Assignment 2

02424 Advanced Dataanalysis and Statistical Modelling

Clara Brimnes Gardner (s153542)
Tobias Engelhardt Rasmussen (s153057)
Christian Dandanell Glissov (s146996)

May 26, 2021

Introduction and Description of Data

This project will concern itself with two different data sets. One is daily ozone measurements in 1976 from the area of Upland in California which is east of Los Angeles. The other is the level of clothing worn at a laboratory. A short description of the variables in each of the data sets in given in tables 1 and 10.

Ozone

Variable	Type	Description
`Ozone`	Continuous	O_3-concentration [ppm] (parts per million)
`Temp`	Continuous	Temperature in Fahrenheit
`InvHt`	Continuous	Inversion base height [ft]
`Pres`	Continuous	Daggett pressure gradient [mm Hg]
`Vis`	Continuous	Visibility [miles]
`Hgt`	Continuous	Vandenburg 500 milibar height [m]
`Hum`	Continuous	Humidity percentage
`InvTmp`	Continuous	Inversion base temperature in Fahrenheit
`Wind`	Continuous	Wind speed [mph]

Table 1: *List of included variables in the ozone data set. The Daggett pressure gradient is the pressure difference to the Daggett field station. The* ***`InvHt`*** *is the height at which the inversion layer starts. An inversion layer is when temperature rises with altitude due to faster cooling at lower levels. The* ***`InvTmp`*** *is the temperature at this height.*

Ozone model - Part 1

In this section we model the ozone concentration, `Ozone`, by using a set of explanatory variables, representing weather different measures of weather conditions. This first part will concern itself with only simple additive models, but of both classical GLM structure and various generalized linear model structures. In the end a model structure is chosen based on how well the model structure seems to fit the data.

Presentation of Data

Table 2 shows the summary statistics for the continuous variable in Table 1. It should be noted that apart from `InvTmp` the data set consists of only integer measurements.

Variable	Variance	Min	25% Quantile	Median	Mean	75% Quantile	Max
`Ozone`	64.18	1	5	11.78	10	17	38
`Temp`	209.06	25	51	62	61.75	72	93
`InvHt`	3254004.23	111	877.5	2112.5	2572.9	5000	5000
`Pres`	1275.72	-69.00	-9.00	24.00	17.37	44.75	107.00
`Vis`	6298.38	0.0	70.0	120.0	124.5	150.0	350.0
`Hgt`	11174.23	5320	5690	5760	5750	5830	5950
`Hum`	394.61	19.00	47.00	64.00	58.13	73.00	93.00
`InvTmp`	190.50	27.50	51.26	62.15	61.01	70.52	91.76
`Wind`	4.48	0.000	3.000	5.000	4.848	6.000	11.000

Table 2: *Summary statistics of the continuous variables in Table 1. Note that almost all the variable can only be positive by definition, and that since the temperature is given in fahrenheit, it is unlikely to be negative in California. Only pressure appears to have negative measurements.*

The scatterplots, densities and correlations can be seen in Figure 1.

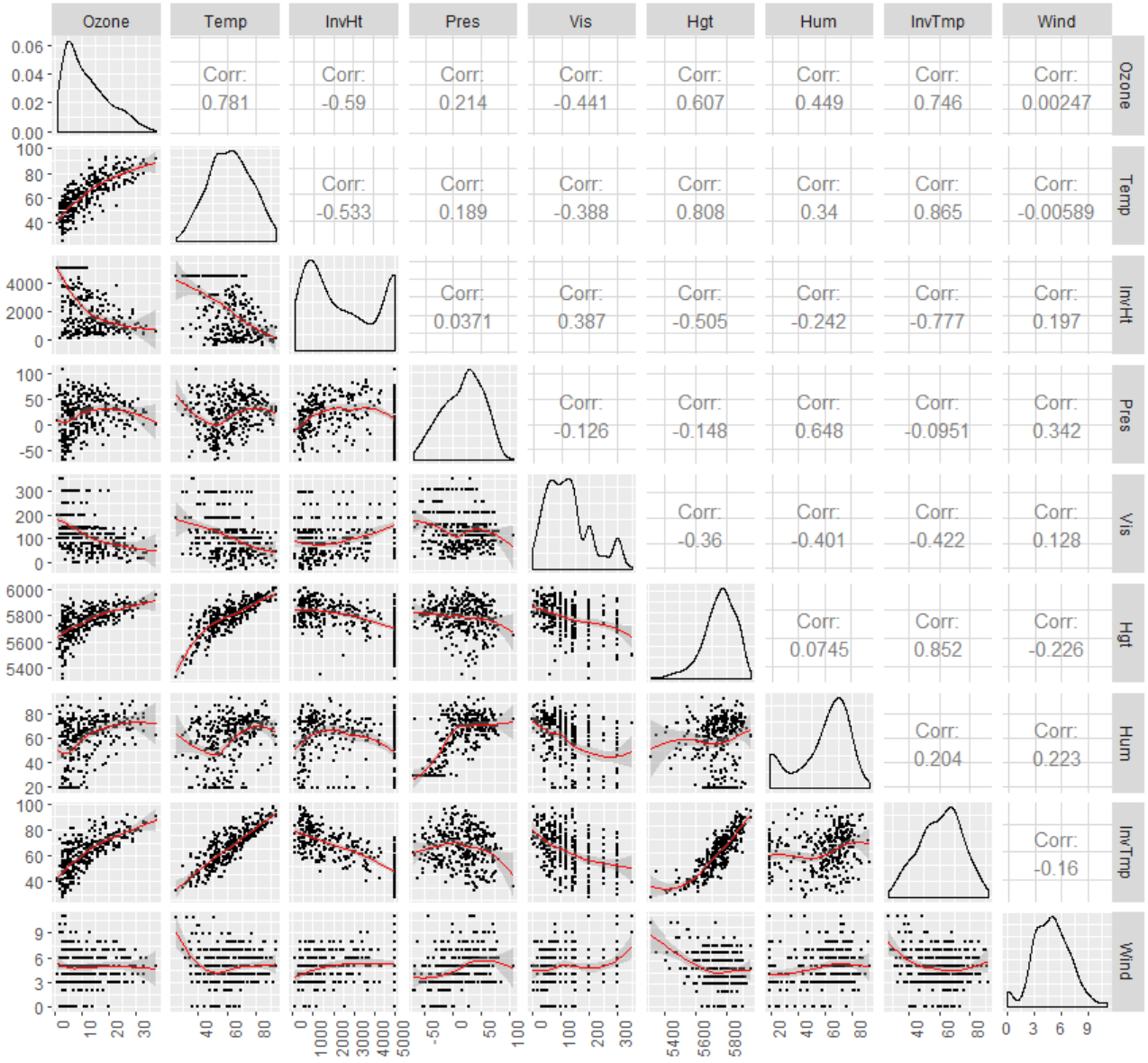

Figure 1: *Scatterplots, densities and correlation of the continuous variables in Table 1*

From Figure 1 strong correlation can be observed between some of the variables, such as between `Hgt` (height) and `Temp` (temperature), and between `InvTemp` (inversion base temperature) and `Temp`. `Ozone` consists mainly of integer values, which causes the more or less straight lines seen in some of the plots. It is also seen that the response variable `Ozone` looks very skewed, while some of the explanatory variables have a fairly regular shape. By investigating the histogram in Figure 2 of `Ozone` it is possible to get an idea of which model might be appropriate in modelling the problem.

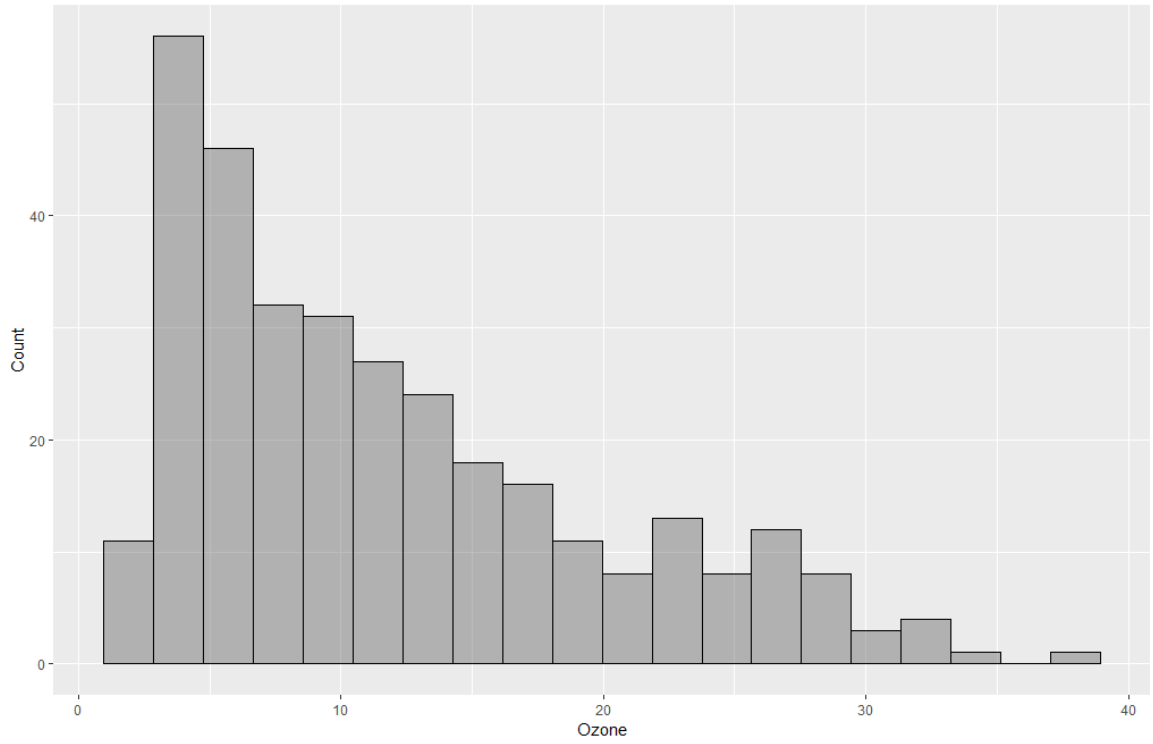

Figure 2: *Histogram of the Ozone variable. A very skewed distribution is observed.*

A plot of the autocorrelation of the response variable is shown in Figure 3. From this plot it is seen that the correlation between consecutive time points is high.

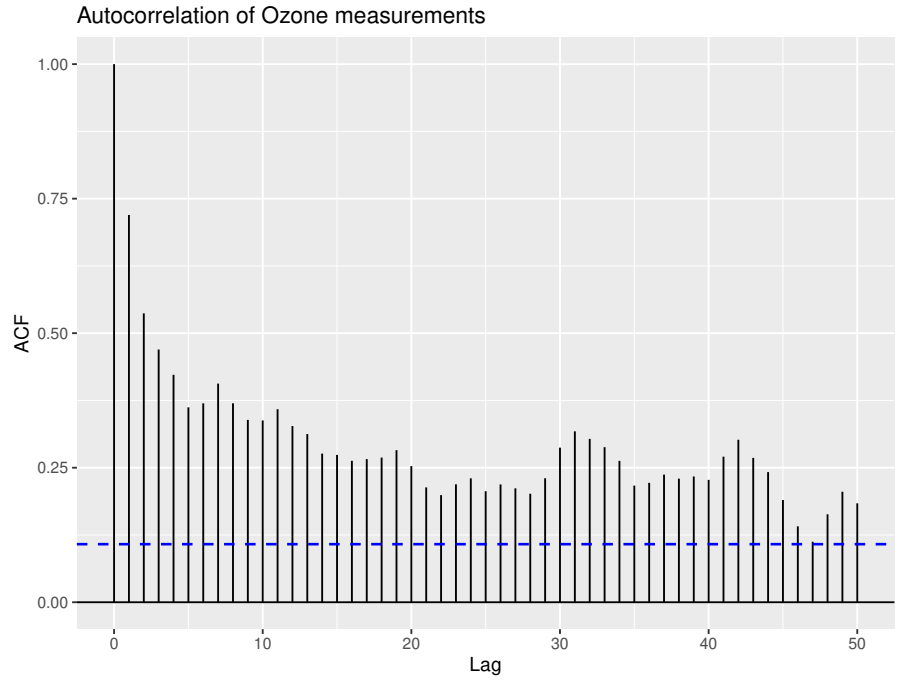

Figure 3: *The autocorrelation of* `Ozone`

A general linear model

A simple general linear model is fitted to the ozone data, where ozone will be the response variable and the rest of the variables predictors with no higher order polynomials and no interaction terms. Note that due to the high correlation between some of the parameters, mentioned earlier, the varibles `InvTmp` and `Hgt` will not be included in the model. The initial model is chosen to be:

$$\texttt{Ozone}_i = \beta_0 + \beta_1\ \texttt{Temp}_i + \beta_2\ \texttt{InvHt}_i + \beta_3\ \texttt{Pres}_i + \beta_4\ \texttt{Vis}_i + \beta_5\ \texttt{Hum}_i + \beta_6\ \texttt{Wind}_i + \epsilon_i \tag{1}$$

Where $\epsilon_i\ N(0, \sigma^2),\ i = 1 \dots 330$. The significance of each of the parameters is assessed using the likelihood ratio test and type II partitioning. The initial anova-table with relevant p-values is given in Table 3.

