KU LEUVEN

FACULTY OF SCIENCE

2017-18 GLM Course KULeuven

Generalized Linear Models

Laura GUERRERO VELASQUEZ
laurajuliana.guerrerovelasquez@student.kuleuven.be
Lyse Naomi WAMBA MOMO
lysenaomi.wambamomo@student.kuleuven.be
Daniel Gerardo GIL SANCHEZ
daniel.gilsanchez@student.kuleuven.be
Gabriel BÉNÉDICT
gabriel.benedict@student.kuleuven.be

Prof. Emmanuel Lesaffre

Academic year 2017-2018

Part I
Medical Care for the Elderly

Fit a Poisson regression model in a frequentist manner by modeling the total number of visits, i.e. nvisit, on the selected covariates: numchron, adldiff, age, gender, married, faminc, employed, privins, and medicaid.

A dataset containing 500 patients older than 65 is analyzed to draw inference on the effect of certain patients' characteristics on the number of medical visits (`nvisit`) during the years 1987 and 1988. The following covariates are available: the number of chronic conditions (`numchron`), namely cancer, heart attack, gall bladder problems, emphysema, arthritis, diabetes, other heart disease. Whether the patient has a condition that limits daily activities (`adlidiff`), age, `gender`, marital status (`married`), family income in 10k$ (`faminc`), employment status (`employed`), private insurance coverage (`privins`) and Medicaid coverage (`medicaid`). The data is an excerpt of the National Medical Expenditure Survey (NMES) (Deb & K. Trivedi, 1997). In the following, a descriptive analysis is performed to identify patterns in the data. Then, different regression models for count data are conducted to investigate the association between the total number of visits and the patients' characteristics. This analysis is conducted in the statistical software `R`.

Regarding univariate descriptive statistics of continuous variables, in the `number of visits` there are subjects that never visited the hospital or the emergency room whereas there is a person that visited the hospital 154 times. The average `number of visits` is 9.47 and the standard deviation is almost 14. This difference between the unconditional mean and variance is the first symptom of overdispersion in the data. The average `number of chronic diseases` in this sample is 1.60, there are people without any disease and some others have up to 8 chronic diseases. The individuals age ranges from 66 to 109 years old, where the mean is about 74 years. The average `family income` is 2.43 and the standard deviation is 2.61. Figure 1 shows a boxplot for each variable. It is important to mention that there is a subject which family income is negative; since there is no reason to justify the deletion this value, the observation is analyzed as if it were a correct income.

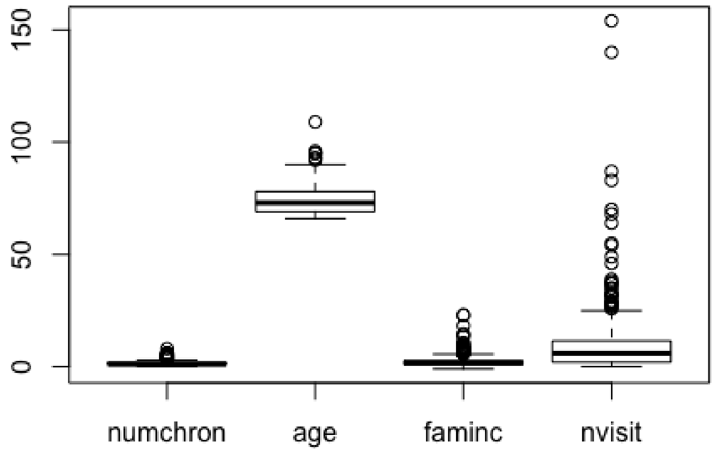


Figure 1: Overview of Continuous Variables

Table 1 shows univariate descriptive statistics for categorical variables considered in the analysis. The proportion of subjects who have `conditions that limits daily activities` is 20.4%. Almost 60% of the sample are `females`. Approximately 57% of people are `married`. Most of them (88.4%) are `unemployed`. 73% of the sample is covered by a `private health insurance`, whereas only 10% is covered by `Medicaid`.

	adldiff	male (`gender`)	married	employed	privins	medicaid
No	398	294	216	442	135	450
Yes	102	206	284	58	365	50

Table 1: Absolute frequencies of Categorical Variables

From a bivariate perspective, the correlation coefficient between the response variable, `nvisits`, and the continuous covariates is not larger than 0.2 in any case (See Figure 2). There are low negative correlations between the `number of chronic conditions` and `family income` (-0.05), age and `family income` (-0.11). With respect to the relationship between the response and categorical variables, a comparison in the mean of the `number of visits` within each variable categories leads to concluded that there are no large differences.

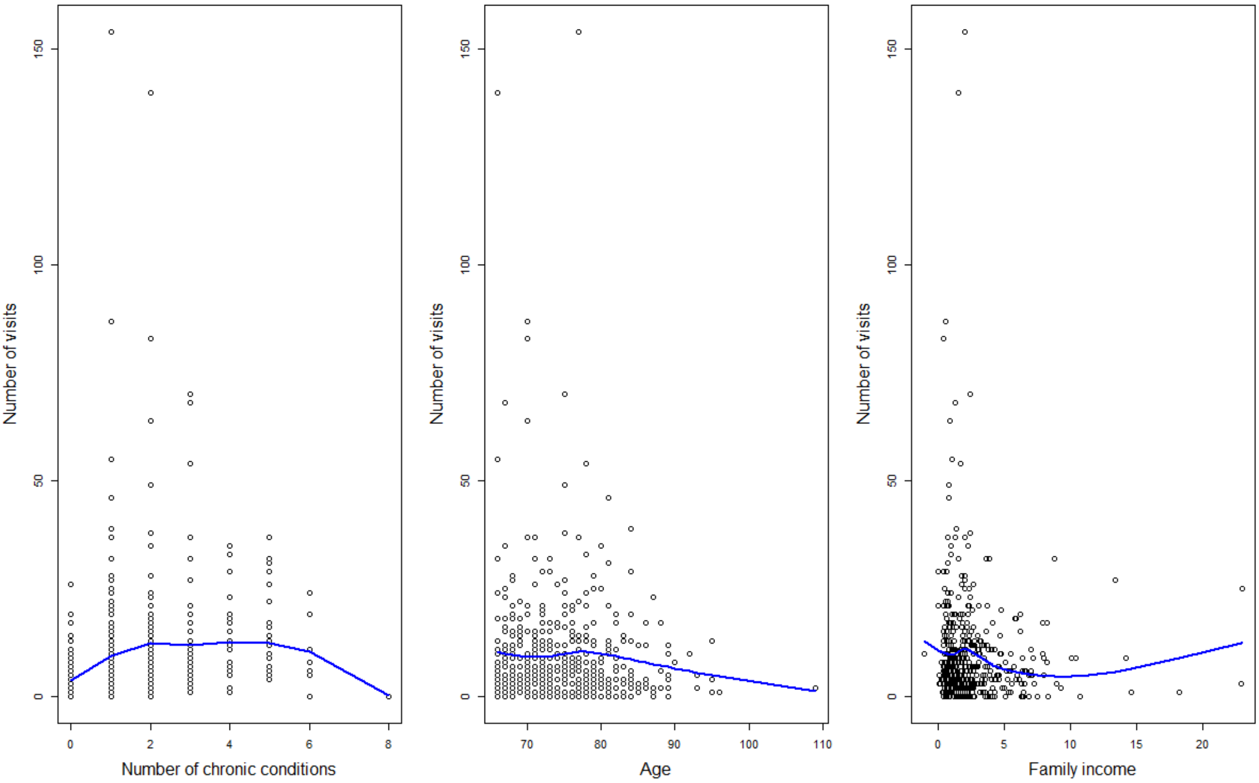


Figure 2: Scatter plots

As it is well known, all models can be summarized in a unified framework that consists of a systematic and distribution part. The systematic part describes the mean structure and the distribution part describes the individual variation of the response around the mean. The success of the model depends on the selection of each part. The distributional part is the primary concern because the inference and conclusions that can be done after having the results, rely on the correctness of this part of the model.

Consequently, there are different kinds of models that can be fitted when the response variable is a count, as it is in this case. The criteria to correctly select the distribution part of the model depends on the distribution of the `number of visits` and whether there is overdispersion in the data or not. Hence, the first step is to assume equidispersion, that is, the conditional mean and the conditional variance are the same, which is an assumption of the Poisson distribution. Therefore, a **Poisson regression model** is performed in a **frequentist manner** using the covariates mentioned.

Considering this model, if the difference between the conditional mean and the conditional variance is large, another parameter needs to be considered with the purpose of modelling the overdispersion. Two different models can be used to fix the overdispersion problem, a Negative Binomial Regression model or a Quassi-

Poisson Regression model.

It is important to mention that in several cases the distribution of the response variable may be affected by the amount of zeros that can arise from the collection of the data. If the frequency of zeros is large in comparison to the distribution of the non-zero values in the number of visits, an extension of the previous models is needed. Two models can be considered, namely the Zero-Inflated Poisson Regression model and the Zero-Inflated Negative Binomial Regression model. The advantage of these distributions is that they allow to fit a model that correctly mirrors the large frequency of zeros.

Table 2 shows the structural form for each model where p is the number of variables, $\mathbf{x}$ is a matrix with $p+1$ columns, containing the intercept and the covariates and β is the corresponding coefficient parameter vector.

