1, \ldots, \beta_p$ are the parameters of the model and ε is a random disturbance or error. As already said the aim of this article is to detect and correct violations of the basic linear model assumptions. Estimation results of a regression model are valid only if the standard regression assumptions are satisfied. In other words the properties of least squares estimators and the statistical analysis are based on the following assumptions:

1. **linearity**: the model is linear in regression coefficients, that means a single observation can be written as

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \varepsilon_i \qquad i = 1, 2, \ldots, n$$

2. **errors**: the errors $\varepsilon_1, \ldots, \varepsilon_n$ are assumed to be independently and identically distributed (iid) normal random variables each with mean zero and a constant variance σ^2;

3. **predictor variables**: the predictor variables $X_1, X_2, \ldots, X_p$ are nonrandom, measured without errors and are assumed to be linearly independent of each other;

4. **about the observations**: all observations are equally reliable and have equal influence in determining the least squares results and the conclusions drawn from the results.

Gross violations of the model assumptions can, however, seriously distort conclusions (S. Chatterjee, A. S. Hadi, B. Price 2000, p. 85 [2]).

3 Data analysis

3.1 Graphs before fitting a model

3.1.1 One-dimensional graphs

Data analysis usually begins with the examination of each variable in the study, in order to have a general idea about the distribution of each single variable. We use histograms and box plots to examine each variable. The purpose is to analyze if variables must be transformed, in order to reduce skewness and achieve symmetry. The first two rows of histograms in Figure 1 show distribution of

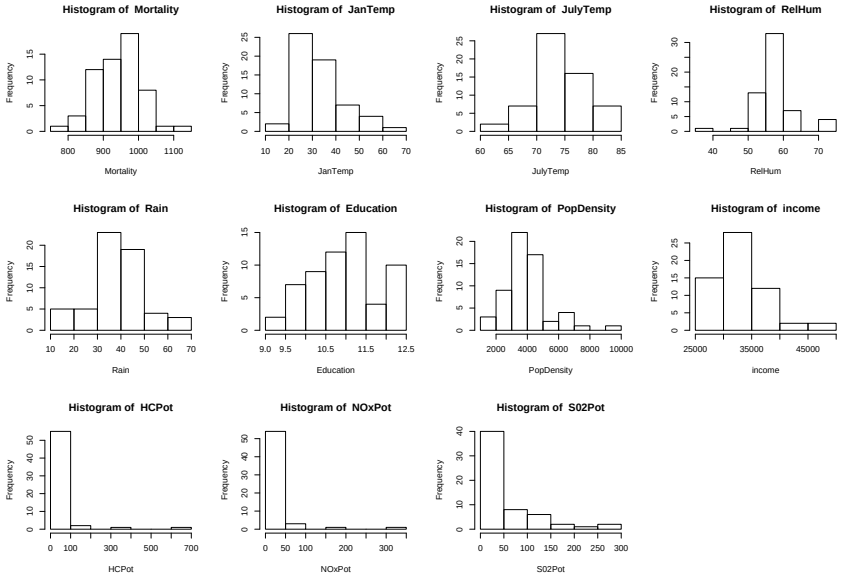

Figure 1: Histograms of all the variables

variables almost symmetric. However, in the last row of Figure 1 distributions of the three air pollution variables look very asymmetric and skewed to the right. This suggests that variable transformation is needed in order to satisfy the standard assumption about the errors made in regression analysis. Another graphical way to see variable's distribution is shown in Figure 4 in appendix **C**, where box plots are used.

Logarithmic transformation of data is particularly useful when the variable analyzed has a large standard deviation compared to its mean (S. Chatterjee, A. S. Hadi, B. Price 2000, p. 170 [2]). Table 1 indicates that HCPot, NOxPot and S02Pot have a large value of stdev/mean ratio, confirming again that variable transformation is needed. The logarithmic transformation of these three variables enables to reach a more symmetric distribution (see Figure 2) and to reduce stdev/mean ratio[3], so that the magnitude of this ratio is now similar among all the variables.

[3]Stdev/mean ratio of the log-transformation of HCPot, NOxPot and S02Pot are 0.4030, 0.5078 and 0.4459 respectively

Table 1: Summary statistics of the variables

	Stdev	Mean	Stdev/Mean
Mortality	62.42	941.17	0.07
JanTemp	10.15	33.80	0.30
JulyTemp	4.60	74.41	0.06
RelHum	5.38	57.75	0.09
Rain	11.57	38.51	0.30
Education	0.85	10.97	0.08
PopDensity	1441.69	3910.49	0.37
income	4473.10	33246.66	0.13
HCPot	92.64	38.47	2.41
NOxPot	46.67	22.97	2.03
S02Pot	63.55	54.66	1.16

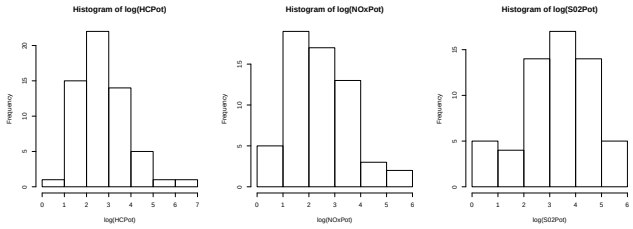

Figure 2: Histograms of logarithmic transformation of HCPot, NOxPot and S02Pot

3.1.2 Two-dimensional graphs

The purpose of these pairwise scatterplots are to explore the relationships between each pair of variables.
In multiple regression, the scatterplots of Y versus each predictor variable may or may not show linear patterns. The absence of such a pattern does not imply that the linear model is incorrect. The pairwise scatterplot of the predictors should show no linear pattern because the predictors are assumed to be linearly independent (see assumption 3).
For reasons of clear graphical presentation, Figure 3 shows pairwise scatter plots of the Y and three predictor variables. A complete pairwise scatter plots is shown in appendix **C**, Figure 7. In addition to graphical analysis the relation of two variables can be checked with the correlation matrix (see Table 6 in appendix **B**).
The first row of Figure 3 indicates a presumed linear positive relation between

Y (Mortality) and the three predictor variables, but as said before, this is not necessary in multiple linear regression. However, the most important information of these scatter plots are those between all the predictor variables (second trough tenth row of Figure 7 on page 24). Figure 3 is an extrapolation of Figure 7 and we can see a linear pattern between log(HCPot), log(NOxPot) and log(S02Pot). This seems to be a violation of the independence assumption due to collinearity problem. Even if the correlation matrix and pairwise scatter plots confirm a strong linear relationship among these three variables, it doesn't mean that multicollinearity problems will surely arise (see section 3.2.2 for multicollinearity check).

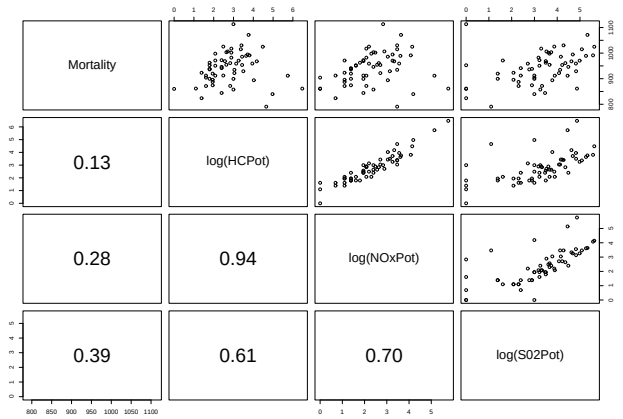

Figure 3: Pairwise scatterplots of Mortality, log(HCPot), log(NOxPot) and log(S02Pot)