Parameter	Sum Sq	Df	F value	Pr(>F)
Temp	4620.51	1	226.14	0.0000
InvHt	590.07	1	28.88	0.0000
Pres	5.62	1	0.28	0.6002
Vis	62.07	1	3.04	0.0823
Hum	312.53	1	15.30	0.0001
Wind	9.36	1	0.46	0.4991
Residuals	6599.42	323		

Table 3: *The initial anova-table for the model given in Equation 1*

The following parameters are removed in the given order:

1. `Pres`
2. `Wind`
3. `Vis`

This results in a model on the following form:

$$\texttt{Ozone}_i = \beta_0 + \beta_1\ \texttt{Temp} + \beta_2\ \texttt{InvHt} + \beta_3\ \texttt{Hum} + \epsilon_i \tag{2}$$

With the anova-table given in Table 4, and the estimates given in Table 5.

Parameter	Sum Sq	Df	F value	Pr(>F)
`Temp`	4998.17	1	244.18	0.0000
`InvHt`	769.08	1	37.57	0.0000
`Hum`	683.17	1	33.37	0.0000
Residuals	6673.07	326		

Table 4: *The anova of the reduced model given in Equation 2.*

Parameter	Estimate	Std. Error
(Intercept)	-10.494	1.6160
`Temp`	0.329	0.0211
`InvHt`	-0.001	0.0002
`Hum`	0.077	0.0134

Table 5: *The estimates of the model given in Equation 2*

To assess how well the model fits the data, a diagnostics plot is made, which is given in Figure 4. From Residuals vs Fitted values plot it is clear, that the model is not appropriate for fitting the given data, even though the rest of the plots look reasonably nice. An attempt to fix this problem could be to transform the response variable, hence this will be attempted using a box-cox transformation in the following section.

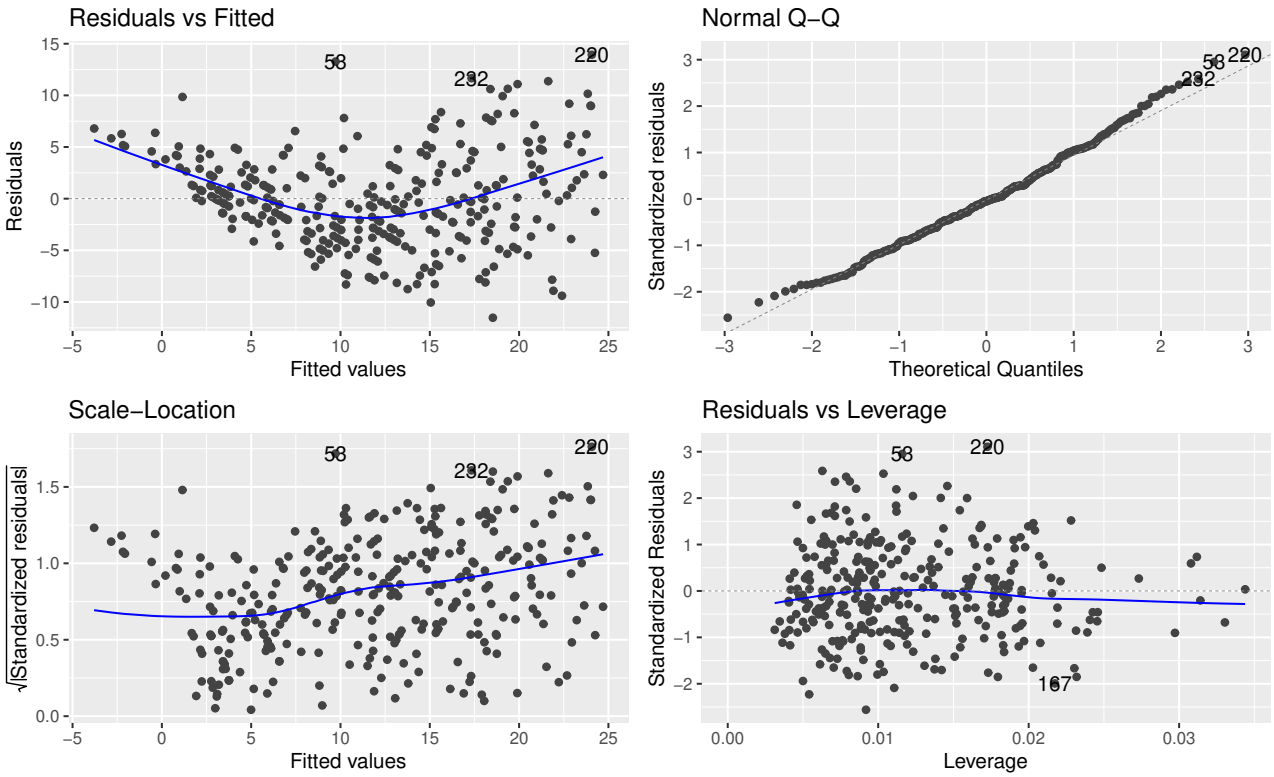

Figure 4: *The diagnostics plot of the model given in Equation 2*

Box Cox Transformation

In order to fix the problem we are trying the box-cox transformation given by:

$$z_i = \begin{cases} \frac{y_i^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0 \\ \ln y_i & \text{if } \lambda = 0 \end{cases} \tag{3}$$

In which y_i is the non-transformed response, which in the case is `Ozone`. That is we need to estimate the value of λ that results in the highest log-likelihood, when the model in Equation 2 is used. Figure 5 shows the profile log-likelihood given different values of λ. The λ that maximises the log-likelihood is found to be $\lambda_{opt} = 0.2828$, hence the response variable will be transformed using this value, i.e.

$$\texttt{Ozone}_{z,i} = \frac{\texttt{Ozone}_i^{0.2828} - 1}{0.2828} \tag{4}$$

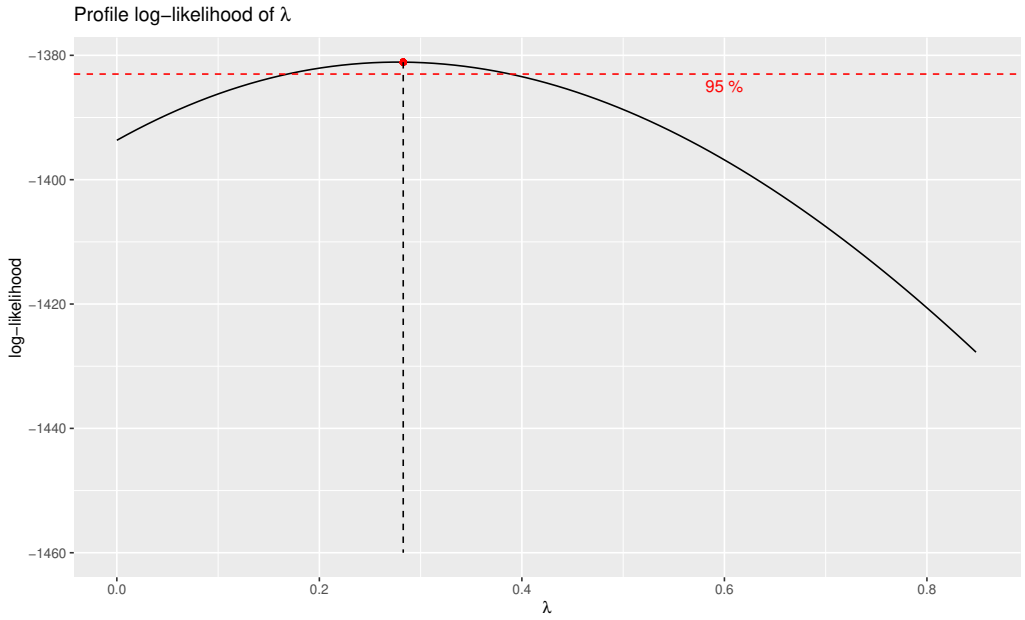

Figure 5: *The profile log-likelihood given λ. The value of λ that maximizes the log-likelihood is found to be 0.2828*

After having made this transformation the model is trained using the same structure. The anova-table and the estimates are shown in Table 6. Since all the variables still seem to be significant, the model is not reduced. The diagnostics plot of the model using the transformed response is shown in figure Figure 6, and the final model becomes:

$$\texttt{Ozone}_{z,i} = -0.4785146 + 0.0564111 \cdot \texttt{Temp}_i - 0.0002076 \cdot \texttt{InvHt}_i + 0.0129173 \cdot \texttt{Hum}_i + \epsilon_i \tag{5}$$

Where $\epsilon_i \sim N(0, 0.74^2),\ i = 1 \dots 330$

	Sum Sq	**Df**	**F value**	**Pr(>Chisq)**
`Temp`	146.383	1	264.284	$2.2 \cdot 10 \cdot \{-16\}$
`InvHt`	32.867	1	59.339	$< 1.608 \cdot 10^{-13}$
`Hum`	19.039	1	34.374	$1.115 \cdot 10^{8}$
	180.56	326		

Parameter	**Estimate**	**Std. Error**
Intercept	-0.4785	0.26590
`Temp`	0.0564	0.00346
`InvHt`	-0.0002	0.00003
`Hum`	0.0129	0.00220

Table 6: *The anova table and estimates for the model in Equation 5*

There still seems to be a bit of structure in the residuals, but the residuals of the model using the transformed response seems significantly better than the one using the response directly. The structure left in the data might be due to the dependence between observations, since these are taken on consecutive days, and this is not taken care of. Also `Ozone` consists only of integer values, which means that a lot of the measurements are repeated, hence even after transformation these will result in the linear structure in the residuals. Looking at the ACF and PACF of the residuals in Figure 7, there seem to be some significant autocorrelation between the residuals up to a lag 1.

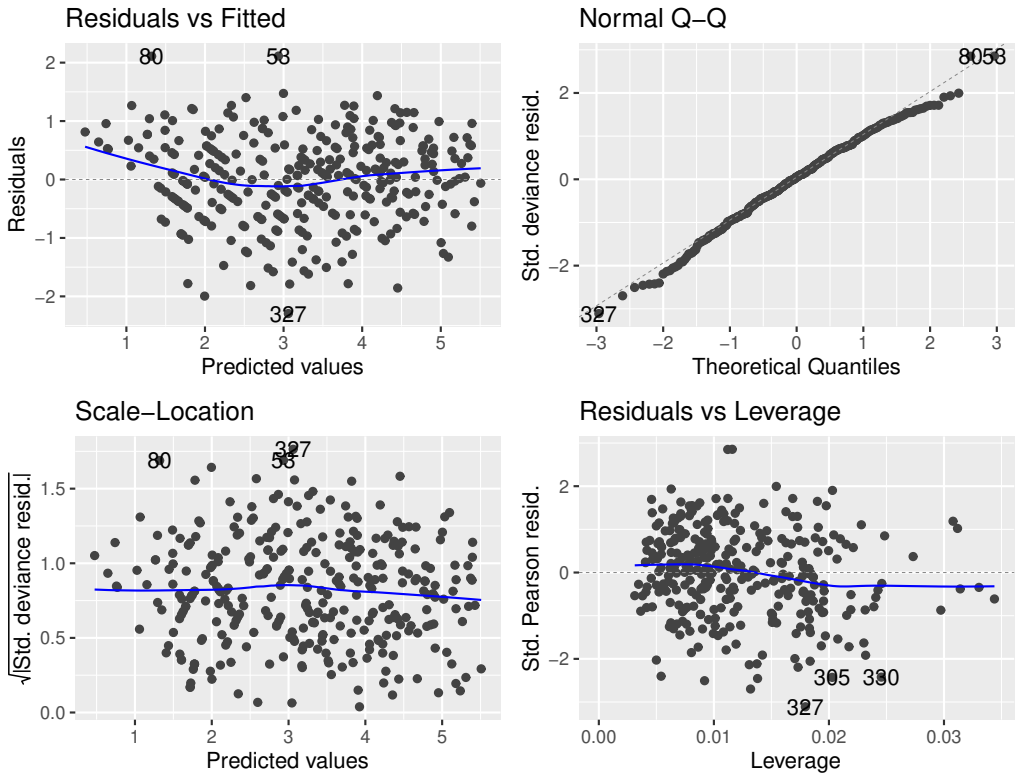

Figure 6: *The diagnostics plot of the model given in Equation 5. Note that these residuals look better than before transformation in Figure 4*