Distribution	**Expression**
Poisson	$\log(E(y \mid \mathbf{x})) = \beta'\mathbf{x}$, with $y \sim Poisson(\lambda)$ and $\lambda = \mu = \sigma^2$
Quasi-poisson	$\log(E(y \mid \mathbf{x})) = \beta'\mathbf{x}$, with $y \sim Poisson(\lambda)$ Quasi-likelihood, so $\lambda = \mu \neq \sigma^2 = \phi\lambda$
Zero Inflated Poisson	$\begin{cases} \pi = \exp(\beta'\mathbf{x})/(1+\exp(\beta'\mathbf{x})), \text{ with } \pi \text{ the probability of zeros} \\ \log(E(y \mid \mathbf{x}, y>0)) = \beta'\mathbf{x}, \text{ with } y \sim \text{Zero-Inflated Poisson}(\lambda,\pi) \text{ and } \lambda = \mu = \sigma^2 \end{cases}$
Negative binomial	$\log(E(y \mid \mathbf{x})) = \beta'\mathbf{x}$, with $y \sim NB(p,r)$, $\mu = \frac{pr}{1-p}$ and $\sigma^2 = \frac{pr}{(1-p)^2}$
Zero inflated negative binomial	$\begin{cases} \pi = \exp(\beta'\mathbf{x})/(1+\exp(\beta'\mathbf{x})), \text{ with } \pi \text{ the probability of zeros} \\ \log(E(y \mid \mathbf{x}, y>0)) = \beta'\mathbf{x}, \text{ with } y \sim \text{Zero-Inflated } NB(p,r,\pi), \mu = \frac{pr}{1-p}, \\ \qquad \text{and } \sigma^2 = \frac{pr}{(1-p)^2} \end{cases}$

Table 2: Considered Distributions

Following this notation, the first model considered, the **Poisson regression model**, can be written as:

$$\begin{aligned} log(E(nvisits)) = \beta_0 + \beta_1 numchron + \beta_2 adldiff + \beta_3 age + \beta_4 gender + \beta_5 married + \\ \beta_6 faminc + \beta_7 employed + \beta_8 privins + \beta_9 medicaid \end{aligned} \tag{1}$$

Such a model results in all variables being significant considering a significance level of 5% ($\alpha = 0.05$), the residual deviance being 5166.9 and 490 degrees of freedom. A rule of thumb to identify whether there is overdispersion in the model is that the residual deviance and the degrees of freedom should be equal or relatively close. This is clearly not the case, in fact the estimate of the overdispersion parameter is $5166.9/490 = 10.54$, which is very far from one.

Fit a negative binomial model and a quasi-Poisson model in a frequentist manner if there is overdispersion.

Model 1 presents overdispersion so a more flexible model is needed to correctly specify the distribution part of the model. As mentioned before, a **Negative Binomial regression** model can be a direct solution but the zero-inflated models are also considered even though the frequency of zeros (52) is not disproportionally high in comparison to the frequency of appending values. In Figure 3 a histogram of the response is presented along with the fit of the distributions considered.

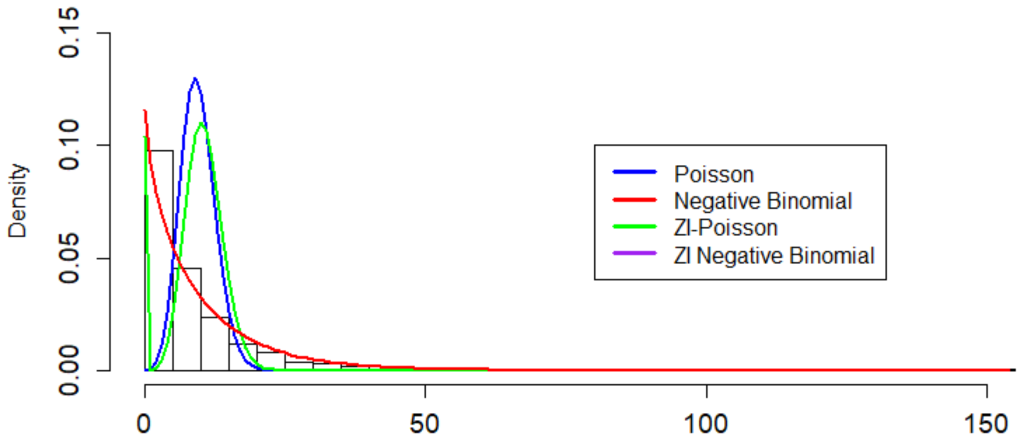


Figure 3: Histogram of number of visits. Poisson, Negative Binomial and Zero-Inflated fit

As can be seen on Figure 3, the Poisson distribution does not fit the amount of zeros and low counts correctly, overestimates the mid-range values and the tail of the distribution is not fitted. The zero-inflated Poisson, fits better the amount of zeros but again the mid-values of the response are overestimated and the tail is underestimated. In contrast, the Negative Binomial distribution fits correctly the whole distribution of the response, the amount of zeros is correctly specified, the mid-values are well estimated and the tail is fitted better than the other distributions. The zero-inflated negative binomial estimates the distribution as the Negative Binomial because the probability of having extra zeros is very close to zero. For these reasons, the **Negative Binomial distribution** is the distribution that fits the response variable best.

Note that the **Quasi-Poisson regression** model is not involved in the comparison of the previous distributions. The reason is because this method is based on quasi-likelihoods instead of proper likelihoods. The advantage though, is that it can be used when the interest of the model is to calculate the effect of covariates on the mean function with a correct variance function. As a consequence, the estimates of the coefficients are similar to the **Poisson Regression** model but the standard errors are corrected, implying that the inference about the covariates is not affected by the misspecification of the distribution of the response variable. In table 3 a comparison of the estimates between the **Poisson**, the **Quasi-Poisson** and the **Negative Binomial** is shown.

Variable	Poisson			Quasi-Poisson			Negative Bin.		
	Coef.	S.E.	p-value	Coef.	S.E.	p-value	Coef.	S.E.	p-value
Intercept	3.53	0.2	< 0.001	3.53	0.98	< 0.001	3.31	0.61	< 0.001
numchron	0.17	0.01	< 0.001	0.17	0.03	< 0.001	0.23	0.03	< 0.001
adldiff	0.17	0.04	< 0.001	0.17	0.21	0.43	0.14	0.13	0.28
age	-0.03	0	< 0.001	-0.03	0.01	0.05	-0.02	0.01	< 0.001
gender	0.15	0.03	< 0.001	0.15	0.12	0.22	0.15	0.11	0.15
married	-0.16	0.03	< 0.001	-0.16	0.12	0.17	-0.16	0.11	0.15
faminc	-0.02	0.01	< 0.001	-0.02	0.03	0.42	-0.01	0.02	0.74
employed	-0.53	0.06	< 0.001	-0.53	0.22	0.02	-0.61	0.16	< 0.001
privins	0.46	0.04	< 0.001	0.46	0.14	< 0.001	0.43	0.12	< 0.001
medicaid	0.15	0.06	0.01	0.15	0.2	0.46	0.09	0.18	0.63

Table 3: Comparison of estimates

There are several things to note about this comparison. First, as it was mentioned before, the estimates of the coefficients are identical in the **Poisson** and the **Quasi-Poisson** model. Second, in the **Poisson** model all covariates are significant (significance is measured with the p-value of a Wald-test, see Appendix). Third, the estimation of robust standard errors in the **Quasi-Poisson** model impacts directly the significance of the covariates. In this model only the `number of chronic disease`, the `age`, the `employment` status and the coverage of a `private health insurance` are significant. Fourth, the **Negative Binomial** has different estimates for each coefficient in comparison with the other models, but it has the same significant variables as the **Quasi-Poisson** model. It is important to notice here, that changing the distribution part of the model, impacts the inference about the covariates. In the **Poisson** model all covariates are significant because when the distribution part is wrongly specified, the systematic part tries to compensate the lack of fit.

A second concern after fitting the response is to decide which model is best to continue the analysis. This decision is based on the AIC (see Appendix) of each model, the residual deviance and the fit of the distribution part on the data (See Table 4 and Figure 3). Of course, the Quasi-Poisson model cannot be compared in terms of AIC and fit because of its dependence to a quasi-likelihood. As it is expected, the **Negative Binomial model** has the lowest AIC, the smallest residual deviance and the better fit. So this model is further considered for variable selection.

Note that the Quasi-Poisson model would be useful if the approximation of the distribution of the response variable is unknown. In this case, since the Negative Binomial gives a good approximation, the Quasi-Poison is not further considered.

Model	AIC	Residual deviance	Residual df
Poisson	6863.71	5166.94	490
Quasi-Poisson	NA	5166.94	490
NB	3244.99	568.74	490

Table 4: Comparison of models

Do model selection, check model adequacy, and graphically check model prediction

Variable selection is performed with an automated selection procedure, namely stepwise selection, based on BIC, because the penalization in overparametrization is larger than in AIC. All main effects are considered as well as their second-order interactions. Although a previous study does not include interactions (Deb & K. Trivedi, 1997), there is no evidence against the existence of an interaction. The resulting model is:

$$\begin{aligned} log(E(nvisit)) = \beta_0 + \beta_1 numchron + \beta_2 adldiff + \beta_3 age + \beta_4 employed + \\ \beta_5 privins + \beta_6 adldiff \cdot age \end{aligned} \tag{2}$$

Note that the form of the model is the same as in the **Poisson regression** model. The dispersion parameter in the **Negative Binomial regression** does not affect the expected counts, but it affects the estimated variance. Table 5 shows the estimate of each parameter. All variables are significant except for the intercept and age, which is kept in the model to respect the principle of marginality.