3.2 Initial model fit to the data

3.2.1 Estimation of the full model

After the explorative analysis developed in the preceding points, we begin to estimate the initial model with all variables and a logarithmic transformation of three predictor variables suggested by the skewness and asymmetry of their distributions. The initial, full model, is described by the following equation:

$$\begin{aligned} \text{Mortality} &= \beta_0 + \beta_1 \text{JanTemp} + \beta_2 \text{JulyTemp} + \beta_3 \text{RelHum} + \\ &+ \beta_4 \text{Rain} + \beta_5 \text{Education} + \beta_6 \text{PopDensity} + \beta_7 \text{income} + \\ &+ \beta_8 \log(\text{HCPot}) + \beta_9 \log(\text{NOxPot}) + \beta_{10} \log(\text{S0xPot}) + \varepsilon \quad (2) \end{aligned}$$

The signs of the estimated coefficients (see Table 2) of each variable are consistent with our expectations except for two of them: population density (PopDensity) and hydrocarbon pollution potential (log(HCPot)). Even though we got these results, the model with all the variables seems to emphasize some problems. More precisely, the model indicates that only two predictor variables have significant coefficients (p-value <0.05). Since the set of variables to include in the model has not yet been decided, we have to develop a variable-selection

Table 2: Estimate of the initial, full model

	Estimate	Std. Error	t value	Pr(>|t|)	VIF
(Intercept)	532.1539	227.0582	2.34	0.0233	
JanTemp	−1.2788	0.8929	−1.43	0.1586	2.43
JulyTemp	5.2621	1.9685	2.67	0.0102	2.42
RelHum	1.5530	1.3310	1.17	0.2491	1.52
Rain	2.0788	0.7213	2.88	0.0059	2.06
Education	−13.1533	9.4379	−1.39	0.1698	1.90
PopDensity	−0.0021	0.0047	−0.44	0.6584	1.36
income	−0.0010	0.0016	−0.62	0.5361	1.59
log(HCPot)	−4.4789	18.9542	−0.24	0.8142	13.52
log(NOxPot)	32.5840	17.7349	1.84	0.0724	13.12
log(S02Pot)	4.2160	8.7052	0.48	0.6304	4.70
Adj. R^2:	**0.4961**				

procedure in order to choose variables that will be included in the equation. Our method consists in detecting problems of collinearity, delete variables causing it (one at a time) until we obtain a set of variables without major problems of collinearity. After that, we apply a backward elimination procedure with the purpose to find out a good model (not the "best" model, but a "good" model).

3.2.2 Multicollinearity and variable selection

The *variance inflation factor*[4] (VIF) is an indicator that measures linear relationship between the predictor variables. A value of VIF equal to 1 means a complete absence of linear relationship between a predictor variable and all the others. Values of variance inflation factors greater than 10 is often taken as a signal that the data have collinearity problems (S. Chatterjee, A. S. Hadi, B. Price 2000, p. 240 [2]). The VIF's for the full model (see Table 2) indicate two values greater than 10: log(HCPot) and log(NOxPot)[5]. This suggests to leave out one of these two variables and recompute the VIF's to check the collinearity on the reduced subset of predictor variables.

Deleting log(HCPot) (which has the greatest value of VIF) we obtain new values of VIF's contained in a range between 1.36 and 4.68 (see Table 3). This result is quite satistactory because it seems that strong multicollinearity is no more present. We can conclude that the data in the reduced dataset are not seriously collinear and for this reason we can apply a variable selection procedure.

[4] $VIF_j = 1/(1 - R_j^2) \qquad j = 1, 2, \ldots, p$ where p is the number of predictor variables and R_j^2 represents the square of the multiple correlation coefficient that results when the predictor variable X_j is regressed against all the other predictor variables.

[5] Collinearity problem of these two variables could also be seen in a pairwise scatter plot on Figure 7, p. 24.

Table 3: VIF's after elimination of log(HCPot)

	JanTemp	JulyTemp	RelHum	Rain	Education
VIF	2.13	2.18	1.51	1.97	1.79
	PopDensity	income	log(NOxPot)	log(S02Pot)	
VIF	1.36	1.58	4.68	4.63	

3.2.3 Backward elimination procedure

We start the backward elimination procedure with the following equation:

$$\begin{aligned} \text{Mortality} &= \beta_0 + \beta_1 \text{JanTemp} + \beta_2 \text{JulyTemp} + \beta_3 \text{RelHum} + \\ &+ \beta_4 \text{Rain} + \beta_5 \text{Education} + \beta_6 \text{PopDensity} + \beta_7 \text{income} + \\ &+ \beta_8 \log(\text{NOxPot}) + \beta_9 \log(\text{S0xPot}) + \varepsilon \end{aligned} \tag{3}$$

The procedure consists of fit the model in equation (3) and successively drop

Table 4: Steps of the backward elimination procedure

	Step 1	Step 2	Step 3	Step 4	Step 5
(Intercept)	524.7522	517.8053	551.3539	542.0700	686.2891
t-values	(2.356)	(2.349)	(2.742)	(2.710)	(4.337)
JanTemp	−1.3527	−1.3468	−1.5257	−1.5976	−1.3815
	(−1.634)	(−1.640)	(−2.262)	(−2.399)	(−2.151)
JulyTemp	5.4079	5.3889	5.2606	5.2909	4.2538
	(2.921)	(2.935)	(2.937)	(2.966)	(2.736)
RelHum	1.5319	1.5824	1.4466	1.4498	
	(1.165)	(1.217)	(1.165)	(1.172)	
Rain	2.1131	2.0725	2.1419	2.1820	2.2858
	(3.020)	(3.011)	(3.249)	(3.332)	(3.5111)
Education	−13.6832	−13.2380	−13.7466	−16.5785	−15.9749
	(−1.507)	(−1.479)	(−1.565)	(−2.075)	(−1.996)
PopDensity	−0.0021				
	(−0.452)				
Income	−0.0009	0.0010	−0.0012		
	(−0.613)	(−0.681)	(−0.791)		
log(HCPot)					
log(NOxPot)	29.2235	28.8360	32.1712	31.4057	30.8687
	(2.875)	(2.779)	(5.565)	(5.531)	(5.435)
log(S02Pot)	3.9603	3.2395			
	(0.463)	(0.389)			
Adj. R^2:	**0.5058**	**0.5137**	**0.5218**	**0.5252**	**0.5219**

(one at a time) the less significant variable (the one that has the smallest t-test in the equation), until all the t-tests are significants. The results of the procedure

are summarized in Table 4. After this procedure the model we obtain is:

$$\begin{aligned} \text{Mortality} &= \beta_0 + \beta_1 \text{JanTemp} + \beta_2 \text{JulyTemp} + \beta_3 \text{Rain} + \\ &+ \beta_4 \text{Education} + \beta_5 \log(\text{NOxPot}) + \varepsilon \end{aligned} \tag{4}$$

The signs of the estimated β coefficients are in line with our expectations (see section 2.3). The *added variable plots* for every covariate before the backward elimination procedure show significant linear relationships only for the five predictors we retained in the above model and this is another confirmation that the t-test values are reliable. At the same time, in these plots we see that the variables eliminated show no strong linear relationship with the response variable (see Figure 6 on page 23).

Next section introduces the diagnostics steps necessary to investigate the validity of assumptions and adequacy of the fitted model (equation (4)). Only if the assumptions behind the model are satistactory we can trust in the results we obtained, otherwise we should reformulate the model after having found out what is the problem.

3.3 Graphs after fitting the model to the data

3.3.1 Graphs for checking the linearity and normality assumptions

The first two assumptions we mentioned above are linearity and normality. In order to check them we can examine graphically the *residuals* of the fitted model's estimation.

Figure 5 in appendix **C** represents the ordered standardized residuals[6] versus their normal scores. If the normality assumption holds, then the plot should resemble a straight line with intercept zero and slope one (which are the mean and the standard deviation of the standardized residuals). A look at Figure 5 indicates quite a good normal approximation for most of the residuals but it shows also clearly the presence of some extreme points (outliers).