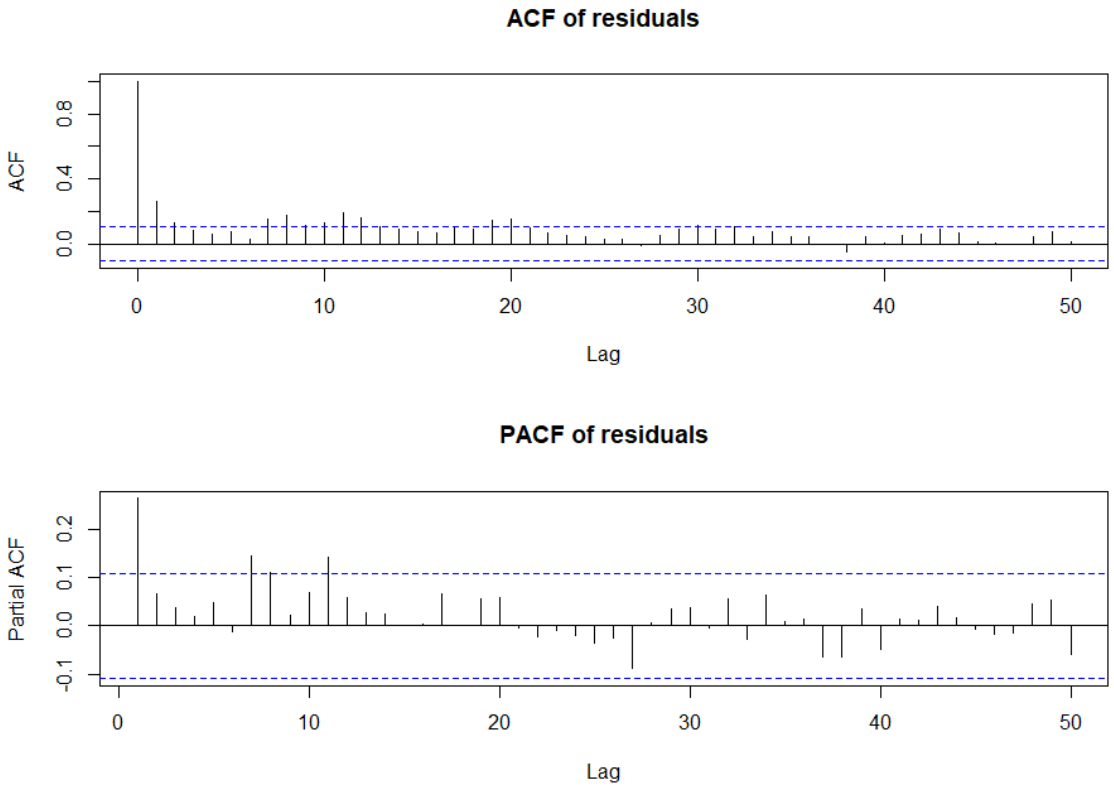

Figure 7: *ACF and PACF of the residuals.*

A generalised linear model

As is seen in Figure 2 the distribution of ozone is very positively skewed and the usual assumptions for the classical generalized linear model (GLM) model might not be fulfilled, such as the requirement of a gaussian distributed response variable. Other modelling methods may be taken into consideration. In this part two other distributions will be looked at for modelling the problem. As the distribution of the ozone variable is positively skewed and continuous the gamma and inverse-gaussian distribution may be appropriate distribution assumptions. For the GLM it is possible to use other distributions from the expoential dispersion family to fit the response. In contrast to a direct mean value μ from the classical GLM, the linear part changes to become a function of the mean value $g(\mu_i) = X\beta$. The GLM for a stochastic variable Y describes some affine hypothesis given by the affine component $\eta = X\beta$, where the component is connected by a link function $g(\mu_i) = \eta$. For both the gamma and the inverse-gaussian a log and inverse link will be used. The inverse link is given as

$$\eta = g(\mu) = \frac{1}{mu} = X\beta \tag{6}$$

$$\mu = \frac{1}{X\beta} \tag{7}$$

And the log link as

$$\eta = g(\mu) = \log(mu) = X\beta \tag{8}$$

$$\mu = \exp X\beta \tag{9}$$

For the the gamma, it is assumed $Y \sim Gamma(\alpha, \frac{\mu}{\alpha})$, where α is the shape parameter and $\frac{\mu}{\alpha}$ is the scale. For inverse-guassian, $Y \sim IG(\mu, \lambda)$, μ and λ is the shape and scale parameters, respectively.

A simple additive model will used for each of the GLM fit. Looking at the initial goodness of fit to deem sufficiency of the models in Table 7 it is clear that the models are sufficient on a $\alpha = 0.05$ significance level.

	Gamma, log	Gamma, inverse	IG, log	IG, inverse
Null Deviance	167	167	20.15	20.15
Residual Deviance, $\mathrm{D}(\boldsymbol{y}; \boldsymbol{\mu}(\widehat{\boldsymbol{\beta}}))$	50.64	57.54	8.354	8.938
$P\left[\chi^2(n-k) \geq \mathrm{D}(\boldsymbol{y}; \boldsymbol{\mu}(\widehat{\boldsymbol{\beta}}))\right]$	1	1	1	1

Table 7: *Goodness of fit test, with* $n - k = 330 - 9 = 321$ *degrees of freedom*

The models are reduced using a type 2 F-test and a summary and residuals plot of all the models can be seen in the appendix. The residuals look decent, but there seem to be a slight decreasing trend in the deviance for the scale-location in the inverse Gaussian models.

Choosing a model

To choose between the classical GLM and the other GLM models the AIC value is used. To compare the transformed Gaussian model the AIC has to be adjusted, this is done by looking at the jacobian of the transformation, $z_n = h(y_n)$

$$\mathrm{AIC}'_z = \mathrm{AIC}_z - 2\sum_{i=1}^{N} \log \left| \frac{dh}{dy} \right|_{y=y_i}$$

Where $\frac{dh}{dy}$ is the jacobian of the transformation and z is the box-cox transformed variable of ozone. In Table 8 the results of the AIC can be seen for each model.

	Gamma, log	Gamma, inverse	IG, log	IG, inverse	Gaussian, boxcox	Gaussian, identity
AIC	1803.609	1848.492	1927.948	1951.85	1794.982	1938.724

Table 8: *AIC of the models*

The Gaussian model with a box-cox transformation is the better model according to the AIC, with a generalized Gamma model being close to the optimal AIC value. Based on the above table the chosen model will be the box-cox transformed Gaussian model.

Evaluating the scaled dispersion matrix

The chosen model is a Gaussian model, looking at the likelihood the unit variance function can be found, first the likelihood is written to the form of an exponential dispersion family, definition 4.2 in the textbook.

$$\begin{aligned} f_Y\left(y;\mu,\sigma^2\right) &= \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{(y-\mu)^2}{2\sigma^2}\right] \\ &= \frac{1}{\sqrt{2\pi}\sigma}\exp\left\{\frac{1}{\sigma^2}\left(\mu y-\frac{\mu^2}{2}\right)-\frac{y^2}{2\sigma^2}\right\} \end{aligned} \tag{10}$$

Here one can directly get the cumulant generator $\kappa(\theta) = \frac{\mu^2}{2}$. The unit variance function is then found by definition 4.3 in the textbook and the canonical link function $\theta = \tau^{-1}(\mu)$

$$V(\mu) = \frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\frac{\theta^2}{2} = 1 \tag{11}$$

The weight matrix can then be found using (4.43) or (4.44) in the textbook as it is for the case of the canonical link, as the weight $w_i = 1$ this gives:

$$\mathbf{W}(\beta) = \mathbf{diag}\left\{V(\mu_i)\right\} = \mathbf{I} \tag{12}$$

To find the scaled dispersion matrix, the model dispersion, σ^2 is needed. It can be found by utilising the Pearson goodness of fit statistic,

$$X^2 = \sum_{i=1}^{n} w_i \frac{\left(y_i-\widehat{\mu}_i\right)^2}{V\left(\widehat{\mu}_i\right)} \tag{13}$$

As seen this is simply the residual deviance divided by degrees of freedom in the Gaussian case as the unit variance is constant, where $X^2 = 180.5661$ for the model in Equation 5, and the model dispersion is estimated with 326 degrees of freedom

$$\widehat{\sigma}^2_{\text{Pears}} = \frac{X^2}{n-k} = 0.5538839$$

The dispersion matrix is then shown in (4.66) in the textbook

$$\mathrm{D}[\widehat{\boldsymbol{\beta}}] = \boldsymbol{\Sigma} = \left[\boldsymbol{X}^T\boldsymbol{I}(\boldsymbol{\beta})\boldsymbol{X}\right]^{-1} \tag{14}$$

With $\mathbf{X}$ as the design matrix. The scaled dispersion matrix is then

$$\boldsymbol{\Sigma}_{scaled} = \boldsymbol{\Sigma}\cdot\widehat{\sigma}^2_{\text{Pears}} \tag{15}$$

The result can be seen below

$$\boldsymbol{\Sigma}_{scaled} = \begin{bmatrix} 7.070640e-02 & -7.479105e-04 & -4.984091e-06 & -1.723307e-04 \\ -7.479105e-04 & 1.204088e-05 & 4.615386e-08 & -1.968279e-06 \\ -4.984091e-06 & 4.615386e-08 & 7.265633e-10 & 4.550527e-09 \\ -1.723307e-04 & -1.968279e-06 & 4.550527e-09 & 4.854146e-06 \end{bmatrix} \tag{16}$$

The solution shows the same as when evaluating it in `R` using `summary(model_gamma)$cov.scaled`

Ozone model - Part 2

In this part the chosen model from *Part 1* will be expanded and developed to make a better model. To do this we will consider both higher order polynomials and interactions. The results will be evaluated and presented. Based on the AICs of the different models attempted earlier given in Table 8, we choose to go with the classical GLM with the box-cox transformation.

Developing the classical GLM with box-cox transformation

Since we are given a fairly big data set, we can afford a initial model with many parameters in terms of degress of freedom. A model with 3-way interactions does not seem to add more information than a model with 2-way interactions, hence the initial final model will be with all 2-way interactions. Note again that the parameters `InvTmp` and `Hgt` will not be considered due to high correlation with other variables. Furthermore, some of the parameters showed a quadratic relationship with the response when fitting a smooth non-linear function, hence each of the squared parameters will also be part of the initial final model. The initial model will include a total of 28 parameters including the intercept.
Before the model is reduced, the box-cox parameter λ is estimated again using this, bigger model. The parameter is estimated to be $\lambda_{opt} = 0.2741$, which is not far from the one estimated earlier, which was expected. The profile-likelihood can be seen in appendix in Figure 23.
The model is reduced using likelihood ratio test and type II partitioning. A list of the order in which parameters are removed is shown in appendix in Table 16. After reducing the model, the final model is found to have the following structure:

$$\texttt{Ozone}_{z,i} = \beta_0 + \beta_1\ \texttt{Temp}_i + \beta_2\ \texttt{InvHt}_i + \beta_3\ \texttt{Pres}_i + \beta_4\ \texttt{Vis}_i + \beta_5\ \texttt{Hum}_i + \beta_6\ \texttt{Wind}_i + \beta_7\ \texttt{Pres}_i^2 + \tag{17}$$

$$\beta_8\ \texttt{Hum}_i^2 + \beta_9\ \texttt{Wind}_i^2 + \beta_{10}\ \texttt{Temp}_i \cdot \texttt{Hum}_i + \beta_{11}\ \texttt{InvHt}_i \cdot \texttt{Hum}_i + \beta_{12}\ \texttt{Vis}_i \cdot \texttt{Wind}_i + \epsilon_i \tag{18}$$

The anova-table and estimates can be found in Table 9.