Regarding interpretation, an increase of one chronic condition in a subject leads to an increase of the expected `number of visits` by a factor of $exp(0.248) = 1.28$ when the remaining covariates are held constant. The difference between being `employed` or not, impacts the expected `number of visits` by a factor of $exp(-0.523) = 0.593$, when the other covariates are fixed. The possession of `private health insurance` impacts increases the expected `number of visits` by a factor of $exp(0.377) = 1.458$, when the remain covariates are held constant. It is logical that having more `chronical diseases` increases the number of visits. Regarding `private health insurance`, someone who possesses one might be more prone to go to the hospital, knowing that he is covered. Finally, `employment` indicates regular activities and might thus decrease the number of days at the hospital.

The interaction between having a condition that `limits daily activities` and age needs special attention. The fact that the interaction is significant implies that when a subject gets older the impact in the expected `number of visits` is different if it has a condition that limits its daily activities. In a similar way, the impact of having a condition that `limits daily activities` can only be interpreted taking into account the age of the subject. For instance, when the age increases by one unit in an individual that does not have any condition that limits their activities, the expected `number of visits` increases by 1% ($exp(0.008)$). Inversely, when a person that is `limited in its daily activities` gets one year older, its expected number of visits decreases 6% ($exp(0.008 - 0.075) = 0.94$).

	Coef.	S.E.	z value	p-value
Intercept	0.921	0.707	1.304	0.192
numchron	0.248	0.033	7.489	< 0.001
adldiff	5.800	1.256	4.617	< 0.001
age	0.008	0.009	0.895	0.371
employed	-0.523	0.153	-3.419	0.001
privins	0.377	0.108	3.505	< 0.001
adldiff:age	-0.075	0.016	-4.577	< 0.001
Dispersion	1.063			

Table 5: Stepwise selected negative binomial model

In this sense, it is necessary to check the model adequacy and prediction. The likelihood ratio test when comparing to the null model gives a p-value < 0.001, which means that the overall fit is statistically significant. Regarding residual diagnostics, Pearson and Deviance residuals are considered to identify non-linear trends in the continuous covariates, Figure 15 show Deviance and Pearson residuals for each covariate (see appendix). The horizontal trend in continuous covariates and similar means in categorical covariates indicates that no additional higher order term is necessary.

To identify influential observations, the use of studentized residuals, Cook's distance and hat values is neccesary. Figure 4 shows that Cook's distance identifies observation 177 and 444 as influential, studentized residuals flags observation 95 and 444 and hat-values identifies observations 182 and 397. Therefore, model (2) is fitted again five times, without each of the influential observations identified, to check whether there is a difference in the estimations. As a result, no substantial differences are found and so the model is valid.

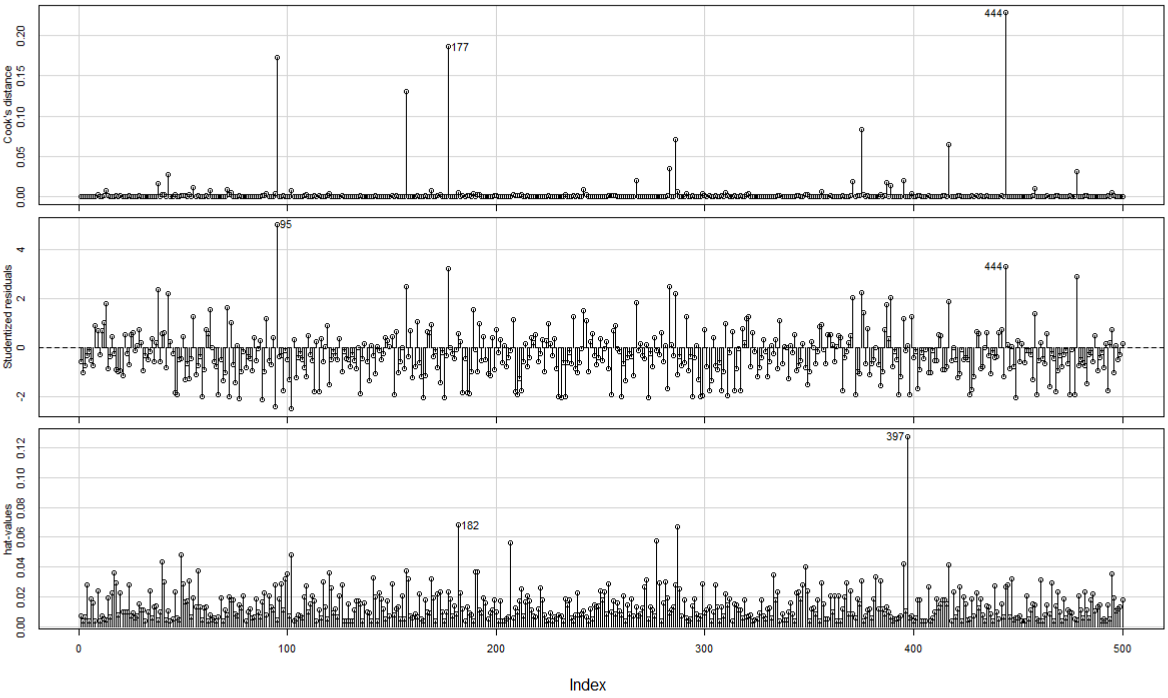


Figure 4: Influential Observations

Model prediction can be tested by plotting the predicted responses against the histogram of the observed responses (see Figure 5). Compared to the blue line in Figure 3 (negative binomial model with no covariates), this model seems to fit slightly closer to the actual data. So it can be concluded that prediction power is overall satisfactory.

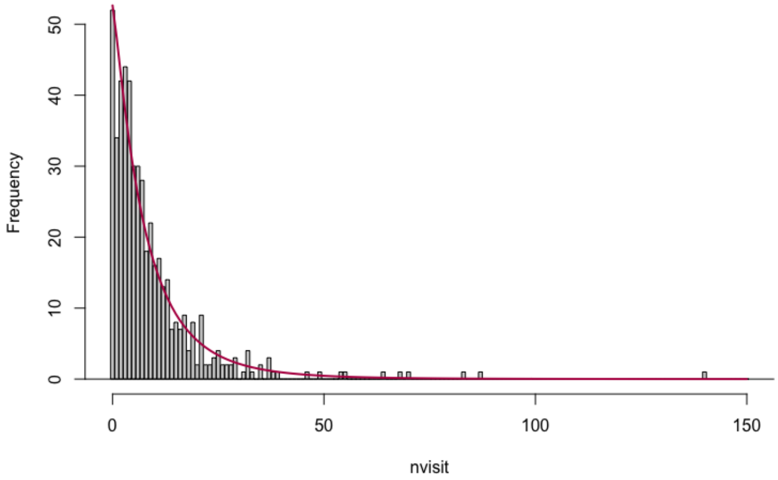


Figure 5: Negative Binomial Stepwise Selected Model – Prediction

Fit the final chosen model in Bayesian manner

Since the final model has an underlying **Negative Binomial** distribution, a **Bayesian Negative Binomial Model** with non-informative normal priors and the same covariates is modelled below. Note that a burn-in of 2.000 out of 10.000 iterations is used. Also, only one set of starting values is used, namely the coefficients of the frequentist model. The convergence is reached, as can be seen on the trace plot of the Markov Chains (see Figure 14 in the Appendix): no time trend over iterations is to be observed, variance is stable and the density is symmetric.

	Mean	**SD**	**Naive SE**	**Time-series SE**
Intercept	1.550	0.540	0.006	0.037
numchron	0.243	0.041	0.000	0.005
adldiffyes	2.806	0.736	0.008	0.042
age	0.000	0.007	0.000	0.001
employedyes	-0.532	0.151	0.002	0.013
privinsyes	0.385	0.110	0.001	0.010
adldiffyes:age	-0.036	0.010	0.000	0.001

Table 6: Bayesian Negative Binomial Model

Inference on the Bayesian model is based on the posterior distribution of each parameter. Their posterior mean is similar to the frequentist model (as expected with non-informative priors for the distributional part and frequentist estimates priors for the systematic part). Since the Bayesian estimates are similar to the frequentist model, the reader is referred to page 6 for inferential considerations, taking into account that now these estimates reflect the posterior distribution of each parameter. Naive SE (also called MC Error or standard error of the mean) and Time-series SE (correcting for autocorrelation of the chain) are systematically lower than the standard deviation of each parameter's posterior distribution (SD in table 6). This indicates that all parameters have sufficiently converged.

To sum up, the interest of investigating the association between the total `number of visits` and the covariates available lead to the following process:

- A **Poisson regression model is fitted** in a frequentist manner, resulting in overdispersion.
- To deal with overdispersion, different models are considered, namely **Quasi-Poisson, Negative Binomial, Zero-Inflated Poisson and Zero-Inflated Negative Binomial.**
- There is no theoretical reason or any improvement in the fit by considering zero-inflated distributions.
- A comparison in the estimates of the Poisson, Quasi-Poisson and Negative Binomial is performed.