Figure 8 in appendix **C** reproduces scatter plots of the standardized residuals versus each of the predictor variables. Any discernible pattern in these plots may indicate violation of normality and/or linearity assumptions. In particular normality assumption should show homogeneity of the variance that means random scattered points almost symmetrically distributed around zero and comprised between +/- 3. Values greater than +/- 3 could be potential outliers. Looking at Figure 8 on page 25, violation of the assumptions seem not to be present. In particular the assumption of constant variance (homoscedasticity) holds and the assumption of linearity seems also to hold, because any discernible pattern is not present. However three observations need to be checked in the outliers detection procedure[7].
Even Figure 9 in appendix **C** indicates a random scatter of points and this is a valid concept to validate the standard assumptions of linearity and normality.

[6] For simplicity we call them "standardized" residuals. Actually, they are the "internally studentized residuals" ($r_i = \frac{e_i}{\hat{\sigma}\sqrt{1-p_{ii}}}$).

[7] These are observations number 27, 36 and 58 that we will investigate further.

This three operations allow us to see which city is poorly fit by the model and may cause problems. For example in Figure 9 on page 26 we notice that the observations number 27, 36 and 58 (representing the cities of Lancaster, New Orleans and York) have the largest residuals in absolute value, close to 3.

3.3.2 Graphs for the detection of outliers and influential observations

Outliers can also occur in the predictor variables. With the analysis of the residuals, these kind of outliers may not be detected because of masking and swamping problems and/or presence of high leverage points. Having an important influence on the regression results, additional measures of the influence of the observations are needed.

As seen in the above point, outliers in the response variable are those with a large value of the standardized residual (+/- 3). While three observations were identified as outliers in the Y-space (observations number 27, 36 and 58), Figure 10 on page 26 indicates six outliers in the X-space and two that lie "near" the line. These points, known as high leverage points, are those that lie over the horizontal dotted line[8]. These eight points were not found as outliers because of their small residuals. Even if small values of residuals is desirable for a good fit of a model, here the reason for the small values of residuals may be due to the fact that these eight observations pull the regression estimation toward them, influencing the estimation results (we will check after this assertion).

Measures of influence

The *influence* of an observation is measured by the effects it produces on the fit when it is deleted in the fitting process (S. Chatterjee, A. S. Hadi, B. Price 2000, p. 103 [2]). Two methods to measure the influence of an observation are the Cook's Distance and the Hadi's Influence Measure (see Figure 11 on page 27). Both graphs indicate that some observation points are to be treated with caution because of their influence on the results. In particular the observations number $27, 28, 31, 33, 36, 48, 49, 55, 57, 58$ are those with the most important influence, according to these two measures. Points with high leverage that are not influential do not cause problems (observations number $16, 46, 47$ and 56). However high leverage points that are influential should be investigated because these points are outlying as far as the predictor variables are concerned and also influence the fit (observations number $28, 31, 49$ and 57). From the Hadi's Influence Measure formula we can draw a graph to aid in classifying unusual observations as high leverage points, outliers, or a combination of both (see figure 12 on page 27). To get an idea of the sensivity of the analysis to these points, the model should be fitted without the offending points and the resulting coefficients examined (S. Chatterjee, A. S. Hadi, B. Price 2000, p. 108 [2]). Deleting the high leverage points that are also influential (one at a time) does not change substantially the regression results obtained with all the obser-

[8]The value of the horizontal line represents 2 times the mean of all the leverage values (p_{ii}). A leverage value indicates the influence of the i-th observation on the i-th fitted value. Observations number $16, 28, 31, 46, 47$ and 57 lie over the line, while observations number 49 and 56 lie just below the line

vations (see Table 7, appendix **B** on page 21). This means that the detected influential points, deleted one at a time, do not affect hardly the regression results and for this reason we decide to keep the observations in the model.

Last but not least we should analyze the problem of the three outliers we found above, that were discovered as influential points too. We decide to estimate our model without these outliers, deleting one at a time and then all together to see if their influence is important for the results. On the contrary to the results obtained in the above paragraph, the deletion of outliers has an important effect for the coefficient of the variable Education and also the adjusted R^2 improves (see Table 8, appendix **B** on page 21).

Looking at the three outliers that represent cities of Lancaster, New Orleans and York, we can say that even if Lancaster has all the values of the predictor variables near the centered values of the sample, the age adjusted mortality is much lower than the average. For New Orleans we notice that the age adjusted mortality rate is the highest in spite of the values of the predictor variables are near the average values. As regards the city of York, we saw that age adjusted mortality is in the average, in spite of extreme values in variables Rain and Education. We can explain these three situations saying that maybe the model is inadequate for this three cities because other reasons, not present in our model and unknown to us, may affect the age adjusted mortality.

3.3.3 Graphs for the effect of variables

As said before, the model has five significant variables. In order to interpret the importance of each variable, we need to compare the t-test in conjunction with the graphical analysis called *added variable plot*.
The slope of the points in the plot gives the magnitude of the regression coefficient of the variable if it were brought into the equation. Thus, the stronger the linear relationship in the added-variable plot is, the more important is the additional contribution of X_j to the regression equation already containing the other predictors. The scatter of the points will also indicate visually which of the data points are most influential in determining this slope and its corresponding t-test (S. Chatterjee, A. S. Hadi, B. Price 2000, p. 110 [2]).

In this case the five added variable plots (see Figure 13 on page 28) seem to validate the t-test results that each predictor variable contributes significantly to the regression equation . The graphs, however, indicate that points $27, 36$ and 58 influence the results and this confirms the conclusions we found with the outlier detection.

4 Summary and conclusions

What we have done until now is a sequence of steps that may be used to fit a linear regression model satistactorily. In particular we started from a set of variables and tried to find out a good linear model that could fit as well as possible the data. Before taking any conclusion, attempt should be made to

validate the fitted model.

4.1 Final model

The sensitivity procedure applied on *variables* allowed us to find the following model:

$$\begin{aligned} \text{Mortality} &= \beta_0 + \beta_1 \text{JanTemp} + \beta_2 \text{JulyTemp} + \beta_3 \text{Rain} + \\ &+ \beta_4 \text{Education} + \beta_5 \log(\text{NOxPot}) + \varepsilon \end{aligned} \tag{5}$$

The sensitivity analysis applied to *observations* has suggested us to drop three of them because of an evident influence on the results. Table 5 gives the estimation of the final model.

Table 5: Estimation of the final model without outliers

	Estimate	Std. Error	t value	Pr(>|t|)	VIF
(Intercept)	868.2495	129.5018	6.70	0.0000	
JanTemp	−1.4937	0.5197	−2.87	0.0059	1.314
JulyTemp	3.2173	1.2313	2.61	0.0118	1.567
Rain	2.3962	0.5204	4.60	0.0000	1.640
Education	−24.8549	6.7718	−3.67	0.0006	1.363
log(NOxPot)	28.7801	4.4758	6.43	0.0000	1.427
Adj. R^2:	**0.6622**				

In order to validate these results we have to check variance inflation factors and residuals distribution. The VIF's (last column of Table 5) are comprised between 1.31 and 1.64, which indicate no strong problems of collinearity. With the deletion of the three outliers also the normal assumption of the residuals seems improved (see Figure 14 on page 29). We also notice that the standardized residuals are contained in a "reasonable" range (-2.49; 2.08) (see Figures 15 and 16, pp. 29 and 30) and no outliers seem to have been masked by the three dropped observations.

4.2 Conclusions

The final model allow us to confirm that the t-values of the five variables increase in absolute values and also the adjusted R^2 grows from 0.52 to 0.66. This means that the coefficients have become more significant and more of the variance of the response variable is linearly explained by the five predictors. The comments on the results should be read under the considerations we made for selecting the final model and their assumptions (e.g. number of variables).

The signs of the coefficients of the variables JanTemp, JulyTemp, Education and NOxPot confirm our intuitions (and also some other researches done in this field) about the relation they have with mortality. As regards the coefficient of the variable Rain we found a positive relationship with mortality. As said this variable could have a positive or a negative relationship with mortality because rain could have positive influence by too dry seasons and may have negative

impact if mortal cars accidents are related with rain.

Anyway in order to obtain more reliable information about the relationship of mortality with the variables we used in the model a time series analysis would be suggested. With an analysis of cross-section it may happen that some data information are distorted by exceptional events (e.g. exceptional rain seasons, very hot summers, etc.).