Parameter	Sum Sq	Df	F value	Pr(>F)
Temp	93.51	1	220.12	0.0000
InvHt	17.67	1	41.59	0.0000
Pres	2.64	1	6.21	0.0132
Vis	1.56	1	3.67	0.0565
Hum	1.16	1	2.73	0.0996
Wind	1.41	1	3.33	0.0690
$\texttt{Pres}^2$	5.13	1	12.07	0.0006
$\texttt{Hum}^2$	4.24	1	9.98	0.0017
$\texttt{Wind}^2$	2.77	1	6.51	0.0112
Temp:Hum	5.22	1	12.28	0.0005
InvHt:Hum	8.42	1	19.82	0.0000
Vis:Wind	2.25	1	5.30	0.0220
Residuals	134.67	317		

(a) *The anova-table of the final model*

Parameter	Estimate	Std. Error
(Intercept)	0.919112	0.816259
Temp	0.011864	0.011291
InvHt	0.000087	0.000063
Pres	0.005294	0.002124
Vis	-0.003326	0.001129
Hum	0.023118	0.016910
Wind	0.087366	0.057808
$\texttt{Pres}^2$	-0.000106	0.000030
$\texttt{Hum}^2$	-0.000391	0.000124
$\texttt{Wind}^2$	-0.014850	0.005819
Temp:Hum	0.000637	0.000182
InvHt:Hum	-0.000005	0.000001
Vis:Wind	0.000466	0.000202
Residual error	0.651783	

(b) *The estimates and standard errors of the final model*

Table 9: *Statistics on the final model*

To assess how well the model seems to fit the data, we will inspect the AIC, the diagnostics plot, and the autocorrelation. The AIC is found to be 1728.9, which is better than the simple models considered earlier. The diagnostics plot, the autocorrelation and the partial autocorrelation can be found in Figure 8 and Figure 9 respectively. The residual plots of the predictors can be seen in Figure 24.

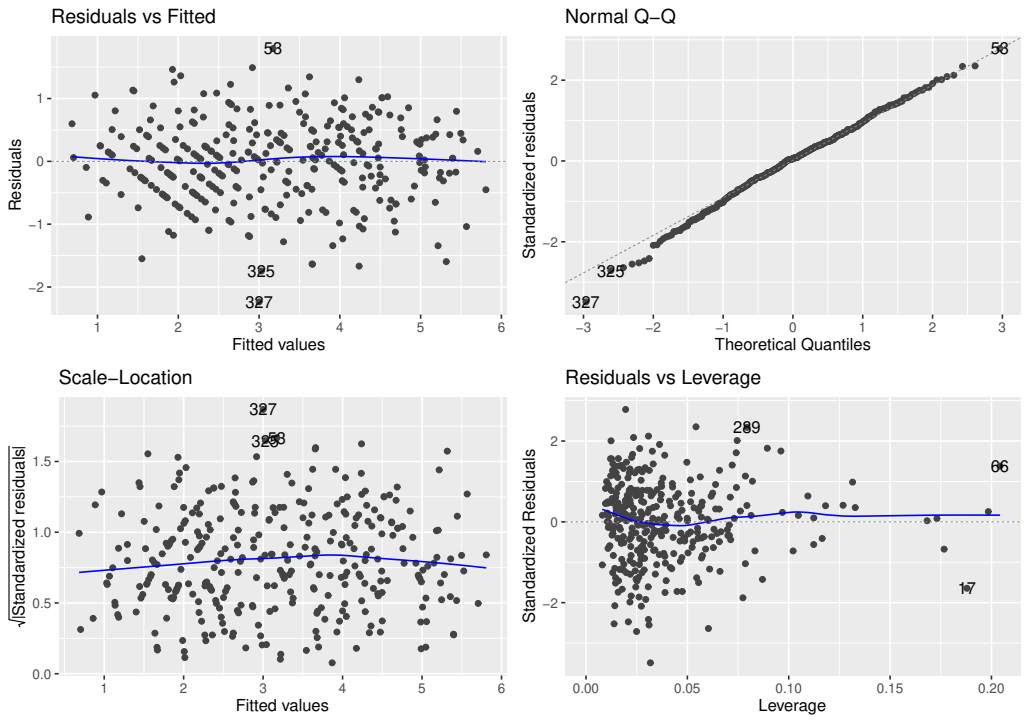

Figure 8: *Diagnostics plot for the final model*

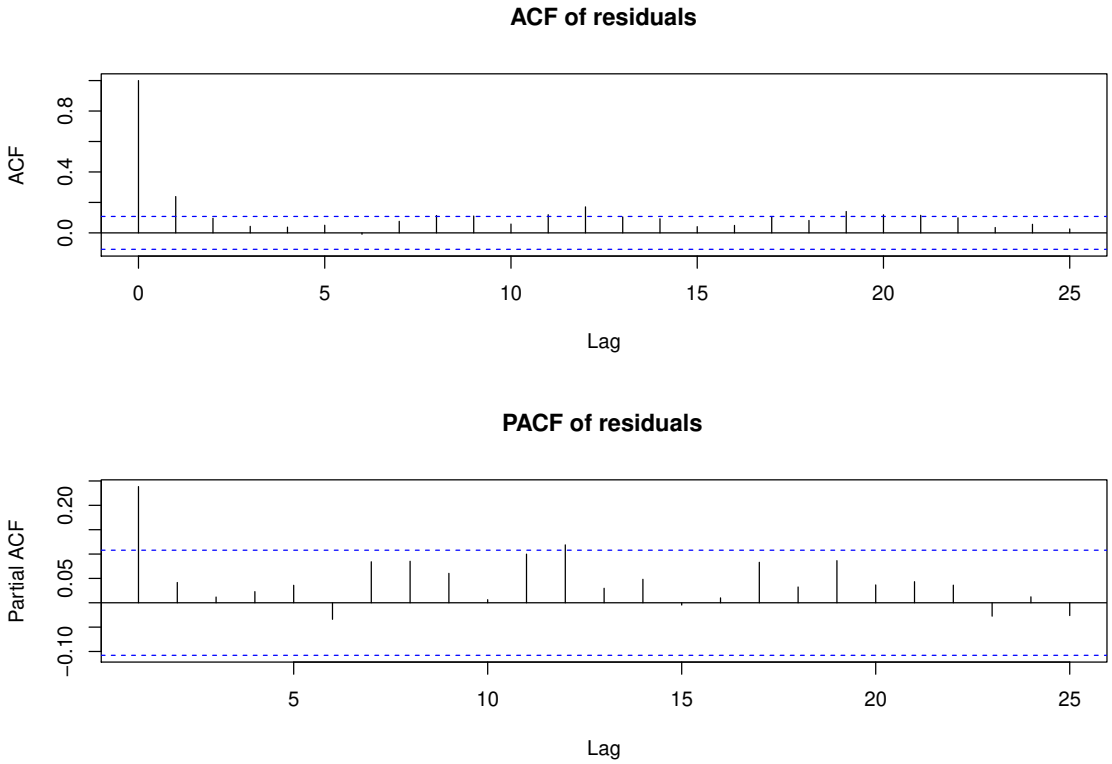

Figure 9: *Autocorrelation and partial autocorrelation of the residuals of the final model*

The diagnostics plot in Figure 8 show very nice residuals apart from the linear structure, that is due to the integer valued `Ozone` measurements. If this is due to rounding at some point, or imprecise measuring instruments is not given, however it makes some observations appear multiple times. The ACF and PACF still show some correlation in lag 1, which makes it a problem to verify the assumption on independent observations. One could consider adding an auto regressive term with one lag (AR(1)) for the response variable, this has been implemented, but due to simplicity reasons and to constrain ourselves within the scope of the course the results of the PACF and ACF is only shown in the appendix Figure 25, where one can see that there is no longer a significant lag in the ACF and PACF. We conclude that our final model is reasonably appropriate in describing the trends in the data. The model shows that all the variables, apart from the ones removed beforehand, are significant in describing the ozone concentration. This makes plotting the predictions extensive and complicated. One plot is made for each variable with the value of the rest of the variables set to its mean. Figure 10 shows the predictions given the temperature. The rest of the plots can be found in appendix.
Looking at all the predictions it is clear that many point fall outside of the prediction interval. This is due to the rest of the variables being held constant, which means that some of the variation will not be visualised in the plots. Overall the model seems to capture the fluctuations of the data well.

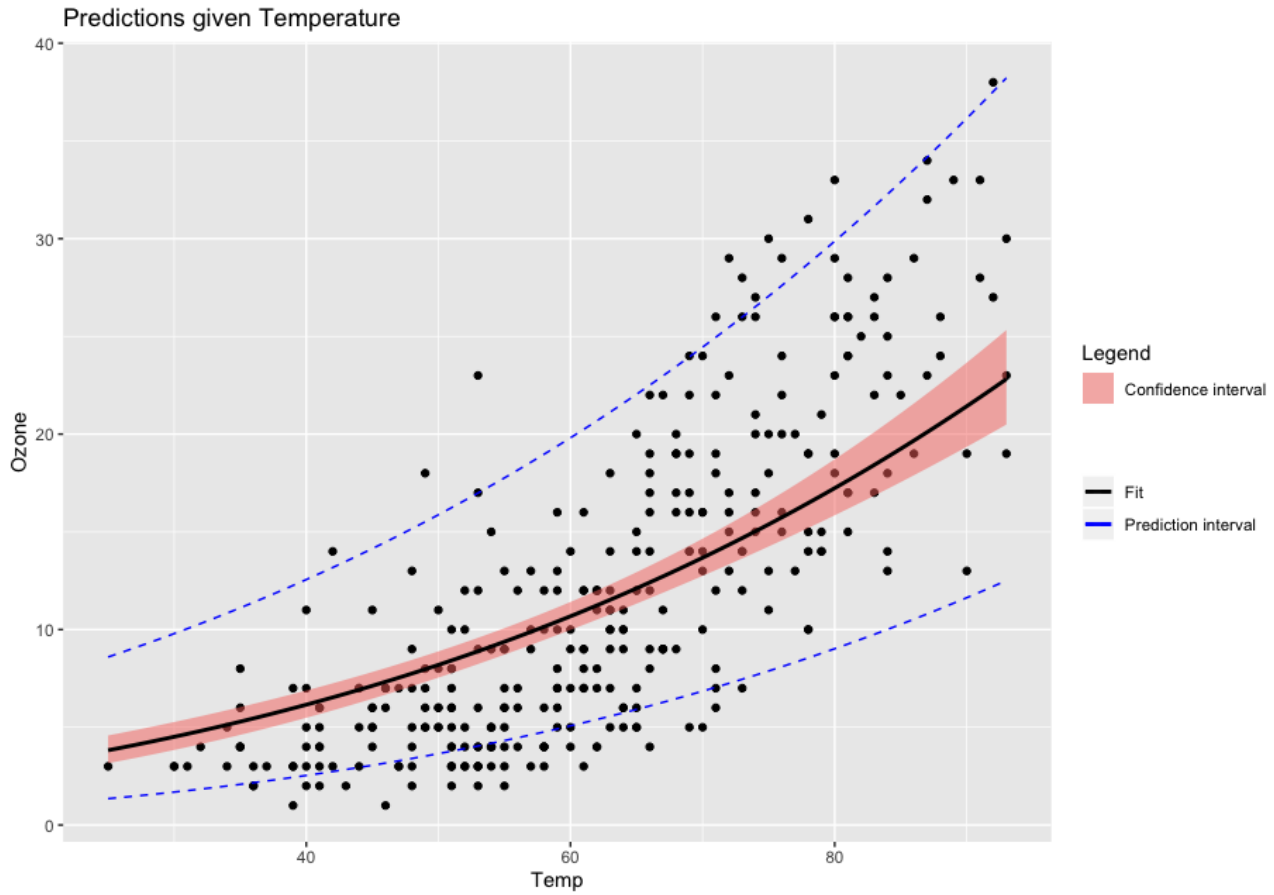

Figure 10: *The predictions of the final model given temperature, the confidence interval, and the prediction interval. Notice that a lot of points fall out of the prediction interval because the remaining variables are held constant.*

Part 3: Clothing insulation level: Count data

Presentation of Data

In this part of the assignment we consider a modified version of the clothing data-set considered in the first assignment. Table 10 shows descriptions of the different variables.