- The AIC criteria suggests that the best model is the **Negative Binomial**. So this model is considered for further analysis.

- **The stepwise procedure suggests some covariates that are significant**: number of chronic conditions, whether the subject has a condition that limits daily activities, age, employment status, and private health insurance.

- **To check model adequacy**, several measures were considered: Likelihood Ratio to test the overall significance of the model. Pearson and Deviance residuals to assess the fit of the continuous covariates. Finally, Cook's distance, hat values and studentized residuals to identify influential observations.

- **To check model prediction,** a comparison between observed and fitted number of visits is shown graphically.

- **The final model is fitted in a Bayesian manner** with non-informative priors, resulting in similar conclusions.

Part II

Air quality for the New York metropolitan area

Make a descriptive analysis to explore the relation between the covariates and the outcome

Daily air quality measurements were taken from May 1, 1973 to September 30, 1973 in New York City, considering `Ozone` concentration, `Solar radiation`, `Wind` speed and `Temperature`; a description of the available variables can be found in Table 7.

Variable	Description	Min	Max	Mean(Sd)
Ozone	Mean ozone in parts per billion (ppb)	1	168	42.1(33.27)
Solar.R	Solar radiation in Langleys	7	334	184.8(91.15)
Wind	Average wind speed in miles per hour	2.3	20.70	9.94(3.55)
Temp	Maximum daily temperature in degrees Fahrenheit	57	97	77.79(9.53)
Month	Month of the year	5	9	-
Day	Day of month	1	31	-

Table 7: Variables description

The following analysis is made in the statistical software `R`. The variable `Solar radiation` is the one who presents more dispersion, whereas `Wind` speed is the most stable. Overall, the measurements were taken in warm days with an average `Temperature` of 77.8 degrees Fahrenheit.

Due to its harmful impact on lung function, the primary interest of this paperwork is to model `Ozone` concentrations. To begin with, Figure 6 shows how skewed the distribution of Ozone is, given the presence of atypical high measurements. In particular, on August 25th the mean Ozone was 168 ppb and on the 1st of July it was 135 ppb, while the minimum value in all measurements is one ppb.

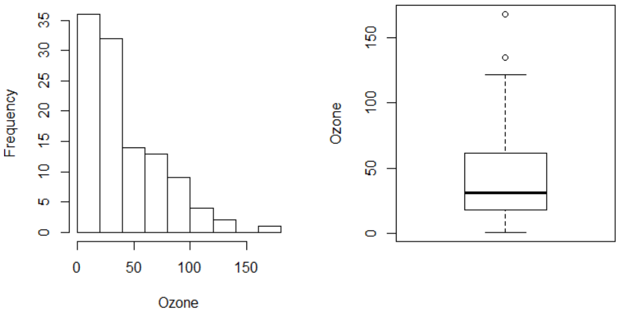
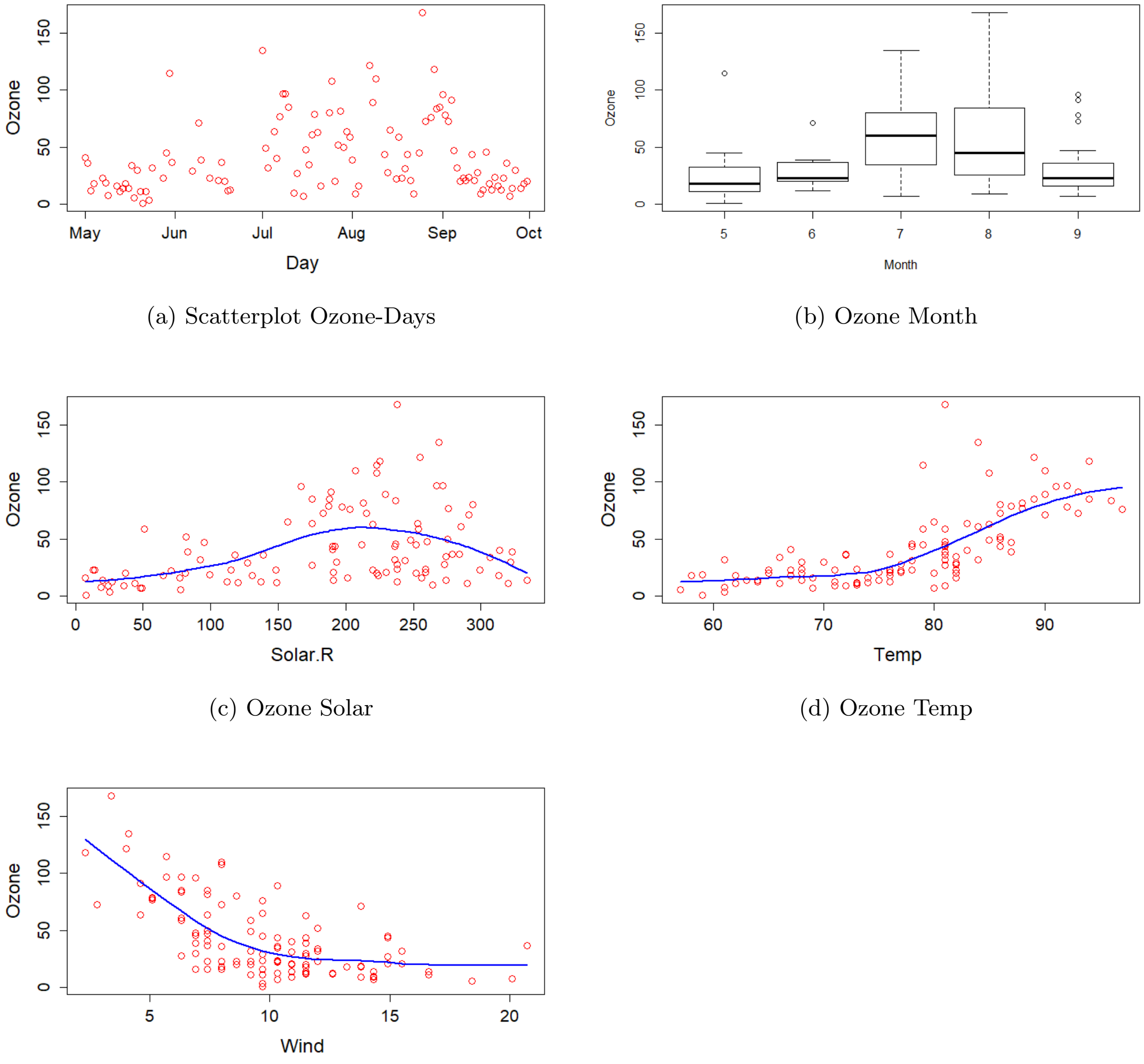


Figure 6: Ozone distribution

In regard to the association between the variables, the outcome `Ozone` is highly correlated to `Temperature` (0.698) and `Wind` (-0.612), but it is clear from Figures 7d and 7e that this association is not strictly linear. Here, LOESS (locally weighted smoothing) is used, to create a smooth line through the scatterplot to see the possible relationship between the variables.



(a) Scatterplot Ozone-Days

(b) Ozone Month

(c) Ozone Solar

(d) Ozone Temp

(e) Ozone Wind

Figure 7: Exploration relation between the covariates and the outcome

	Ozone	**Solar.R**	**Wind**	**Temp**
Ozone	1	0.348	-0.612	0.698
Solar.R	0.348	1	-0.127	0.294
Wind	-0.612	-0.127	1	-0.497
Temp	0.698	0.294	-0.497	1

Table 8: Correlation matrix

Month	**Mean**	**Std.Dev**
May	24.12	22.88
June	29.44	18.20
July	59.11	31.63
August	60	41.76
September	31.44	24.14

Table 9: Ozone Descriptive statistics by Month

Accordingly, days with high temperatures present also high Ozone levels, while windy days show lower Ozone concentrations. In comparison, the association between Ozone and Solar radiation is positive but weak, a expected behaviour when looking at their scatterplot where a quadratic behavior is present.

As for the time line measurements were taken, warmer days and months (July and August) tend to have higher Ozone concentrations, which is expected given the observed relationship with Temperature before.

Due to the fact that July and August have average `Ozone` levels twice larger than the other months, the outlying observations mentioned above, have now a normal behaviour within these months.

It is important to say, that in June there are twenty-one days without measurements, more specifically the first and last week of June and some other days in the middle. This may be the reason why its mean `Ozone` concentration is so low (see Table 9), because within each month the highest concentrations are measured at the beginning and end of the month.

Model the ozone concentrations as a function of temperature only. Consider different types of splines and select the best alternative.