All the work was done using the statistical software package R[9] and text was written with LaTeX.

. . . statistical model building is an art!
(S. Chatterjee, A. S. Hadi, B. Price 2000, p. 311 [2])

[9] http://www.r-project.org

References

[1] Birkes D., Dodge Y. (1993), *Alternative Methods of Regression*, John Wiley & Sons, Inc.

[2] Chatterjee S., Hadi A.S., Price B. (2000), *Regression Analysis by Example*, New York: Wiley

[3] Chatterjee S., Hadi A.S. (1988), *Sensitivity Analysis in Linear Regression*, New York: Wiley

[4] Dodge Y. and Rousson V. (2004), *Analyse de Régression Appliquée*, Dunod, Paris, 2^e édition.

[5] Green W.H. (1997), *Econometric Analysis*, Prentice-Hall International, New Jersey, Third Edition.

[6] Griffiths W.E., Hill R.C. (1993), *Learning and Practicing Econometrics*, John Wiley & Sons, Inc., New York.

[7] Kalkstein L. S. and Valimont K. M. (1987), *Climate effects on human health.* In Potential effects of future climate changes on forests and vegetation, agriculture, water resources, and human health. EPA Science and Advisory Committee Monograph no. 25389, 122-52. Washington, D.C.: U.S. Environmental Protection Agency.

[8] Maindonald J., Braun J. (2003), *Data Analysis and Graphics Using R*, Cambridge University Press.

[9] Rousseuw P. J. and Leroy A. M. (1987), *Robust regression and outlier detection*, New York: John Wiley & Sons.

[10] Samet J.M., Zeger S.L. , Dominici F., Curriero F., Coursac I., Dockery D.W., Schwartz J., Zanobetti A. (2000), *The National Morbidity, Mortality, and Air Pollution Study Part II: Morbidity and Mortality from Air Pollution in the United States*, in Health Research Report Number 94, Part II June 2000.

[11] Schwartz J., Marcus A. (1990), *Mortality and air pollution in London: a time series analysis*, in American Journal of Epidemiology, Vol 131, Issue 1 185-194.

[12] Venables W.N., Ripley B.D. (2002), *Modern Applied Statistics with S*, Springer-Verlag New York, Inc., Fourth Edition.

A List of the data set

city Mortality JanTemp JulyTemp RelHum Rain Education PopDensity income HCPot NOxPot SO2Pot
Akron, OH 921.87 27 71 59 36 11.4 3243 29560 21 15 59
Albany-Schenectady-Troy, NY 997.87 23 72 57 35 11 4281 31458 8 10 39
Allentown, Bethlehem, PA-NJ 962.35 29 74 54 44 9.8 4260 31856 6 6 33
Atlanta, GA 982.29 45 79 56 47 11.1 3125 32452 18 8 24
Baltimore, MD 1071.29 35 77 55 43 9.6 6441 32368 43 38 206
Birmingham, AL 1030.38 45 80 54 53 10.2 3325 27835 30 32 72
Boston, MA 934.7 30 74 56 43 12.1 4679 36644 21 32 62
Bridgeport-Milford, CT 899.53 30 73 56 45 10.6 2140 47258 6 4 4
Buffalo, NY 1001.9 24 70 61 36 10.5 6582 31248 18 12 37
Canton, OH 912.35 27 72 59 36 10.7 4213 29089 12 7 20
Chattanooga, TN-GA 1017.61 42 79 56 52 9.6 2302 25782 18 8 27
Chicago, IL 1024.89 26 76 58 33 10.9 6122 36593 88 63 278
Cincinnati, OH-KY-IN 970.47 34 77 57 40 10.2 4101 31427 26 26 146
Cleveland, OH 985.95 28 71 60 35 11.1 3042 35720 31 21 64
Columbus, OH 958.84 31 75 58 37 11.9 4259 29761 23 9 15
Dallas, TX 860.1 46 85 54 35 11.8 1441 38769 1 1 1
Dayton-Springfield, OH 936.23 30 75 58 36 11.4 4029 30232 6 4 16
Denver, CO 871.77 30 73 38 15 12.2 4824 39099 17 8 28
Detroit, MI 959.22 27 74 59 31 10.8 4834 33858 52 35 124
Flint, MI 941.18 24 72 61 30 10.8 3694 32000 11 4 11
Grand Rapids, MI 871.34 24 72 61 31 10.9 3226 29915 5 3 10
Greensboro-Winston-Salem-HighP, NC 971.12 40 77 53 42 10.4 2269 29450 8 3 5
Hartford, CT 887.47 27 72 56 43 11.5 2909 37565 7 3 10
Houston, TX 952.53 55 84 59 46 11.4 2647 39558 6 5 1
Indianapolis, IN 968.67 29 75 60 39 11.4 4412 31461 13 7 33
Kansas City, MO 919.73 31 81 55 35 12 3262 30783 7 4 4
Lancaster, PA 844.05 32 74 54 43 9.5 3214 30248 11 7 32
Los Angeles, Long Beach, CA 861.26 53 68 47 11 12.1 4700 36624 648 319 130
Louisville, KY-IN 989.26 35 71 57 30 9.9 4474 29621 38 37 193
Memphis, TN-AR-MS 1006.49 42 82 59 50 10.4 3497 27910 15 18 34
Miami-Hialeah, FL 861.44 67 82 60 60 11.5 4657 32808 3 1 1
Milwaukee, WI 929.15 20 69 64 30 11.1 2934 35272 33 23 125
Minneapolis-St. Paul, MN-WI 857.62 12 73 58 25 12.1 2095 35871 20 11 26
Nashville, TN 961.01 40 80 56 45 10.1 2682 28641 17 14 78
New Haven-Meriden, CT 923.23 30 72 58 46 11.3 3327 34364 4 3 8
New Orleans, LA 1113.16 54 81 62 54 9.7 3172 32704 20 17 1
New York, NY 994.65 33 77 58 42 10.7 7462 36047 41 26 108
Philadelphia, PA-NJ 1015.02 32 76 54 42 10.5 6092 33449 29 32 161
Pittsburgh, PA 991.29 29 72 56 36 10.6 3437 32934 45 59 263
Portland, OR 893.99 38 67 73 37 12 3387 33020 56 21 44
Providence, RI 938.5 29 72 56 42 10.1 3508 30094 6 4 18
Reading, PA 946.19 33 77 54 41 9.6 4843 32449 11 11 89
Richmond-Petersburg, VA 1025.5 39 78 53 44 11 3768 33510 12 9 48
Rochester, NY 874.28 25 72 60 32 11.1 4355 34896 7 4 18
St. Louis, MO-IL 953.56 32 79 57 34 9.7 5160 34546 31 15 68
San Diego, CA 839.71 55 70 61 10 12.1 3033 32586 144 66 20
San Francisco, CA 911.7 48 63 71 18 12.2 4253 47966 311 171 86
San Jose, CA 790.73 49 68 71 13 12.2 2702 41994 105 32 3
Seattle, WA 899.26 40 64 72 35 12.2 3626 37069 20 7 20
Springfield, MA 904.16 28 74 56 45 11.1 1883 29327 5 1 20

```
Syracuse, NY 950.67 24 72 61 38 11.4 4923 30114 8 5 25
Toledo, OH 972.46 26 73 59 31 10.7 3249 30497 11 7 25
Utica-Rome, NY 912.2 23 71 60 40 10.3 1671 27305 5 2 11
Washington, DC-MD-VA 967.8 37 78 52 42 12.3 5308 41888 65 28 102
Wichita, KS 823.76 32 81 54 28 12.1 3665 34812 4 2 1
Wilmington, DE-NJ-MD 1003.5 33 76 56 65 11.3 3152 33927 14 11 42
Worcester, MA 895.7 24 70 56 65 11.1 3678 29374 7 3 8
York, PA 911.82 33 76 54 62 9 9699 28985 8 8 49
Youngstown-Warren, OH 954.44 28 72 58 38 10.7 3451 28960 14 13 39
```