Variable	Type	Description
`clo`	Continuous	Number of times the subject changes clothes
`nobs`	Factor	Number of observations during the day
`tOut`	Continuous	Outdoor temperature
`tInOp`	Continuous	Indoor operating temperature
`sex`	Factor	Sex of the subject
`time`	Continuous	Total time of observation
`subjId`	Factor	Identifier for subject
`day`	Factor	Day (within the subject)

Table 10: *List of included variables in the clothing level data set.*

Figure 11 shows a histogram of the variable `clo`. It is seen that most people do not change clothes during the day. Especially men are reluctant to change, and no men change more than two times. Above 20% of the women changes clothes 1 time and 2 times. There is a very little percentage of women changing clothes 3 or 4 times.

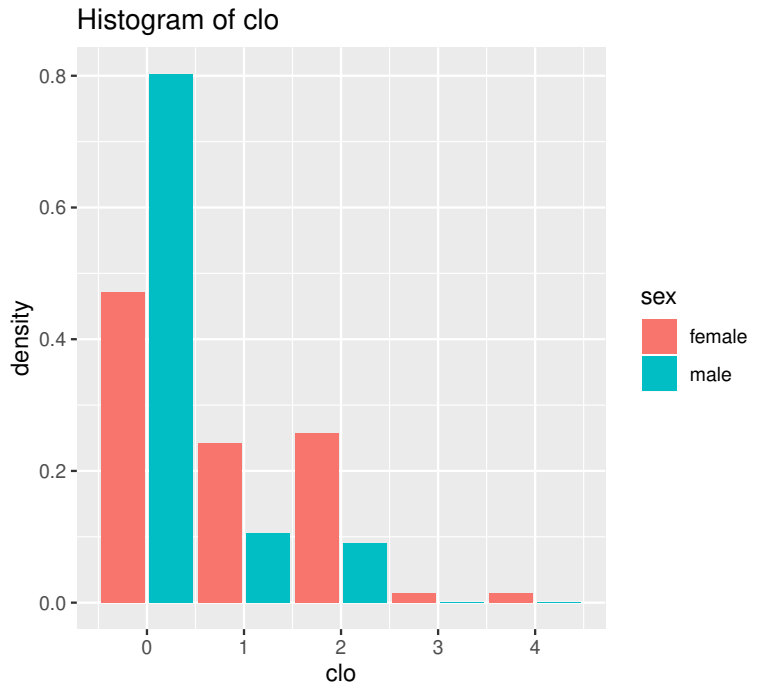

Figure 11: *Histogram of* `clo` *for each sex.*

Figure 12 shows `clo` against the three continuous variables. The upper figure shows that there might be a correlation between `time` and `clo`, as larger observations of `clo` are fond for larger observations of `time`. As there are a number of male observations which are only observed for a short time, it is however difficult to see whether the sex-differences could explain the trend. The two bottom figures shows that there is no obvious connection between `tOut`, `tInOp` and `clo`.

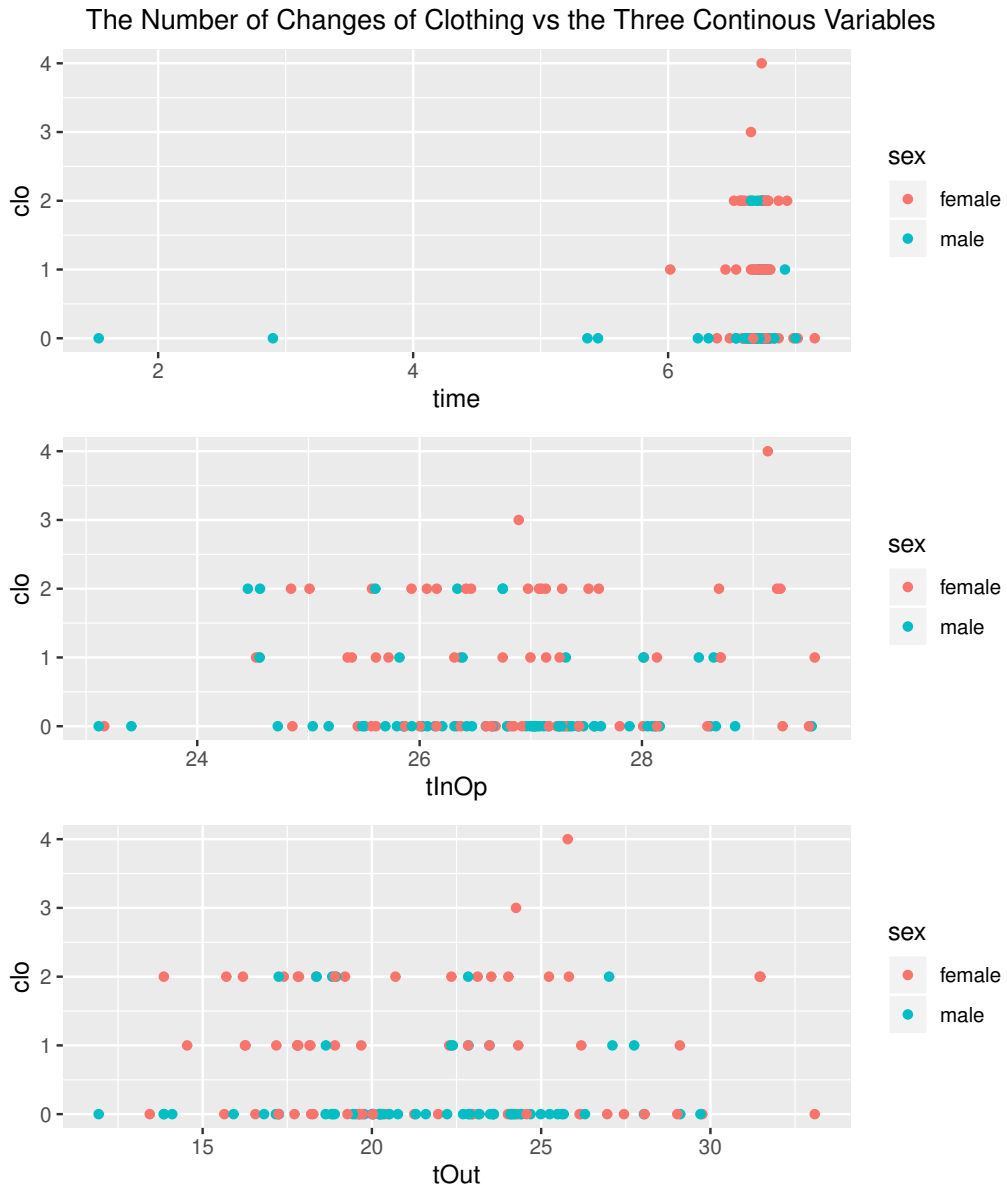

Figure 12: *`clo` against the three continuous variables. The top figure shows `clo` against `time`, the middle figure shows `clo` against `tInOp` and the bottom figure shows `clo` against `tOut`. In all figures the sex is marked with the color of the point.*

Binomial Model

First a generalized linear model based on the Binomial distribution is developed. To create this model the number of clothes changes, `clo`, should be viewed as the number of successes in a number of trial. It is obvious to use the number of observations, `nobs`, as the number of trials. The link-function of the model is chosen to be the canonical link function, which is the logit function. The linear part of the model is assumed to be additive. As the figure 12 indicated that there might be an interaction between `time` and `sex`, the interaction between these to variables are included as well. The linear part of the initial model therefore becomes

$$\log\left(\frac{\mu}{1-\mu}\right) = \eta = \beta_0 + \beta_1(\texttt{sex}) + \beta_2(\texttt{sex}) \cdot \texttt{time} + \beta_3 \cdot \texttt{tInOp} + \beta_4 \cdot \texttt{tOut} \tag{19}$$

The model is reduced by performing successive χ^2-tests. Table 11 shows the tests associated with removing one parameter from (19). The table shows that `tOut` should be removed from the model.

Parameter Removed	Df	Deviance	AIC	LRT	Pr($>\chi^2$)
None		168.66	278.19		
`time`	1	169.29	276.82	0.6324	0.42646
`sex`	1	172.45	279.98	3.7897	0.05157
`tInOp`	1	168.88	276.41	0.2187	0.64007
`tOut`	1	168.79	276.32	0.1295	0.71898
`time` × `sex`	1	172.03	279.56	3.3666	0.06653

Table 11: *The result of χ^2 tests to see if the model in* (19) *should be reduced*

The reduction of the model is continued, by performing successive χ^2-tests. By removing all insignificant effects, the final model becomes

$$\log\left(\frac{\mu}{1-\mu}\right) = \eta = \beta_0 + \beta_1(\texttt{sex}). \tag{20}$$

The residual deviance of the model in (20) is found to be 172.57 on 134 degrees of freedom. The residual deviance should approximately follow a χ^2-distribution on $n - k = 134$ degrees of freedom. The p-value associated with the sufficiency of the model thus becomes

$$P(\chi^2(134) \geq 172.57) = 0.0138.$$

Setting the significance level to be $\alpha = 0.05$, the model is therefore not sufficient.

A way to overcome this problem, is to include an overdispersion term in the model. This does not affect the parameters, $\boldsymbol{\beta}$, but it affects the confidence intervals of them. A model with linear part corresponding to the initial model given in (19) is now fit including over-dispersion. The model is reduced in the same way as the model without overdispersion - the only difference being that the test is an F-test instead of a χ^2-test. Note that the test is not exact. The result of the F-tests associated with the initial model is given in table 12. As for the binomial model without over-dispersion `tOut` should be removed.

Parameter Removed	Df	Deviance	F-value	Pr(>F)
None		168.66		
`time`	1	169.29	0.4875	0.48630
`sex`	1	172.45	2.9211	0.08982
`tInOp`	1	168.88	0.1685	0.68209
`tOut`	1	168.79	0.0998	0.75259
`time` × `sex`	1	172.03	2.5949	0.10963

Table 12: *The F-tests associated with the initial Binomial model including overdispersion*

The reduction of the model leads to the same linear structure as the one for the binomial model without over-dispersion given in (20). It is repeated below for reference

$$\log\left(\frac{\mu}{1-\mu}\right) = \eta = \beta_0 + \beta_1(\texttt{sex}). \tag{21}$$

The model with over dispersion cannot be tested for sufficiency. The assumptions on the residuals can however be checked through a diagnostics plot. One should however note that since `nobs` is a very low number (not larger than 6), the diagnostics plots are not expected to be very good. The fact that `sex` is the only explanatory variable, also means that there only are two different predicted variables, which makes some of the plots difficult to read. Figure 13 shows the diagnostics plots. The two plots in the left pane are very difficult to read, due to there only being two different predictions. The QQ-plot in the upper right corner, shows large issues with the normal-distribution of the deviance residuals. This is most likely due to the low values of `nobs`.

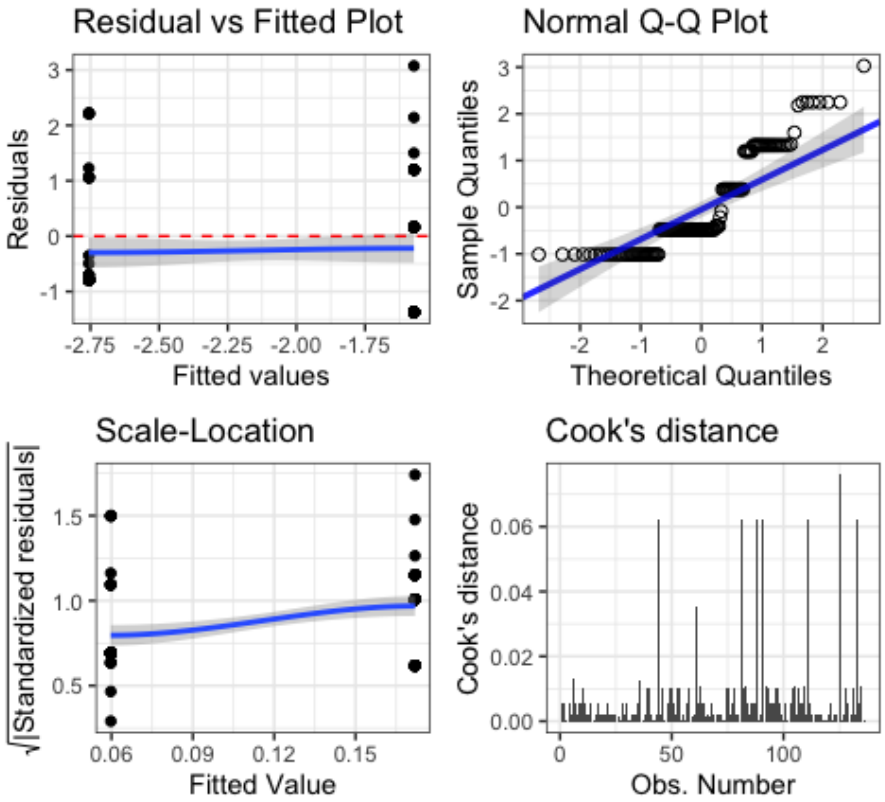

Figure 13: *Diagnostics plot for the Binomial model with over-dispersion and linear structure given in* (21).