From Figure 7d it is obvious that there is a non linear relationship between the `Ozone` concentration and the `temperature`, thus a simple linear model is not useful. In this respect, a smoothing function (spline) will be used to model the relationship among the variables. Both penalized cubic splines and penalized B-splines (P-splines) of degree 3 will be used and compared. This penalization ensures that the number of knots and their position are controlled in an optimal setting leading to a smooth fit. 10 equally-spaced knots are considered throughout. The model fitted here is defined as,

$$Ozone_i = f(Temp_i) + \epsilon_i \tag{3}$$

where $Ozone_i$ is the response variable on the $i-th$ day, $Temp_i$ is the covariate, $\epsilon_i \sim N(0,\sigma^2)$ and f is a smooth function defined as:

$$f(x) = \sum_{j=1}^{d} \gamma_j B_j(x)$$

is a complete basis represented as a linear combination of $d = m+l-1$ basis functions for penalized B-splines, where m is the number of knots and l is the degree (3 in this case). In general, a B-Spline of order l is obtained recursively by

$$B_j^l(x) = \frac{x - \kappa_{jl}}{\kappa_j - \kappa_{jl}} B_{j-1}^{l-1} + \frac{\kappa_{j+1} - x}{\kappa_{j+1} - \kappa_{j+1-l}} B_j^{l-1}(x)$$

These B-splines consists of 4 polynomial pieces of degree 3 joined in a 2 times continuously differentiable way. In the case of penalized cubic splines, the function f is made up of polynomials splines of degree 3 (order = 4) with knots $\kappa_1 < \kappa_2 < ... < \kappa_m$, twice continuously differentiable and whose third derivative is a step function that jumps at the knots. Due to the penalization, the problem of estimating a generalized additive model reduces to estimate the smoothing parameters which when an optimal one is found, a good balance between model fit and model smoothness is obtained. Appropriate measures for comparing these two splines are the Aikaike's Information Criterion (AIC), Bayesian Information Criterion (BIC) and General Cross Validation function (GCV, see Appendix).

Spline specification	GCV	AIC	BIC
P-Spline	517.91	1010.54	1025.71
Cubic	518.48	1010.66	1025.71

Table 10: Comparison GAM models with response Ozone as a function of Temperature

From Table 10, it can be observed that the best choice of spline is the penalized B-spline with lower GCV and AIC scores. Usually one seeks the lowest values when using these criteria. The GCV criterion for example, returns the smoothing parameter λ that minimizes the expected prediction error. Figure 8 shows the fit of model (3) with P-splines.

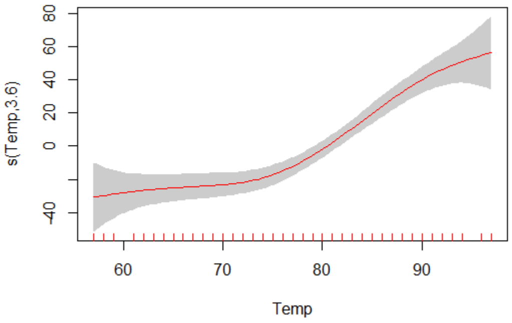


Figure 8: fit of Ozone against Temperature using P-splines

The smooth suggests that Ozone concentrations increase at low rate as the temperature reaches the 80 degrees Fahrenheit, but after this value the rate of increase grows drastically as the temperature is higher.

Include the other covariates in the model and determine what variables show an impact on ozone concentrations.

Now, other covariates are used to model Ozone concentrations. The model is now defined as a semiparametric model, by the combination of an additive (non parametric) and a parametric part:

$$Ozone_i = f(Temp_i) + f(Solar.R_i) + f(Wind_i) + f(Day_i) + \beta_0 + \beta_1 Month + \epsilon_i \tag{4}$$

The smooth functions f are defined as before, and represent the additive part of the model, where Temperature, Solar radiation, Days and Wind speed are assumed to have a nonlinear relationship with the response. The parameters β_0 and β_1 correspond to the partial linear (parametric) part of the model, where the variable *Month* is used as a categorical predictor.

Different smoothing functions are investigated to model the relationship between the response and the covariates. In order to make a standardized comparison, the number of knots is set to 10 in all cases, given variables Temperature and Wind have fewer unique values, and are set as equidistant knots. Consequently, the comparison is focused in the type of spline used.
First, each of the covariates is modeled independently with the outcome to investigate the most appropriate type of spline, as in the previous item where Ozone was modeled with Temperature as single predictor. Table 11 shows the comparison of splines in terms of GCV, AIC and BIC.

		Cubic			**P-Spline**	
Variable	**GCV**	**AIC**	**BIC**	**GCV**	**AIC**	**BIC**
Solar.R	892.11	1070.94	1084.59	891.28	1070.84	1084.23
Wind	567.01	1020.63	1034.15	566.14	1020.47	1033.71
Day	998.35	1082.95	1107.69	999.37	1083.57	1096.11

Table 11: Comparison splines

In brief, the most appropriate splines are selected as P-splines for Solar.R and Wind, and as Cubic for Days.

Thus, three models are fitted: first, including all variables with cubic spline (Full Cubic); second, including all variables with P-spline (Full P-Spline) and third, using the previous spline selection for each variables (Combined model). The result of the comparison for these models is shown in Table 12:

Model	GCV	AIC	BIC
Full Cubic	268.99	936.80	1014.83
Full P-Spline	311.60	948.17	1016.31
Combined Model	291.12	940.45	1009.42

Table 12: Comparison models

The full cubic spline model turns out to be the best model in terms of smoothness (GCV) and AIC, while the Combined Model does best in terms of BIC. The reason could be, that modelling all variables simultaneously could lead to different associations between a covariate and the `Ozone` concentration, when correcting for the values of the other variables due to multicolinearity. So the chosen model is Full Cubic. The results for the linear part of the model are shown.

Term	Estimation	P-value
Intercept	49.39(4.69)	< 0.0001
June	-11.79(7.68)	0.1286
July	-7.87(6.67)	0.2415
August	-3.91(6.91)	0.5726
September	-14.10(5.70)	0.0155

Table 13: Full Cubic Spline Estimation Linear Terms

Estimates for the `Months` follow the trend observed in the descriptive part. Higher concentrations are predicted in August and July, followed by June, September and May.

Table 14 shows the smoothing functions, where it is concluded that all the smooth terms are highly significant to explain `Ozone` concentrations.

Term	edf	F	P-value
s(Temp)	3.19	10.04	< 0.0001
s(Solar.R)	3.04	3.47	0.0126
s(Wind)	8.16	8.87	< 0.0001
s(Day)	8.39	3.99	0.0003

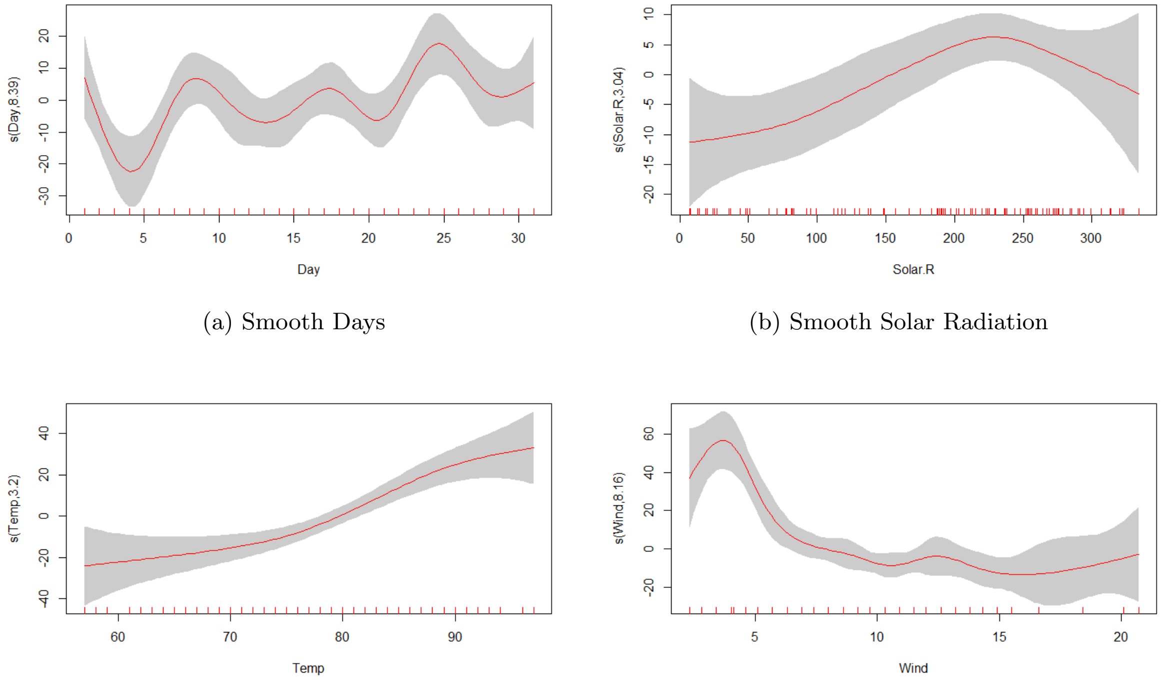
Table 14: Full Cubic Spline Estimation Smooth Terms

The estimated smooths functions are shown in Figure 9, where the values of the smoothed predictors are plotted against the partial residuals, i.e., the residuals after removing the effect of all other covariates. Among all terms, `Days` is the one with the wiggliest behaviour. Here, `Ozone` levels are high at the start of the month, then decrease to its lowest value, around the end of the first week, and increase again having the biggest value around day 25 of the month.

In terms of `solar radiation`, `Ozone` concentration is constantly increasing until solar radiation is about 225 Langleys, after this point `Ozone` concentrations collapses.

The `temperature` smooth suggests that `Ozone` concentrations increase as `temperature` is higher, but this behaviour is steeper when `temperature` is above 80 degrees Fahrenheit.