B Tables

Table 6: Correlation matrix of all the variables

	Mortality	JanTemp	JulyTemp	RelHum	Rain	Education
Mortality	1.00	−0.02	0.32	−0.10	0.43	−0.51
JanTemp	−0.02	1.00	0.32	0.09	0.06	0.11
JulyTemp	0.32	0.32	1.00	−0.44	0.47	−0.27
RelHum	−0.10	0.09	−0.44	1.00	−0.12	0.19
Rain	0.43	0.06	0.47	−0.12	1.00	−0.47
Education	−0.51	0.11	−0.27	0.19	−0.47	1.00
PopDensity	0.25	−0.08	−0.01	−0.15	0.08	−0.24
income	−0.28	0.20	−0.19	0.13	−0.36	0.51
log(HCPot)	0.13	0.23	−0.41	0.18	−0.48	0.18
log(NOxPot)	0.28	0.18	−0.30	0.10	−0.39	0.03
log(S02Pot)	0.39	−0.31	−0.31	−0.11	−0.13	−0.25

	PopDensity	income	log(HCPot)	log(NOxPot)	log(S02Pot)
Mortality	0.25	−0.28	0.13	0.28	0.39
JanTemp	−0.08	0.20	0.23	0.18	−0.31
JulyTemp	−0.01	−0.19	−0.41	−0.30	−0.31
RelHum	−0.15	0.13	0.18	0.10	−0.11
Rain	0.08	−0.36	−0.48	−0.39	−0.13
Education	−0.24	0.51	0.18	0.03	−0.25
PopDensity	1.00	−0.00	0.26	0.34	0.45
income	−0.00	1.00	0.29	0.25	−0.10
log(HCPot)	0.26	0.29	1.00	0.94	0.61
log(NOxPot)	0.34	0.25	0.94	1.00	0.70
log(S02Pot)	0.45	−0.10	0.61	0.70	1.00

Table 7: Results of omitting influential points in the reduced model

	All obs.	−31	−28	−57	−49
(Intercept)	686.2891	677.4772	691.7299	695.5930	642.8550
t-values	(4.337)	(4.274)	(4.406)	(4.460)	(4.005)
JanTemp	−1.3815	−1.0456	−1.1686	−1.4637	−1.5789
	(−2.151)	(−1.438)	(−1.782)	(−2.307)	(−2.411)
JulyTemp	4.2538	4.1070	4.1618	3.5717	5.1023
	(2.736)	(2.629)	(2.623)	(2.249)	(3.053)
Rain	2.2858	2.3990	2.1545	2.7692	2.1736
	(3.5111)	(3.628)	(3.300)	(3.916)	(3.333)
Education	−15.9749	−15.1461	−15.6338	−13.4974	−17.1244
	(−1.996)	(−1.882)	(−1.969)	(−1.681)	(−2.142)
log(NOxPot)	30.8687	29.0855	32.6266	30.7757	32.0772
	(5.435)	(4.878)	(5.646)	(5.501)	(5.614)
Adj. R^2:	**0.522**	**0.516**	**0.524**	**0.539**	**0.533**

Table 8: Results of omitting outliers in the Y-direction in the reduced model

	All obs.	−27	−36	−58	−27;−36;−58
(Intercept)	686.2891	789.9047	677.9276	764.0458	868.2495
t-values	(4.337)	(5.255)	(4.524)	(5.125)	(6.705)
JanTemp	−1.3815	−1.2634	−1.7660	−1.3089	−1.4937
	(−2.151)	(−2.119)	(−2.826)	(−2.193)	(−2.874)
JulyTemp	4.2538	3.8522	4.1258	3.7556	3.2173
	(2.736)	(2.664)	(2.801)	(2.586)	(2.613)
Rain	2.2858	2.1540	2.1754	2.6163	2.3962
	(3.5111)	(3.562)	(3.521)	(4.261)	(4.605)
Education	−15.9749	−22.0473	−12.7914	−20.9156	−24.8549
	(−1.996)	(−2.877)	(−1.668)	(−2.752)	(−3.670)
log(NOxPot)	30.8687	29.2265	30.1119	31.1144	28.7801
	(5.435)	(4.878)	(5.591)	(5.899)	(6.430)
Adj. R^2:	**0.522**	**0.579**	**0.514**	**0.593**	**0.662**