The estimates of the parameters are given in table 13. Notice that due to the parameterisation the common intercept is left out.

Parameter	**Estimate**	**Std. Error**	**t value**	**Pr(>|t|)**
$\hat{\beta}_1(female)$	-1.5721	0.1645	-9.557	<2e-16
$\hat{\beta}_1(male)$	-2.7560	0.2743	-10.047	<2e-16

Table 13: *The parameters for the final binomial model with over-dispersion included. The linear part is given in* (21).

The dispersion parameter is estimated to be $\hat{\sigma^2} = 1.34$. From the parameters the probability of changing clothes at an observation, μ, can be calculated, by using the inverse of the logit-transform. Thus

$$\hat{\mu}_{female} = \frac{\exp(\hat{\beta}_1(female))}{1+\exp(\hat{\beta}_1(female))} = \frac{\exp(-1.5721)}{1+\exp(-1.5721)} = 0.17,$$
$$\hat{\mu}_{male} = \frac{\exp(\hat{\beta}_1(male))}{1+\exp(\hat{\beta}_1(male))} = \frac{\exp(-2.7560)}{1+\exp(-2.7560)} = 0.060.$$

The confidence intervals of the predictions are found by finding the confidence intervals in the link domain, and applying the inverse logit-transformation. Thus

$$g\left(\hat{\mu_i} \pm u_{1-\alpha/2}\sqrt{\hat{\sigma^2}\hat{\sigma}_{ii}}\right)^{-1}, \tag{22}$$

where g denotes the logit transformation. $\sqrt{\hat{\sigma_{ii}}}$ is found to be 0.16 and 0.27 for women and men respectively. $\hat{\sigma}$

denotes the over-dispersion parameter. The results are the following confidence intervals

$$\begin{aligned} \text{female: } & [0.13; 0.23] \\ \text{male: } & [0.03; 0.11]. \end{aligned}$$

Prediction intervals for the two probabilities are found by bootstrapping. First a new-data-set is simulated. As we are only interested in a prediction interval for males and females, there are only two observations in the data-set. The number of observations are set to 6, as this is the most common choice in the original data-set. The model is fitted to the data-set, and new predictions are made. By repeating this procedure $k = 1000$ times boot-strapping confidence interval is found by extracting the relevant quantiles. The 95% confidence interval is simulated to be

$$\begin{aligned} \text{female: } & [0; 0.667] \\ \text{male: } & [0; 0.50]. \end{aligned}$$

Poisson Model

A generalised linear model based on the Poisson model is now developed. Let μ denote the predicted value of `clo`. The link-function is chosen to be the log-function, which is also the canonical link-function. It seems reasonable that the number of clothes changes should scale with the total time of measurement. That is

$$\log\left(\frac{\mu}{\texttt{time}}\right) = \mathbf{X}\boldsymbol{\beta}$$

This corresponds to including `time` as an offset in the model. The linear part of the model is assumed to be additive without any interactions. Thus the linear part of the initial model becomes

$$\log(\mu) = \eta = \log(\texttt{time}) + \beta_0 + \beta_1 \cdot \texttt{tInOp} + \beta_2 \cdot \texttt{tOut} + \beta_3(\texttt{sex}) + \beta_4 \cdot \texttt{nobs} \tag{23}$$

The model is reduced by performing successive χ^2-test. Table 14 shows the results of the first round of tests. A type II partioning has been used. From the table it is seen that `nobs` should be removed from the model.

Parameter Removed	Df	Deviance	AIC	LRT	Pr(>Chi)
None		148.97	275.96		
`sex`	1	167.60	292.59	18.6306	1.587e-05
`tOut`	1	149.01	274.00	0.0366	0.8482
`tInOp`	1	149.14	274.12	0.1625	0.6869
`nobs`	1	148.97	273.96	0.0003	0.9869

Table 14: *The first round of χ^2-tests. A type II partitioning is used.*

The reduction is continued by carrying out succesive tests. The linear part of the final model becomes

$$\log(\mu) = \log(\texttt{time}) + \beta_0 + \beta_1(\texttt{sex}). \tag{24}$$

The residual deviance is found to be 149.15 on 134 degrees of freedom. This can be used to test for model sufficiency. The residual deviance should approximately follow a χ^2 distribution with $n - k$ degrees of freedom. n is the number of observations, k and is the dimension of the tested model. The p-value for the tested model thus becomes

$$P(\chi^2(134) \geq 149.15) = 0.17,$$

which means that the model is accepted. Figure .. shows a diagnostics plot of the model. As for the binomial model, the plot is not expected to be very good, as the counts are very low. The figure shows issues with both left-pane plots, as the residuals seems to change against the fitted values. The QQ-plot also reveals that the deviance residuals does not follow a normal distribution.

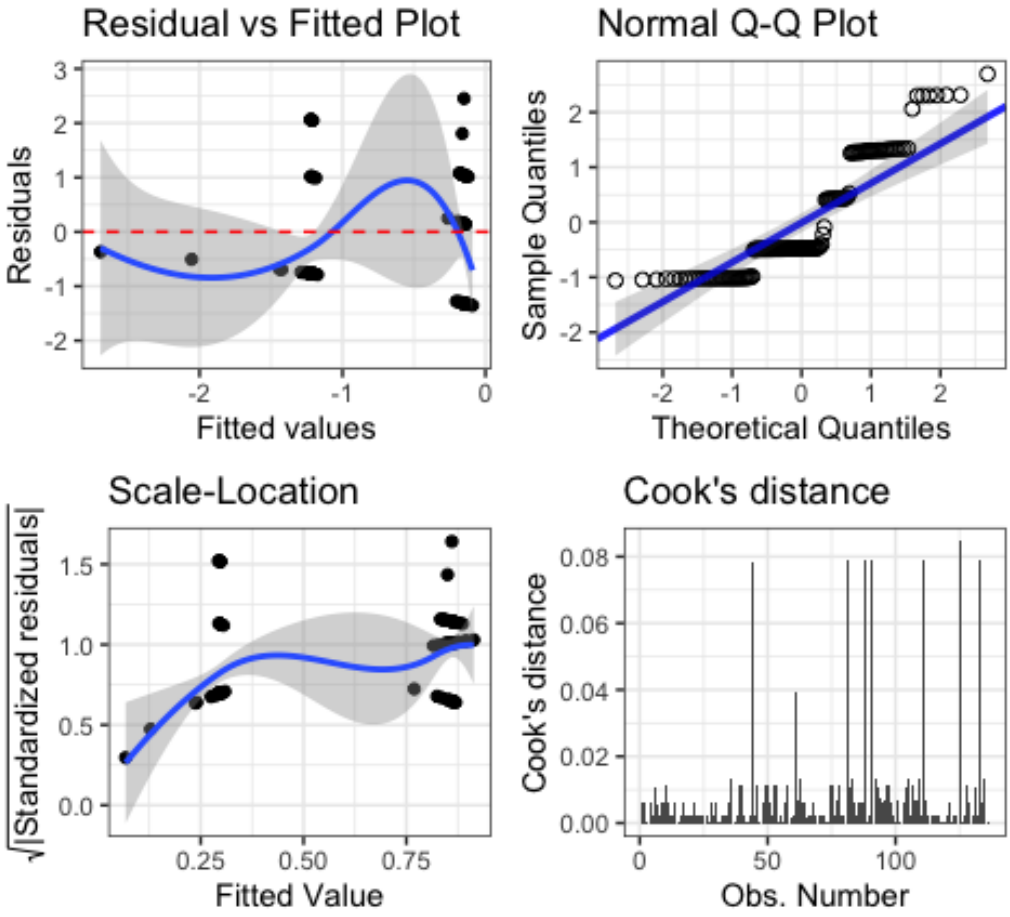

Figure 14: *Diagnostics plot of the final Poisson model with the linear structure given in* (24)

Table 15 shows the parameters of the model.

Parameter	**Estimate**	**Std. Error**	**t value**	**Pr(>|t|)**
$\beta_1(female)$	-2.0573	0.1291	-15.94	<2e-16
$\beta_1(male)$	-3.1205	0.2294	-13.60	<2e-16

Table 15: *The parameters for the final Poisson model. The linear part is given in* (24)

Setting `time` $= 1$ the hourly clothes changing rate, μ can be found for each sex by using the inverse of the log-transformation - that is by taking the exponential.

$$\mu_{female} = \exp(\log(1) + \beta_1(female) = 1 + \exp(-2.0573) = 0.13$$
$$\mu_{male} = \exp(\log(1) + \beta_1(male) = 1 + \exp(-3.1205) = 0.04$$

Taking the time into account figure 15 shows the predicted values using the Poisson distribution. The shaded region represents the confidence intervals, and the dotted lines represent the prediction intervals. The confidence intervals are found using the same method as for the binomial model - see equation (22). The prediction intervals are found by bootstrapping. First a new data-set is simulated using the parameters estimated in the fit. Hereafter the model is fitted to the generated data-set, and new predictions are made. This procedure is repeated $k = 10000$ times, and the prediction intervals are found by taking the 2.5% and 97.5% quantile. As the data is simulated there are some unexptected dives in the data.

Even though only `sex` is a part of the linear structure, the predicted values of `clo` increase with time. This is due to the offset.

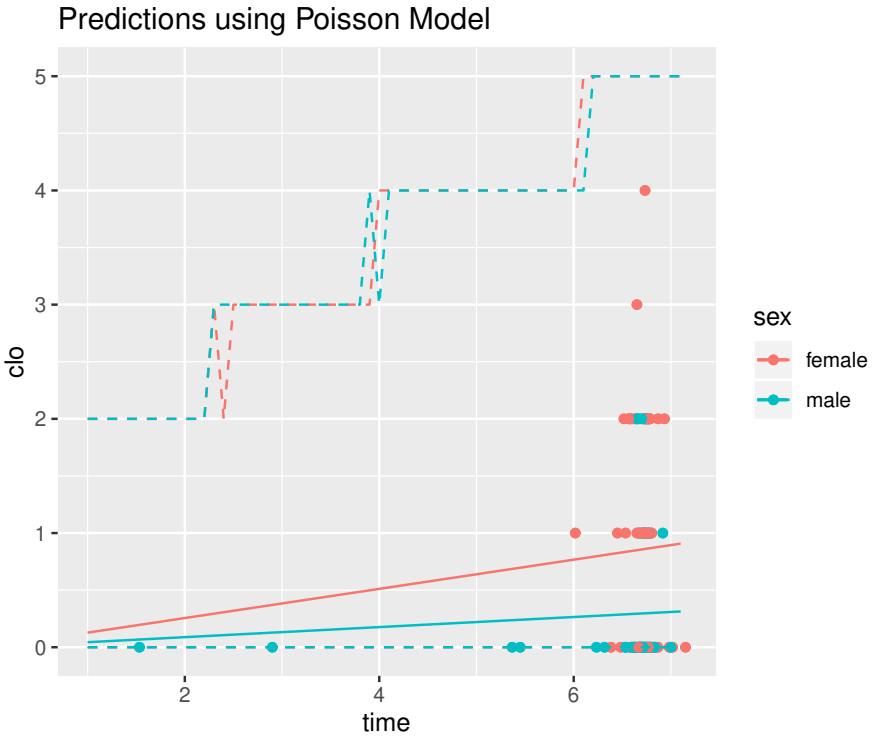

Figure 15: *Predictions along with a confidence interval. The dots represent the observations.*

Discussion of the two Models

The binomial model and the poisson model leads to two different interpretations of the data. In the binomial model the number of clothes changes, `clo`, is seen in relation to the number of observations, `nobs`. That is each observation is viewed as a trial where the subject can either change clothes or not. The number of successes is `clo` and the number of failures is `nobs-clo`. The output of the model will therefore be a probability, which should be multiplied with `nobs` in order to obtain `clo`.

The poisson model on the other hand considers `clo` as count-data, and does not directly link it to `nobs`. Instead `time` is used to scale the rate, as it is assumed that the rate is constant for a given time-interval.

Both models have meaningful interpretations, but considering the analysis in the previous two sections implies that the poisson model is the most appropriate.