Finally, `Wind` speed suggests largest concentrations are found when the wind is at less than five miles per hour, being the peak at three miles per hour. The windier it is, the less `Ozone` concentrations are found until ten miles per hour, beyond this speed, `Ozone` levels stabilize at the average concentration.



(a) Smooth Days

(b) Smooth Solar Radiation

(c) Smooth Temperature

(d) Smooth Wind speed

Figure 9: Smooth Splines Full Cubic model including all covariates

To sum up, all variables impact `Ozone` levels but the magnitude of the effect depends on the scale of each covariate.

Fit the previous model but now taking as response the logarithm of ozone concentration. Is this model better? You may base your decision on model diagnosis.

Due to the observed skewness in the response variable, it seems reasonable to considered a logarithmic transformation to improve the fit of the model. Here the following model is considered,

$$log(Ozone_i) = \beta_0 + \beta_1 Month + f(Temp_i) + f(Solar.R_i) + f(Wind_i) + f(Day_i) + \epsilon_i \qquad (5)$$

It can be noticed that taking the logarithm of the response variable, `Ozone`, makes its distribution less skewed and now appropriately follows a normal distribution as can be seen on Figure 10 below;

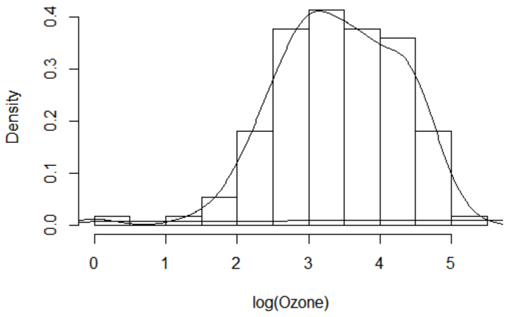


Figure 10: Distribution of log(Ozone)

Cubic splines will be used for modelling just like in the previous question. The estimated coefficients are given as follows;

Term	Estimation	P-value
Intercept	3.504(0.139)	< 0.001
June	-0.172(0.227)	0.450
July	-0.108(0.200)	0.591
August	0.007(0.204)	0.972
September	-0.193(0.169)	0.255

Table 15: Cubic Spline Estimation Linear Terms for (Standard Errors)

Term	edf	F	P-value
s(Temp)	4.206	5.814	< 0.001
s(Solar.R)	2.202	8.964	< 0.001
s(Wind)	2.193	5.142	< 0.001
s(Day)	8.128	1.955	0.041

Table 16: Cubic Spline Estimation for Smooth Terms (Standard Errors)

From Tables 15 and 16, it can be seen that at 5% significance level, there is no influence of the `month` period on the $\log$ concentration of `Ozone`, whereas all smooth terms in Table 16 have a significant effect. With effective degrees of freedom greater than 1 for the latter terms, one concludes that all smooth variables assume a non-linear relationship with $log(ozone)$. Model diagnostics for both fits are shown in Figures 11 and 12.

- The upper leftmost plot of both Figures, shows the normal Quantile-quantile plot from which the lines of dots is expected to lie very closely on the straight line suggesting validity of the distributional assumption. The log-transform fit (Figure 12) appears to satisfy this assumption.

- The upper right plot shows constancy of the variance (homoscedasticity). Here, it is not expected any dependence between the residuals and the linear predictors, that is, no trend in the plot is expected when the distributional assumptions are met. Again there seems to be constant variance in Figure 12 if the point on the bottom left is ignored.

- A symmetric distribution is expected to be seen from the histograms of the residuals on the bottom left plots since in theory these are normally distributed. The histogram in Figure 12 appears more consistent with normality than in Figure 11.

- Lastly, the bottom right plots compare the response and the fitted values. Here, it is expected a positive linear relationship which again appears to be satisfied.

To sum up, the logarithmic transformation of the response provides a better fit since the distributional assumptions appear to better satisfied.

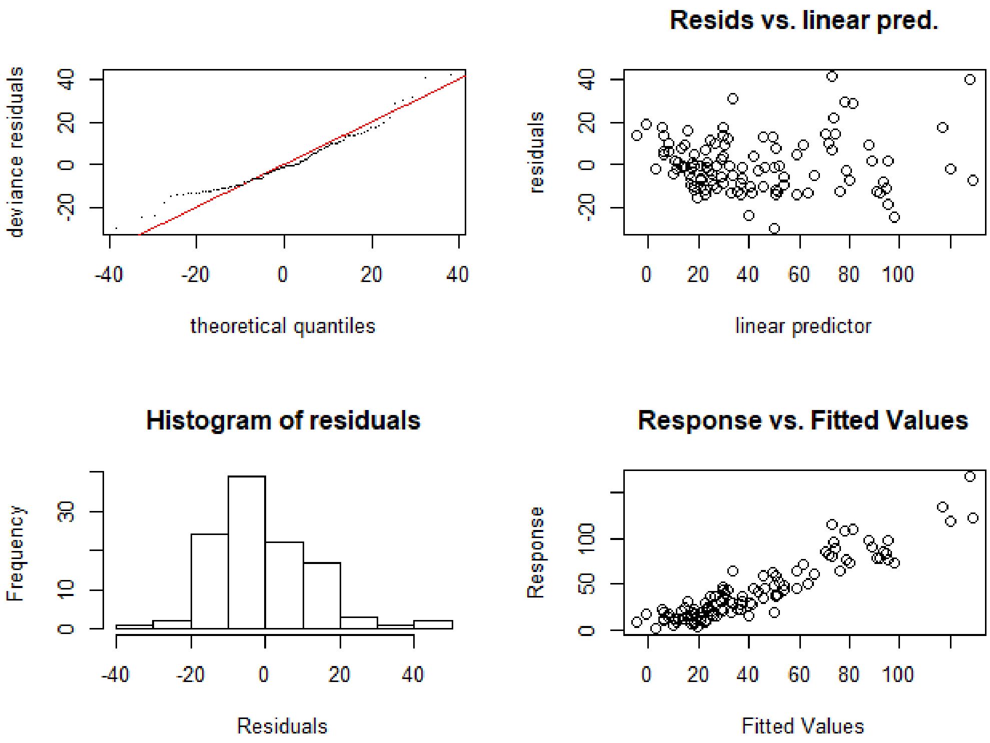


Figure 11: Model diagnostics for fit with ozone concentration

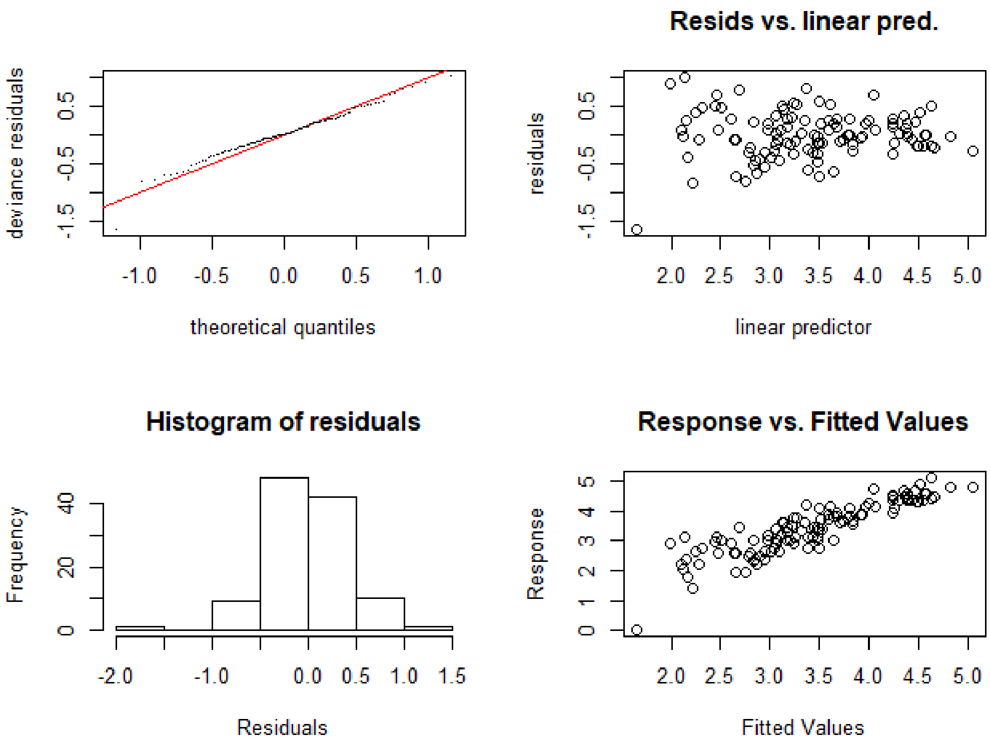


Figure 12: Model diagnostics for fit with log($ozone$)concentration

Lung function has been found to be deteriorated by exposure to ozone levels higher or equal to 70 ppb. Fit a model that relates the probability of having such concentration levels and the covariates considered above

Depending on the value of mean Ozone in air, it can be classified as a risky value for lung function. Now, the variable is dichotomized as $Ozone_c$, where values above 70ppb represent high risk for the lung function.

$$Ozone_c = \begin{cases} 1 & \text{if } Ozone > 70 \\ 0 & \text{if } Ozone \leq 70 \end{cases}$$

Resulting in 24 out 111 days, where high risk Ozone level were observed. The aim is now to model the probability of having risky Ozone levels, associated to the covariates used before. Given the observed variability in the complete data set, again a semi-parametric approach is investigated, more specifically a Generalized Additive Model (GAM) is considered.