C Graphics

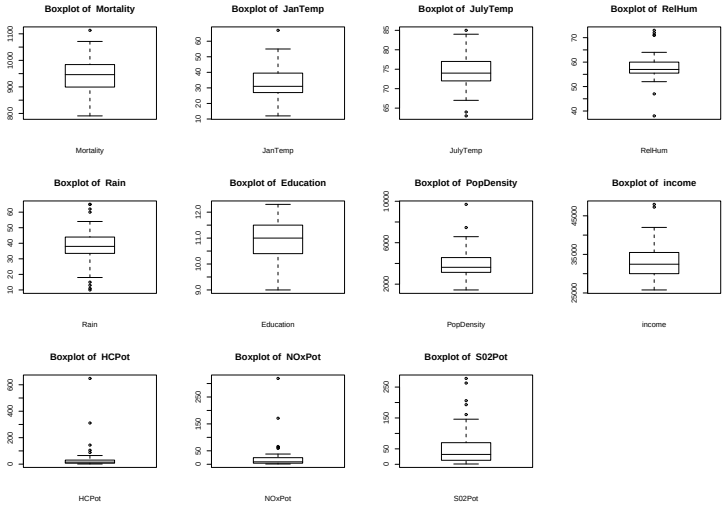

Figure 4: Boxplots of all the variables

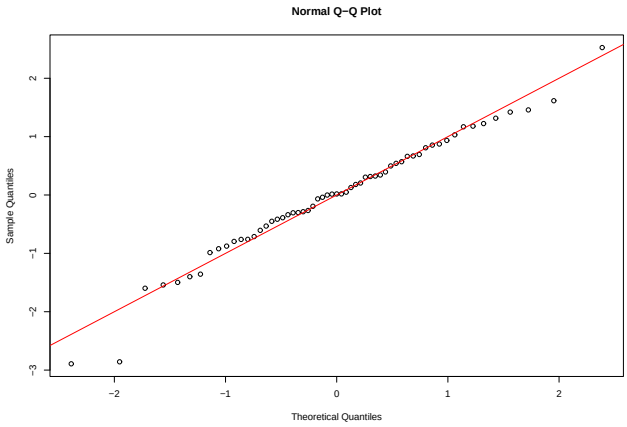

Figure 5: Normal probability plot of reduced model's residuals

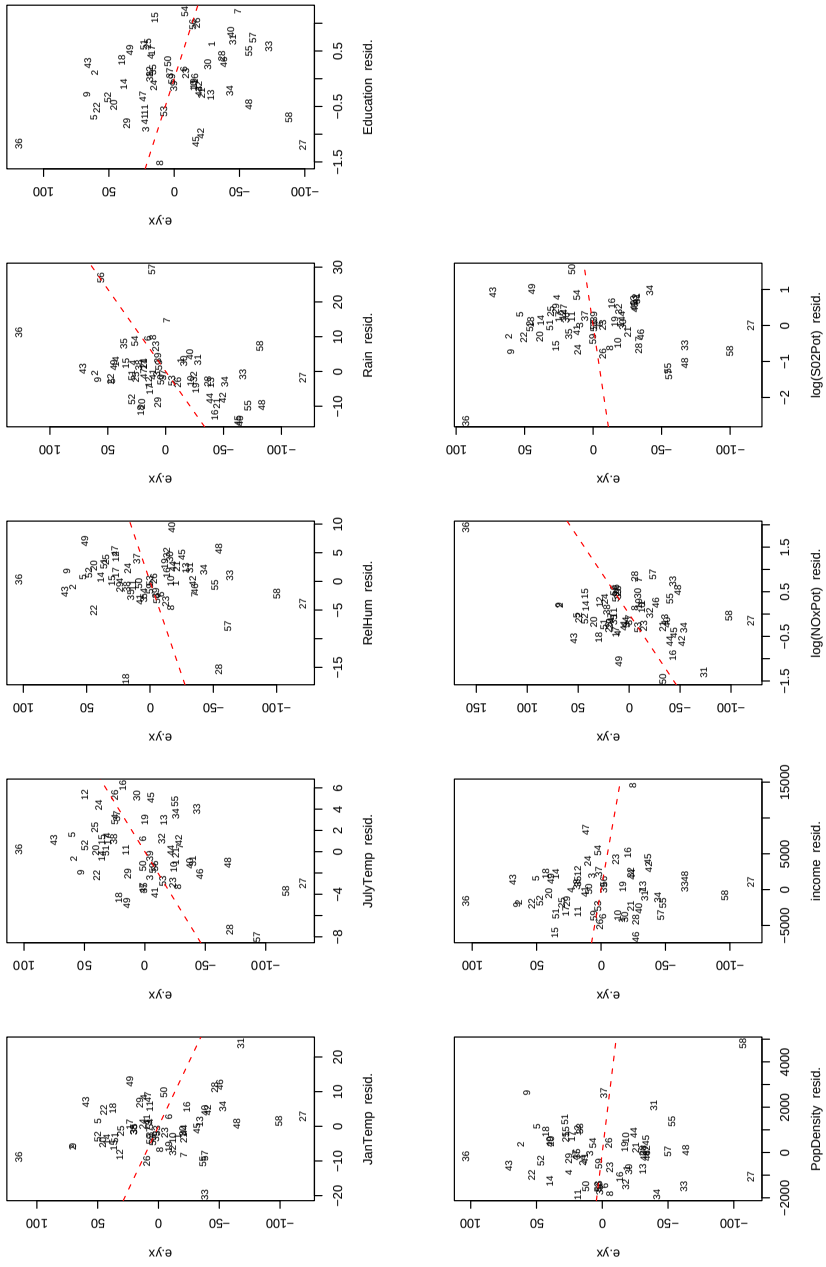

Figure 6: Added variable plots before the backward elimination procedure

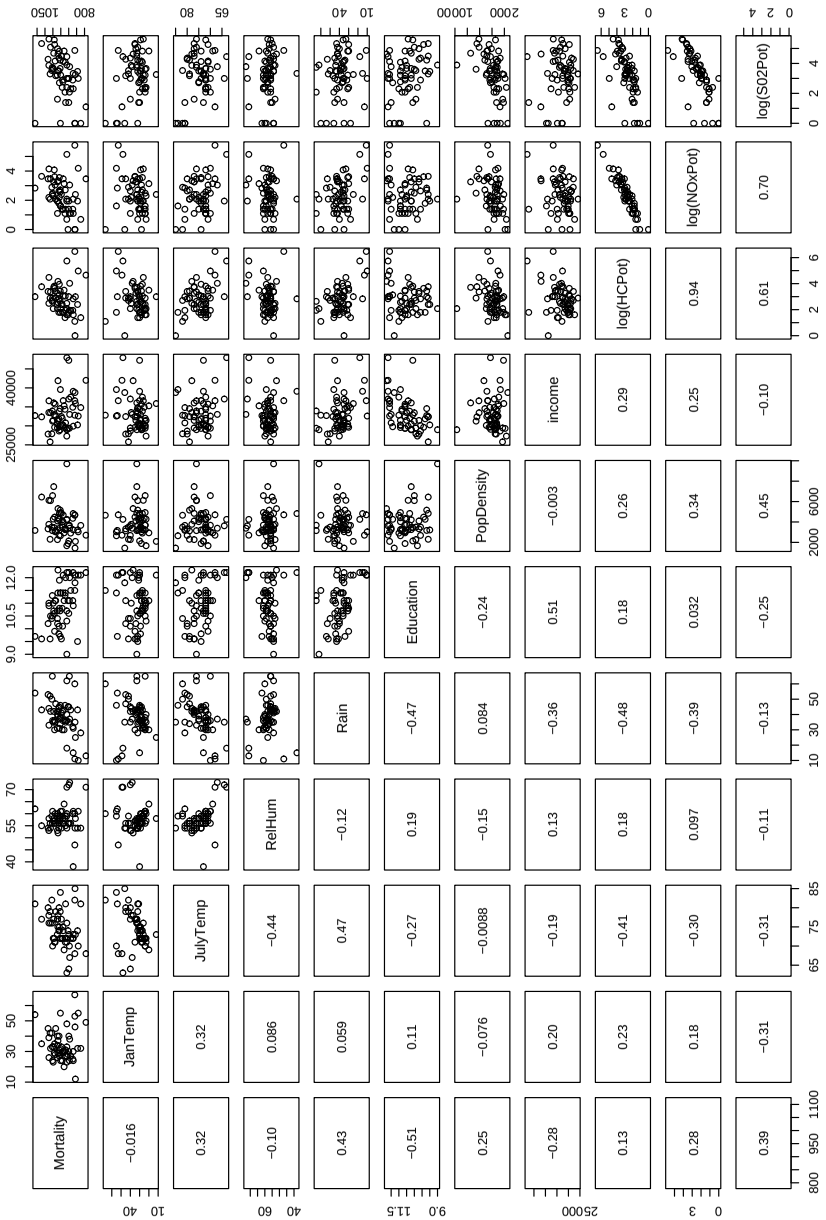

Figure 7: Pairwise scatter plots of all variables

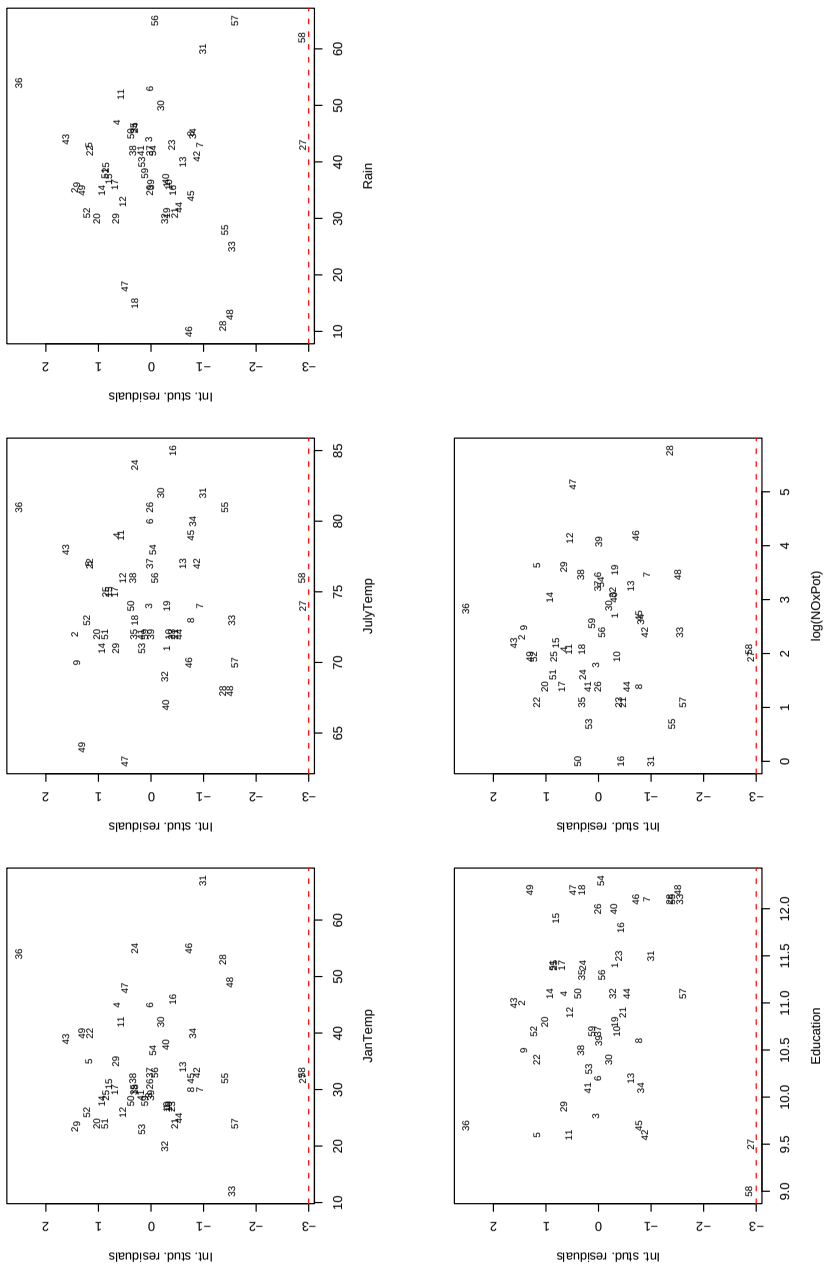

Figure 8: Scatter plots of the standardized residuals vs. each of the predictor