Including SubjId instead of Sex

Both the binomial model and the Poisson model are now fit using `subjId` instead of `sex`. The linear structures for each of the models are chosen to be

$$\text{Binomial Model: } \log\left(\frac{\mu}{1-\mu}\right) = \eta = \beta_0 + \beta_1(\texttt{subjId}) + \beta_2 \cdot \texttt{time} + \beta_3 \cdot \texttt{tInOp} + \beta_4 \cdot \texttt{tOut} \tag{25}$$

$$\text{Poisson Model: } \log(\mu) = \eta = \log(\texttt{time}) + \beta_0 + \beta_1 \cdot \texttt{tInOp} + \beta_2 \cdot \texttt{tOut} + \beta_3(\texttt{subjId}) + \beta_4 \cdot \texttt{nobs}. \tag{26}$$

Notice that the interaction between `subjId` and `time` are not included in the initial binomial model, as this would use a lot of degrees of freedom. The models are reduced by performing successive likelihood-ratio tests.

The final linear structures become

$$\text{Binomial Model: } \log\left(\frac{\mu}{1-\mu}\right) = \beta_0 + \beta_1(\texttt{subjId})$$
$$\text{Poisson Model: } \log(\mu) = \eta = \log(\texttt{time}) + \beta_0 + \beta_1(\texttt{subjId}).$$

The residual deviances are found to be 75.41 on 89 degrees of freedom and 63.48 on 89 degrees of freedom for the binomial model and the Poisson model respectively. The p-values to test for sufficiency is found by comparing with the χ^2-distribution with 89 degrees of freedom in both cases. The p-values become

$$\text{Binomial Model: } P(\chi^2(89) \geq 75.41) = 0.84$$
$$\text{Poisson Model: } P(\chi^2(89) \geq 63.48) = 0.98,$$

meaning that both models are accepted. Figure 16 shows a boxplot of the estimated values for each model in the response domain. It does not makes sense to compare the distributions of the two parameters, as `nobs` influences the predictions in the Binomial case, while `time` influences the predictions in the Poisson case.

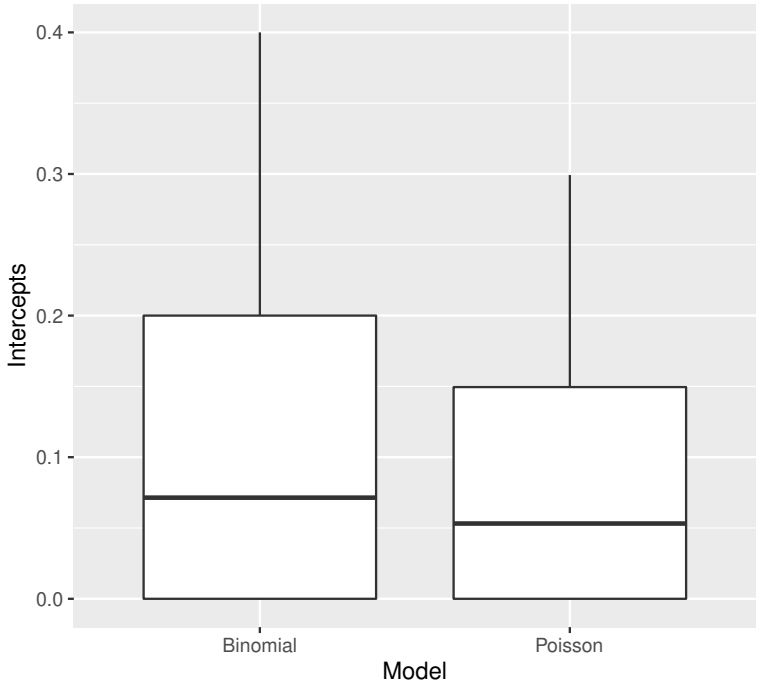

Figure 16: *Boxplot showing the intercepts of each* `subjId` *for the two models*

Figure 17 shows boxplots of the predictions of the two models. It is seen that there is no great difference in the distribution of predictions.

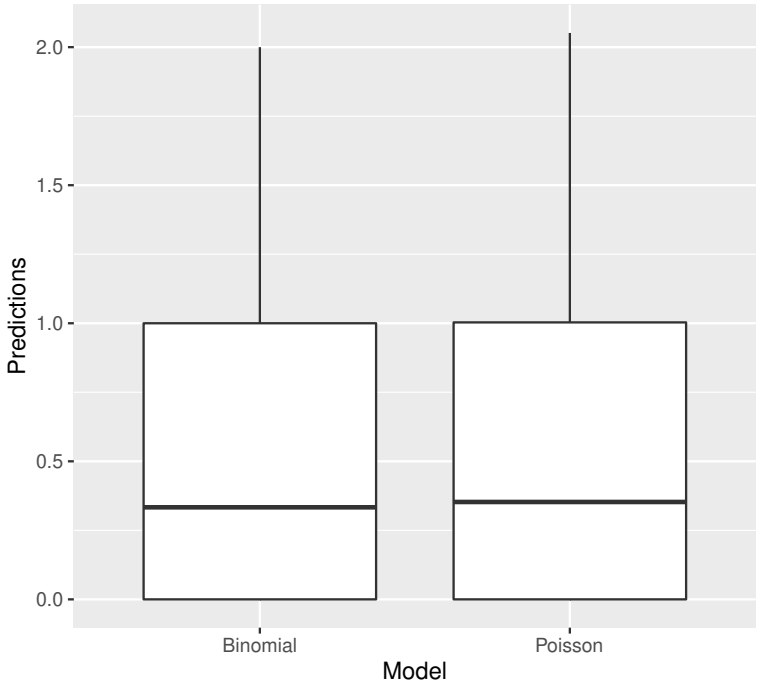

Figure 17: *Boxplots of the predictions for each model*

The standard errors of each estimated intercept depends heavily on the value of the intercept as figure 18 illustrates. The figure shows the estimated intercept in the link domain vs the logarithm of the standard error. It is seen that low estimates leads to higher standard errors. This is due to the link function - for low values, the inverse link-function is very flat, meaning that a small change in the link-domain does not lead to a large change in the response domain, and thus high standard errors.

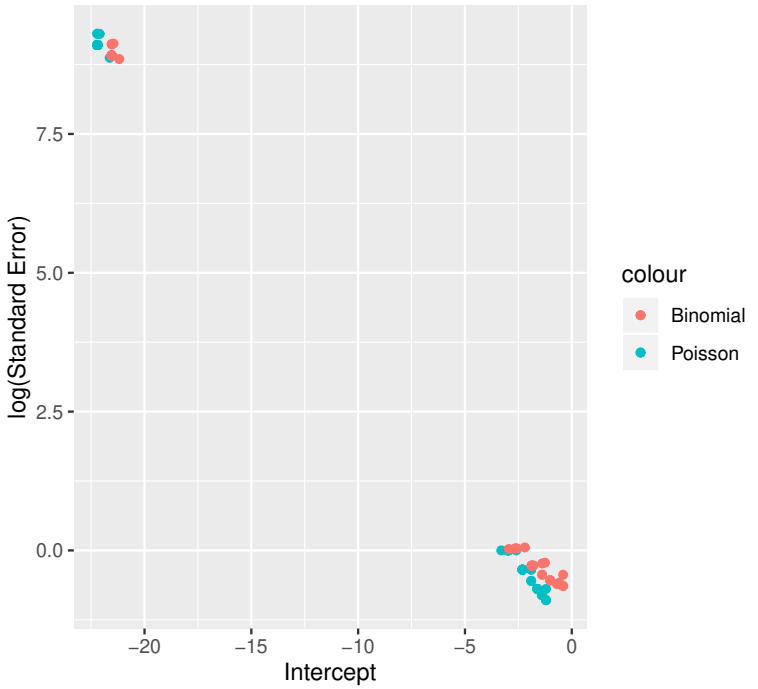

Figure 18: *The estimated intercepts vs. the logarithm of the standard error. A low estimate leads to a high standard error*

Conclusion on Clothing Data

In this part of the assignment the number of clothes changes through a working day has been investigated. Both a Binomial model and a Poisson model has been considered for the data. First models with the `sex` of subject, but not the `subjId`, were considered. Here it was found that the standard binomial model, did not account for all variance in data, and therefore over-dispersion were included. The Poisson model was perhaps more suited, as no over-dispersion needed to be included in this model. Models with `subjId` instead of `sex` were herafter fitted to the data. These models performed better, having much higher p-values in the test for sufficiency.

Appendix

Summary of models

The gamma model with a log link.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.4897	0.1382	3.54	0.0005
Temp	0.0275	0.0018	15.27	0.0000
InvHt	-0.0001	0.0000	-8.28	0.0000
Hum	0.0069	0.0011	6.00	0.0000

The gamma model with a inverse link.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.2433	0.0138	17.61	0.0000
Temp	-0.0018	0.0002	-12.14	0.0000
InvHt	0.0000	0.0000	7.99	0.0000
Hum	-0.0009	0.0001	-6.87	0.0000

The inverse gaussian with a inverse link.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.2403	0.0210	11.47	0.0000
Temp	-0.0022	0.0002	-11.00	0.0000
InvHt	0.0000	0.0000	7.48	0.0000
Pres	-0.0002	0.0001	-2.11	0.0353
Vis	0.0001	0.0000	2.12	0.0351
Hum	-0.0005	0.0002	-2.82	0.0050

The inverse gaussian with a log link.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.8575	0.1397	6.14	0.0000
Temp	0.0280	0.0020	14.27	0.0000
InvHt	-0.0001	0.0000	-9.20	0.0000
Pres	0.0029	0.0006	4.85	0.0000