It is assumed that $Ozone_{ci}$ follows a binomial distribution, and its mean μ_i is linked to the structured additive predictor

$$\eta_i^{struct} = f_1(z_{i1}) + ... + f_q(z_{iq}) + x_i^T\beta$$

by

$$\mu_i = h(\eta_i^{struct})$$

or

$$\eta_i^{struct} = g(\mu_i)$$

Where h is the response function and g is the link function, with $g = h^{-1}$ Therefore, the model is defined as

$$logit(P_{Ozone_{ci}}) = f(Temp_i) + f(Solar.R_i) + f(Wind_i) + f(Day_i) + \beta_0 + \beta_1 Month + \epsilon_i \qquad (6)$$

Where $P_{Ozone_{ci}}$ is the probability of having `Ozone` levels higher than 70 ppb. `Temperature`, `Solar radiation`, `Days` and `Wind` speed are assumed to have a nonlinear relationship with the response, represented by the smooth functions f. While the linear part corresponds to the parameters β_0 and β_1, the latter associated to the linear effect of the predictor `Month`.

The model is fitted including all covariates using Cubic splines (GAM Full Cubic) and using P-Splines (GAM Full P-Splines).

Term	Estimation	P-value
Intercept	-6.49(57.80)	0.911
June	-32.93(81.74)	0.687
July	-43.05(76.46)	0.573
August	-34.91(71.32)	0.624
September	-61.70(1654.98)	0.970

Table 17: GAM Full Cubic Estimation Linear Terms for Logistic Model

Term	edf	Chi.sq	P-value
s(Temp)	1.00	0.42	0.514
s(Solar.R)	1.86	0.20	0.932
s(Wind)	1.00	0.38	0.537
s(Day)	2.97	1.51	0.778

Table 18: GAM Full Cubic Estimation for Smooth Terms Logistic Model

Term	Estimation	P-value
Intercept	-3.67(1292.05)	0.998
June	-55.97(1312.31)	0.966
July	-74.19(1312.22)	0.955
August	-61.63(1306.21)	0.962
September	-112.16(21209.25)	0.996

Table 19: P-Spline Estimation Linear Terms for Logistic Model

Term	edf	Chi.sq	P-value
s(Temp)	1.00	0.07	0.779
s(Solar.R)	1.00	0.02	0.884
s(Wind)	1.00	0.06	0.797
s(Day)	3.30	0.50	0.934

Table 20: P-Spline Estimation for Smooth Terms Logistic Model

When fitting a GAM model using Cubic and P-splines some details shown in Tables 17 to 20 need to be discussed. First, in both models the linear part is characterized by having large estimates and standard errors, which is leading to non significant effect of the variable Month, as it is not the case in the continuous approach. Second, some of the smooth functions have $edf = 1$, Temperature and Wind for the GAM Full Cubic; and for the GAM Full P-splines the functions of the covariates Solar radiation, Temperature and Wind. Thus, it will be more appropriate to allow linear terms for these variables. Finally, it draws attention the fact that the adjusted R^2 and $Deviance$ have perfect values of 1 and 100%, respectively; and both models are resulting in negative UBRE values (GCV in non-continuous case), which is unlikely given it is defined as the summation of squared terms. Therefore, it is concluded that a semi-parametric model has estimation problems and may not be appropriate to model the dichotomization of the Ozone concentration.

A new approach is taken by fitting a generalized linear model (GLM), in particular a Logistic model is used to explain the probability of having Ozone levels above 70 ppb. Now, all covariates are assumed to have a linear relationship with the logit of the response.

$$logit(p_{Ozone_{ci}}) = \beta_0 + \beta_1 Temp_i + \beta_2 Solar.R_i + \beta_3 Wind_i + \beta_4 Day_i + \beta_5 Month \qquad (7)$$

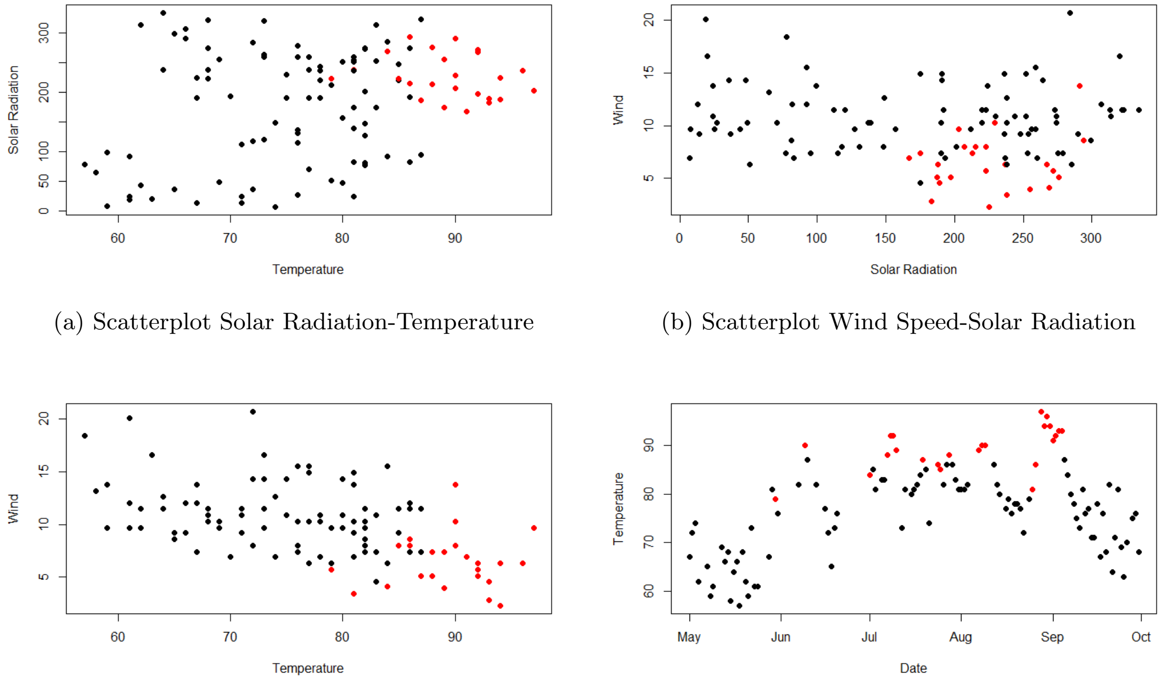
The estimated parameters are shown in Table 21.

Term	Estimate (Std.Error)	P-value
Intercept	-147.60 (71.77)	0.0398
Temp	2.00 (0.98)	0.0426
Solar.R	-0.00083 (0.02)	0.9689
Wind	-1.48 (0.71)	0.0374
June	-11.48 (9.15)	0.2098
July	-13.39 (8.79)	0.1277
August	-9.86 (7.99)	0.2174
September	-20.35 (14.11)	0.1493
Day	0.064 (0.10)	0.5219

Table 21: Logistic model of Ozone levels above 70 ppb

Two new issues arise from the logistic model. First, the estimation procedure of the parameters presents warnings due to observations of different categories that can be almost be perfectly separated by an hyperplane, also known as Quasi-complete separation. Secondly, a consequence of this difficulty is that the estimated regression coefficients and standard errors are inflated, resulting in most of the Wald tests non significant for the covariates.

The separation of the response variable is observed when looking at the relationship between the covariates. The most severe cases of separation of the response in the association of the covariates are shown in Figure 13, where red dots represent risky values of Ozone (Ozone values >70ppb). It is evident why the smoothing functions for some of the covariates were defined as linear, since the probability of having a risky value depends on clear thresholds.



(a) Scatterplot Solar Radiation-Temperature

(b) Scatterplot Wind Speed-Solar Radiation

(c) Scatterplot Wind Speed-Temperature

(d) Scatterplot Temperature-Date

Figure 13: Exploration separation of Ozone levels above 70ppb in covariates association

From this representation, is also clear how the unbalance of the two categories of Ozone may generate problems to estimate the models. Since the frequency of risky values is so low, a longer time line following the levels of Ozone could improve the analysis in order to find how the covariates impact the presence of risky Ozone levels.

References

Deb, P. & K. Trivedi, P. (1997). Demand for medical care by the elderly: A finite mixture approach. *12*, 313–36.

Appendix

Zero Inflated Distributions

Suppose that the base count, y, has f density (Poisson or NB). A separate component that inflates the probability of a zero is added.

$$Pr[y=j] = \begin{cases} \pi + (1-\pi)f(0) & if j = 0 \\ (1-\pi)f(j) & if > 0 \end{cases}$$

Where π is the estimated probability of having zero from a binary model, usually logit.

Akaike's Information Criterion

For n observations, p variables and the maximized value of the likelihood function $\hat{\mathcal{L}}$

$$AIC = 2p - 2\ln(\hat{\mathcal{L}})$$

Bayesian Information Criterion

For n observations, p variables and the maximized value of the likelihood function $\hat{\mathcal{L}}$

$$BIC = \ln(n)p - 2\ln(\hat{\mathcal{L}})$$

Wald Test

$$\frac{\hat{\beta} - \beta_0}{\mathsf{se}(\hat{\beta})} \sim N(0,1)$$

$\beta_0 = 0$ in the case of a significance test

Pearson Residuals

$$p_i = \frac{(y_i - \hat{\mu_i})}{\sqrt{\hat{\omega}_i}}$$

Where $\hat{\omega}$ is an estimate of the variance of the response variable. In Poisson $\omega = \mu$ and in Negative Binomial $\omega = \alpha\mu$ (α is the second parameter in the gamma distribution to get the negative binomial).