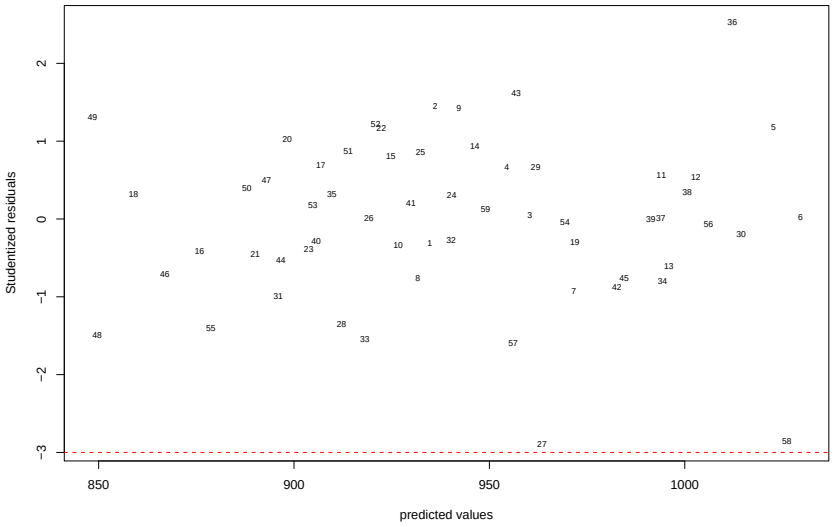

Figure 9: Scatter plot of the standardized residuals vs. the fitted values

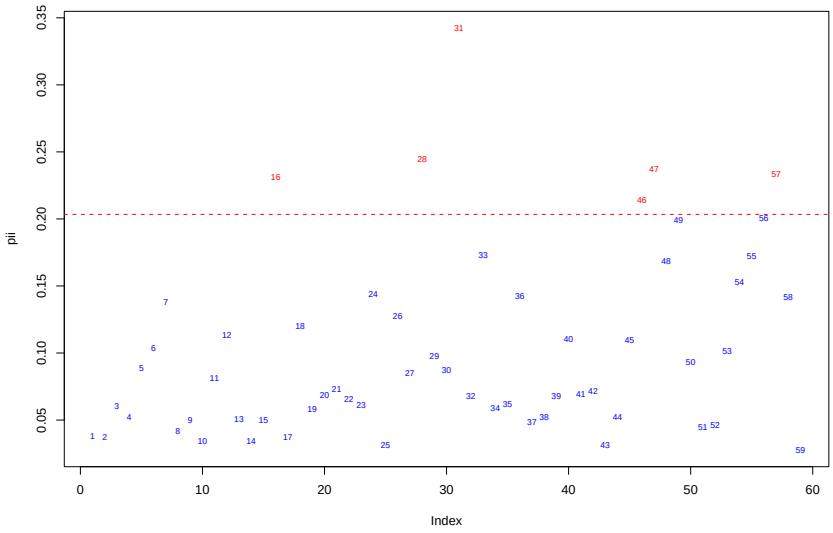

Figure 10: Index plot of the the leverage values p_{ii} (reduced model)

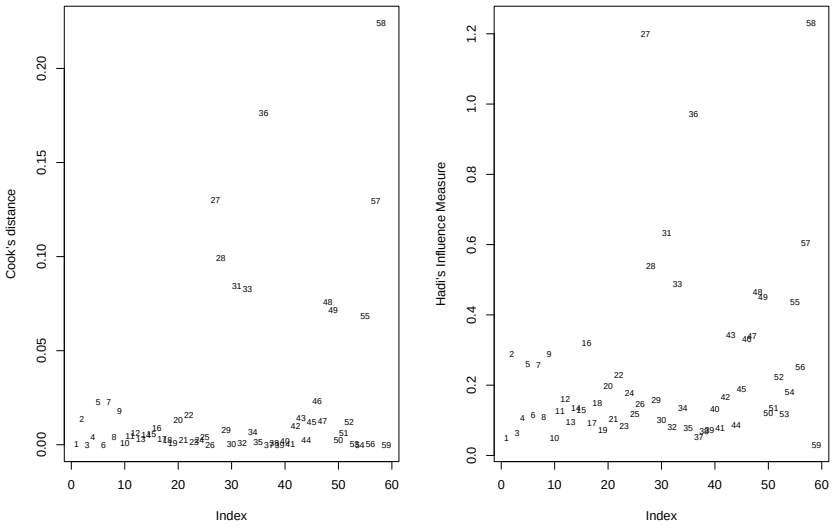

Figure 11: Index plot of the Cook's distance and the Hadi's Influence Measure (reduced model)

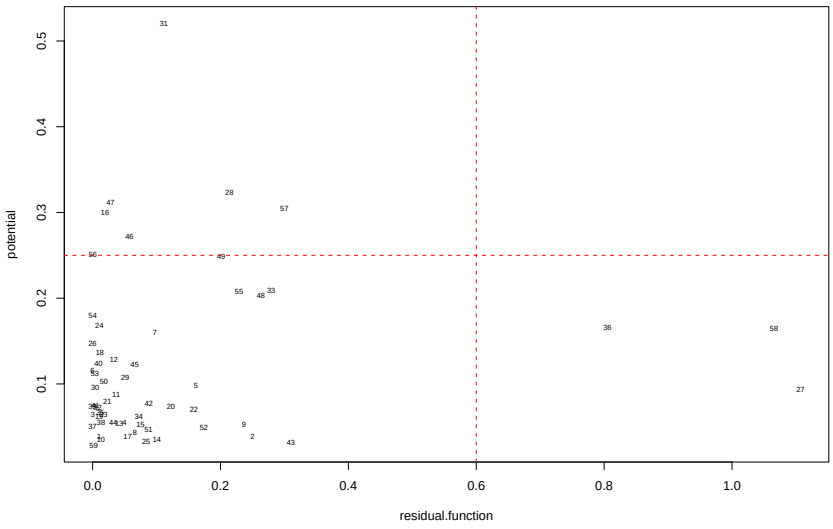

Figure 12: Potential-Residual plot (reduced model)