Deviance Residuals

$$d_i = sign(y_i - \hat{\mu_i})\sqrt{2[l(y_i) - l(\hat{\mu_i})]}$$

where $l(\hat{\mu})$ is the log-density of y evaluated at $\mu = \hat{\mu}$ and $l(y)$ is the log-density evaluated at $\mu = y$.

Standardized residuals, also called Studentized residuals

$$r_{si} = r_i / \sqrt{1 - h_{ii}}$$

Where r_i can be either Pearson or Deviance residuals.

Cook's Distance

$$D_c = \Big(\frac{p_i}{1 - h_{ii}}\Big)^2 \frac{h_{ii}}{p\phi}$$

Where p_i are Pearson residuals, h_{ii} is the i-th elementh of the diagonal of the Hat matrix, p is the number of parameters and ϕ is the dispersion considered in the model.

General CrossValidation function (GCV)

$$GCV = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{y_i - \hat{f}(x_i)}{1 - tr(S)}\right]^2$$

where, n = number of observations, S is the smoother matrix, $\hat{f} = Sy$ is the estimated smooth function and $tr(S)$ is the trace of S

Section 1 - Trace plots

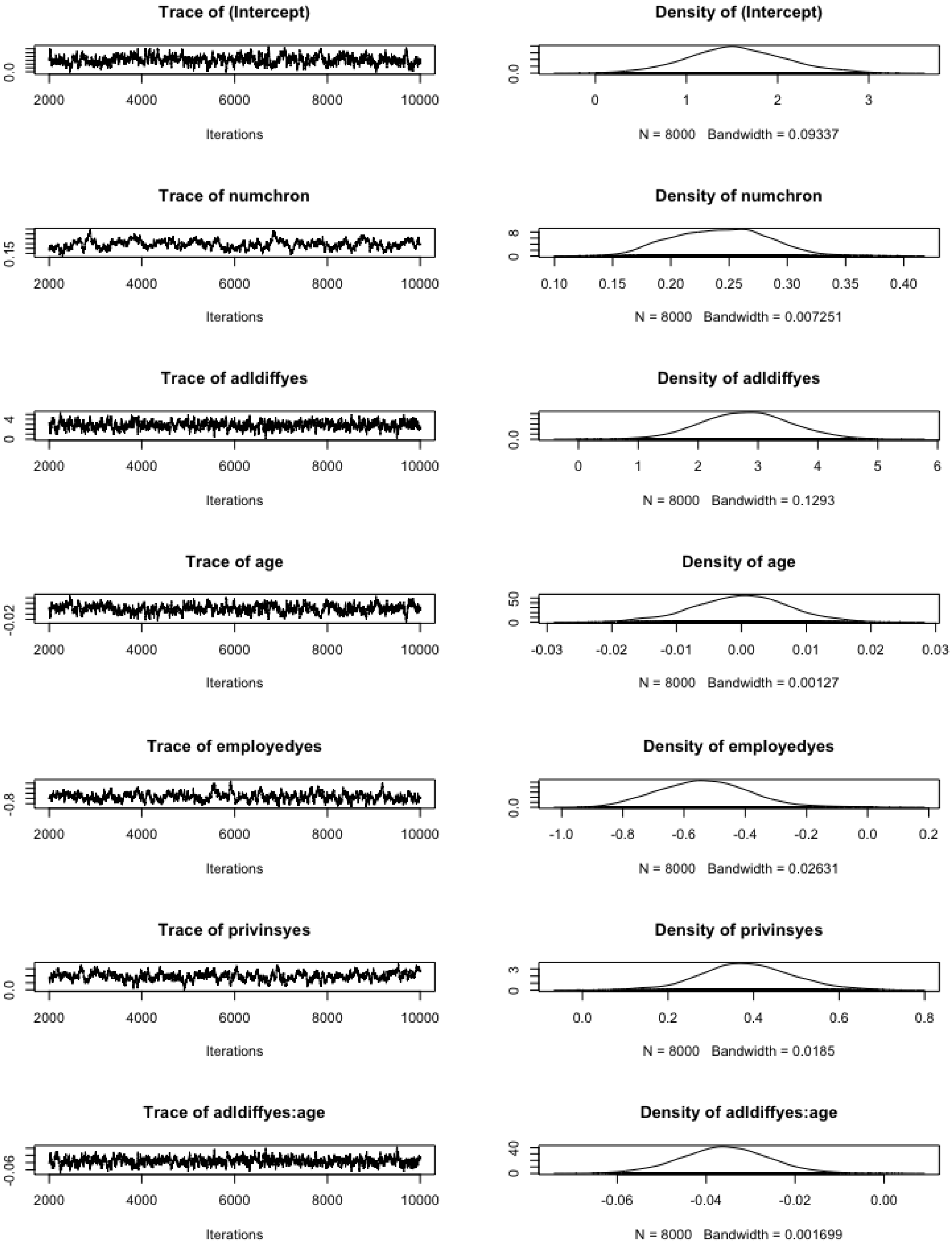


Figure 14: Trace Plots

Section 1 - Residuals

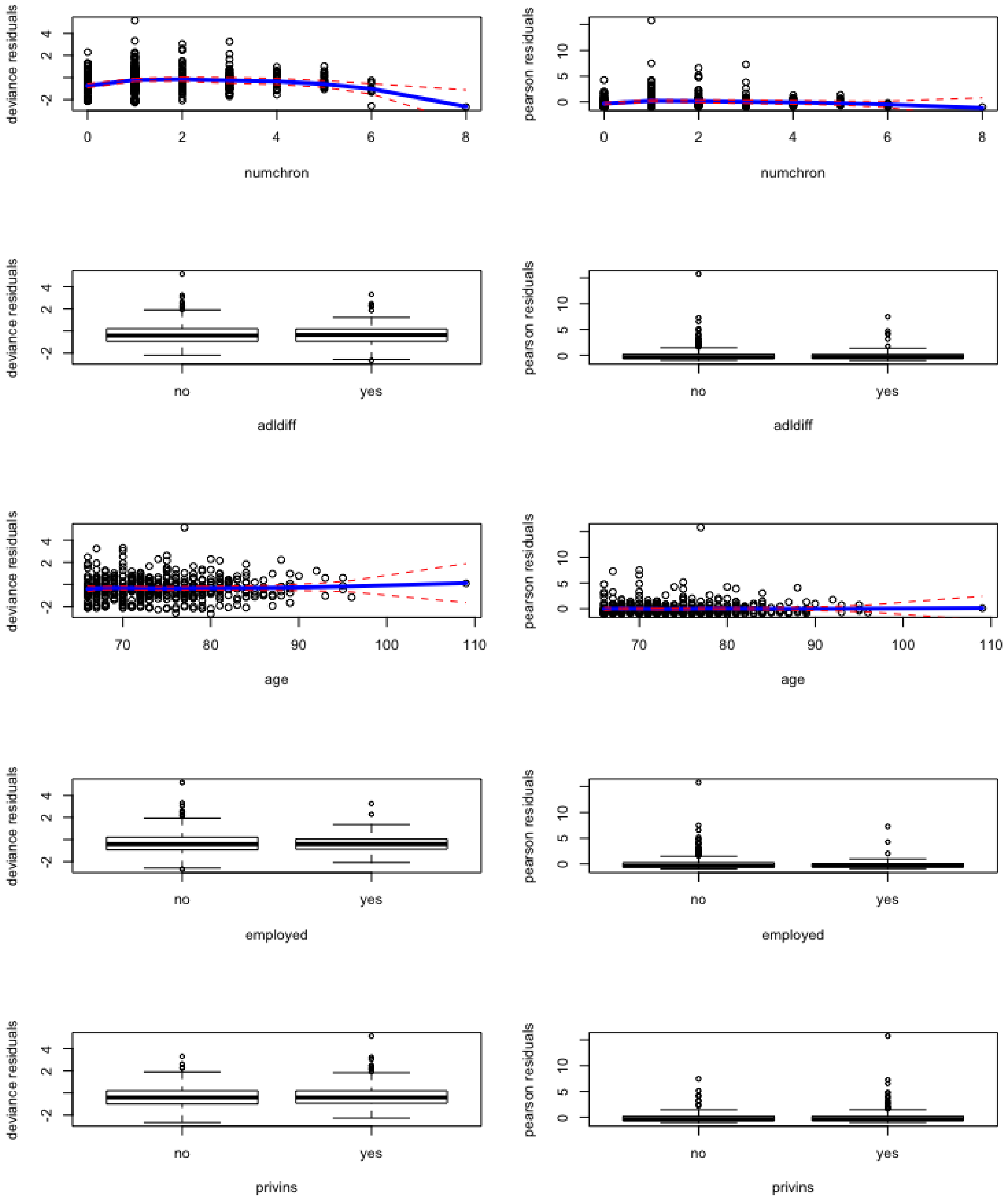


Figure 15: Residuals Plots