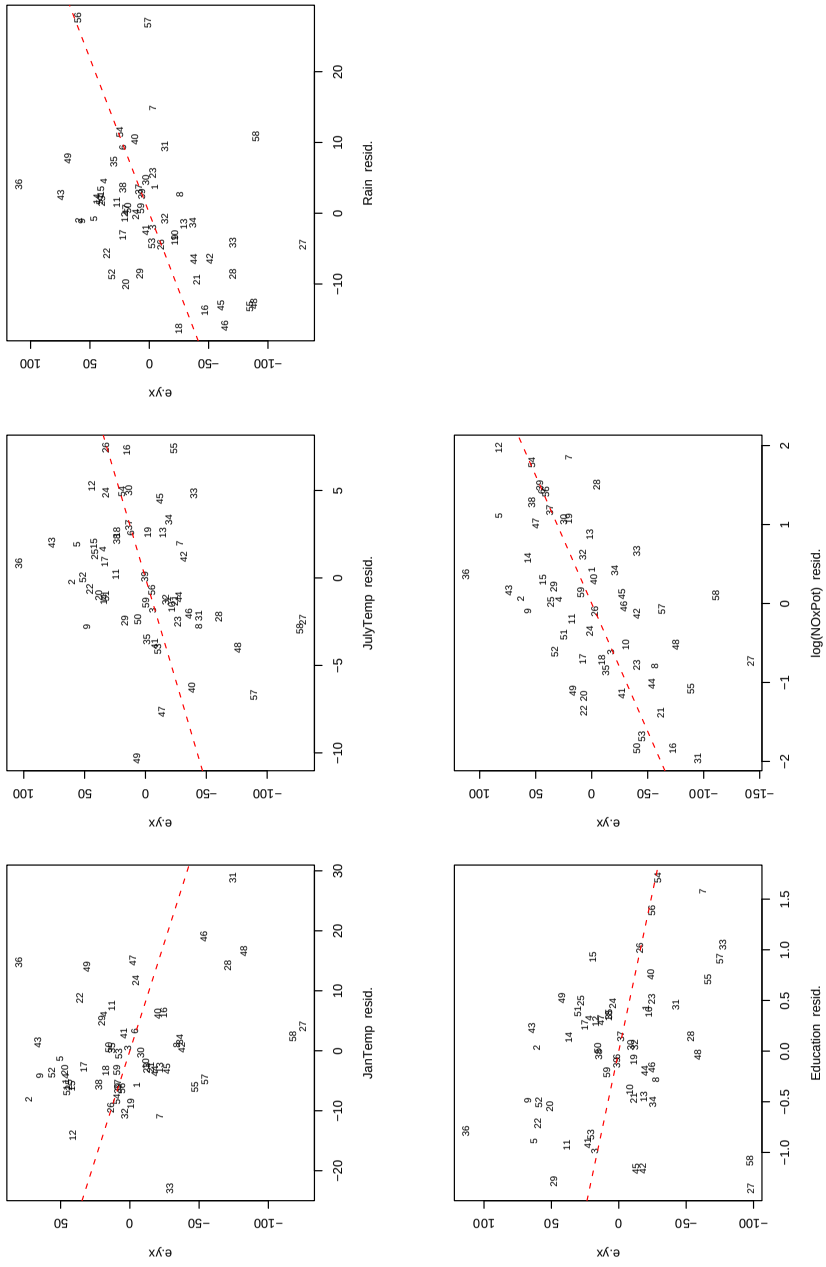

Figure 13: Added variable plot for the reduced model

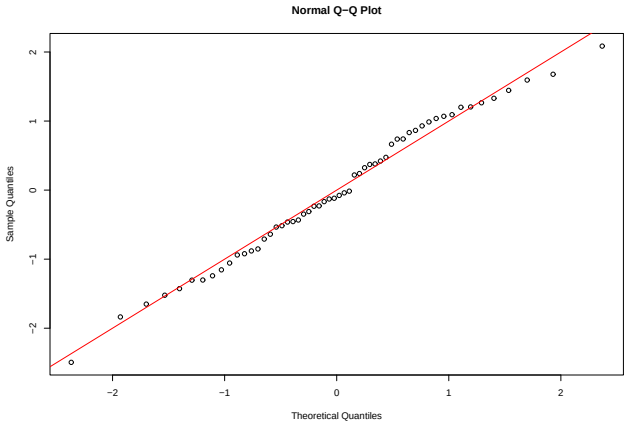

Figure 14: Normal probability plot for the final model

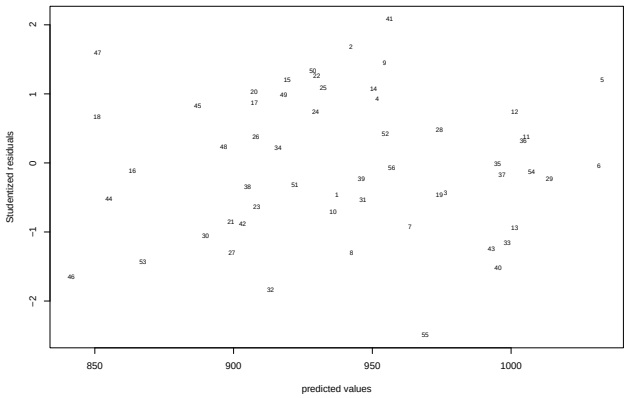

Figure 15: Scatter plot of the standardized residuals vs. the fitted values (final model)

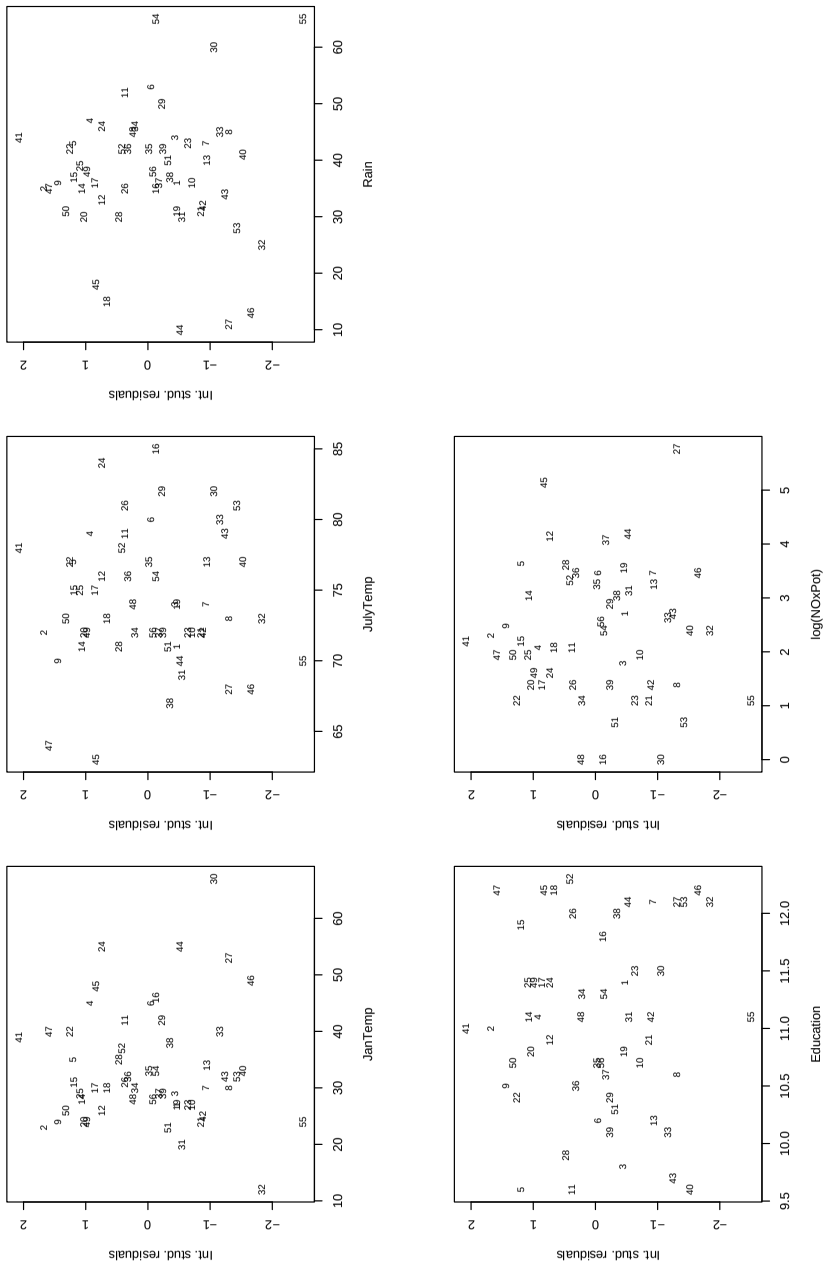

Figure 16: Scatter plots of the standardized residuals vs. each predictor (final model)