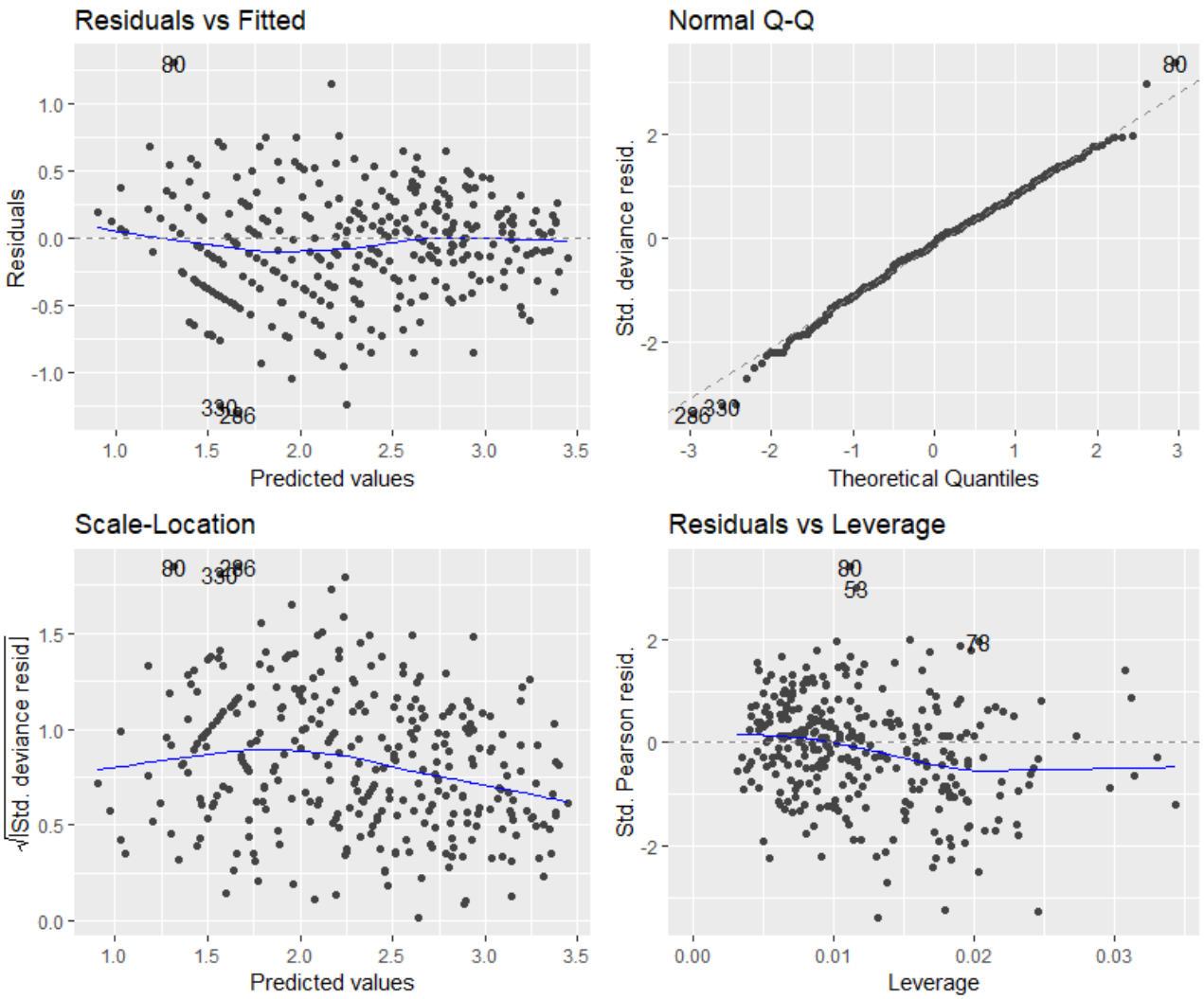

Figure 19: *The Gamma log link residuals*

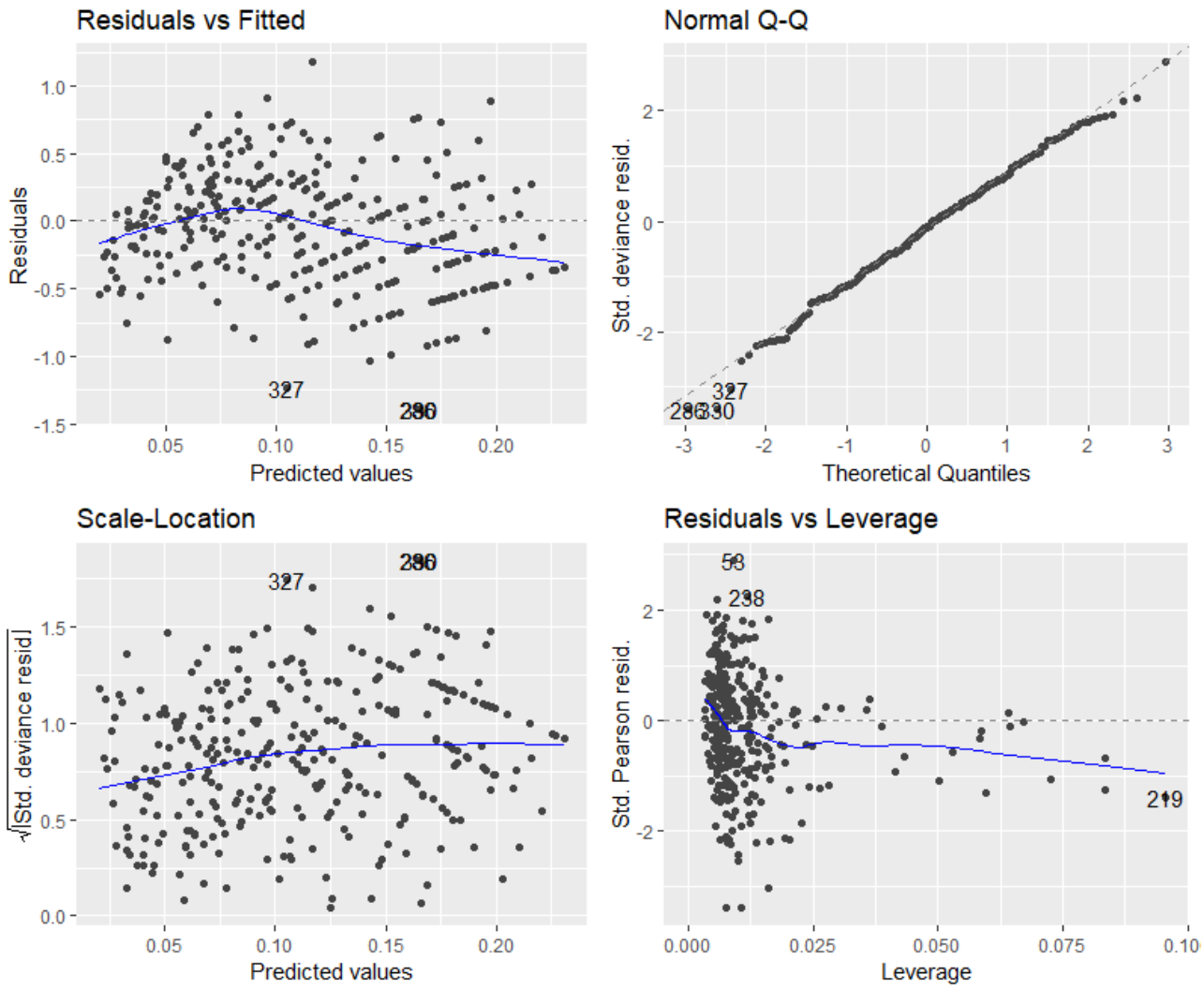

Figure 20: *The Gamma inverse link residuals*

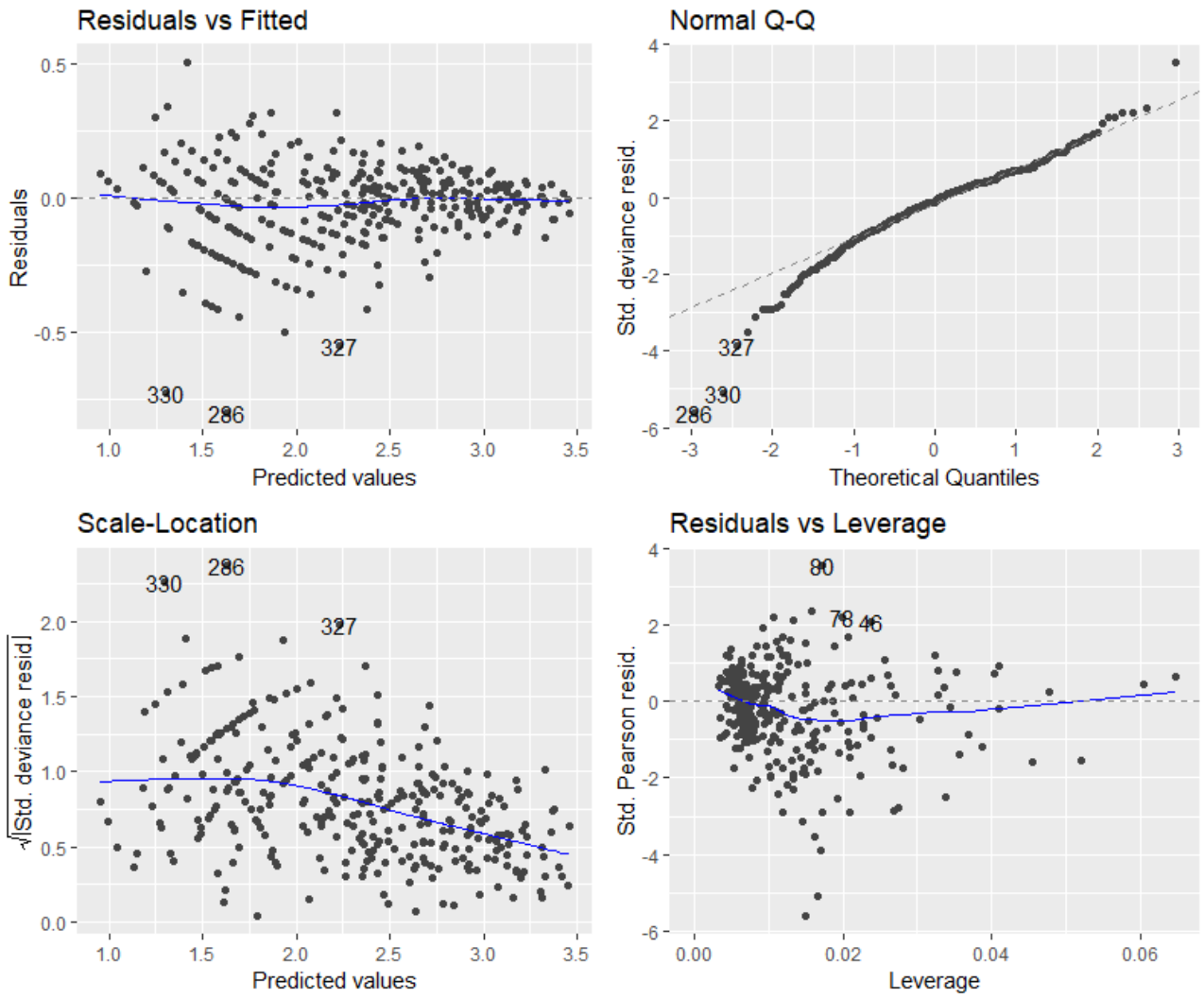

Figure 21: *The inverse gaussian log link residuals*

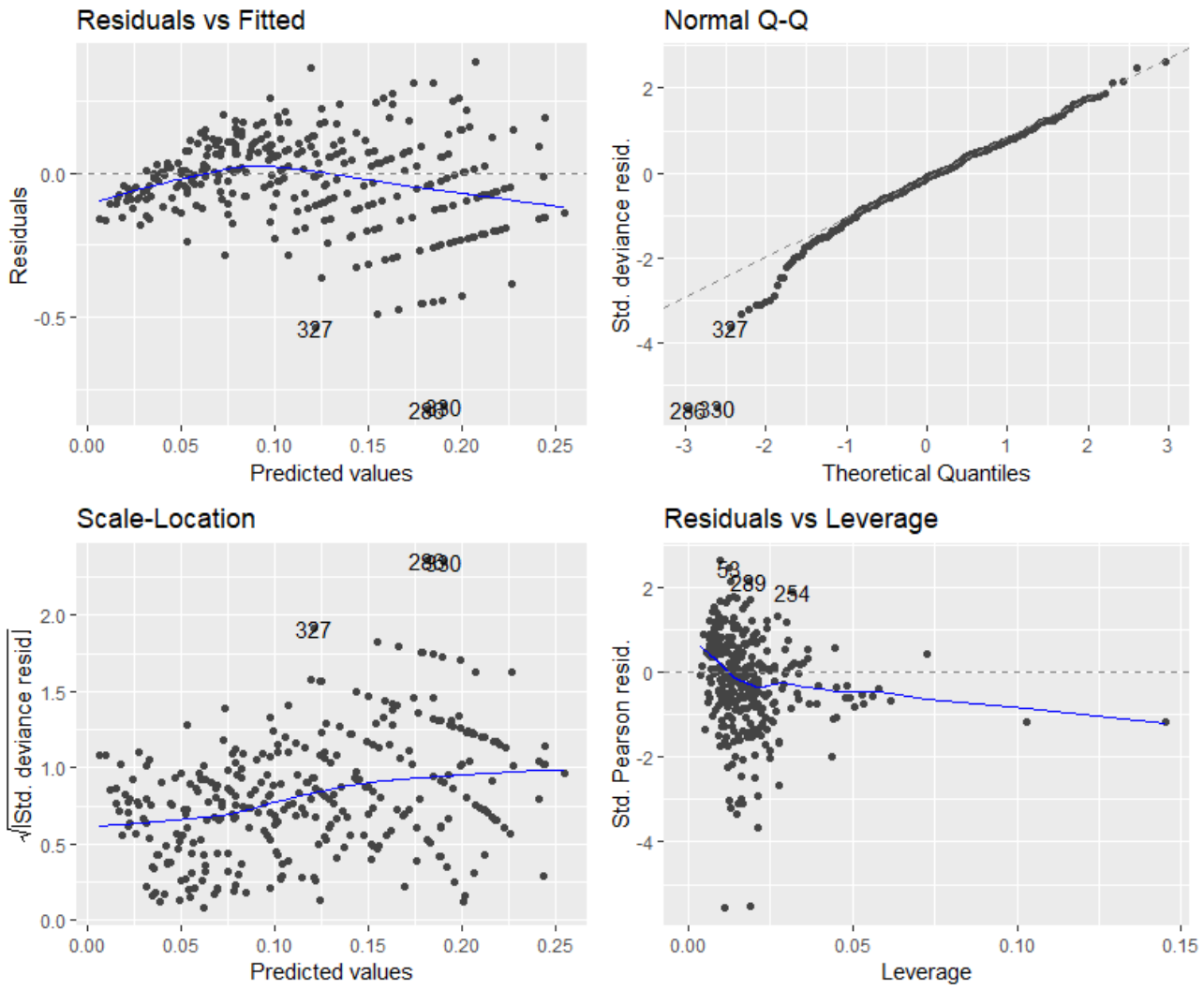

Figure 22: *The inverse gaussian inverse link residuals*

The final model

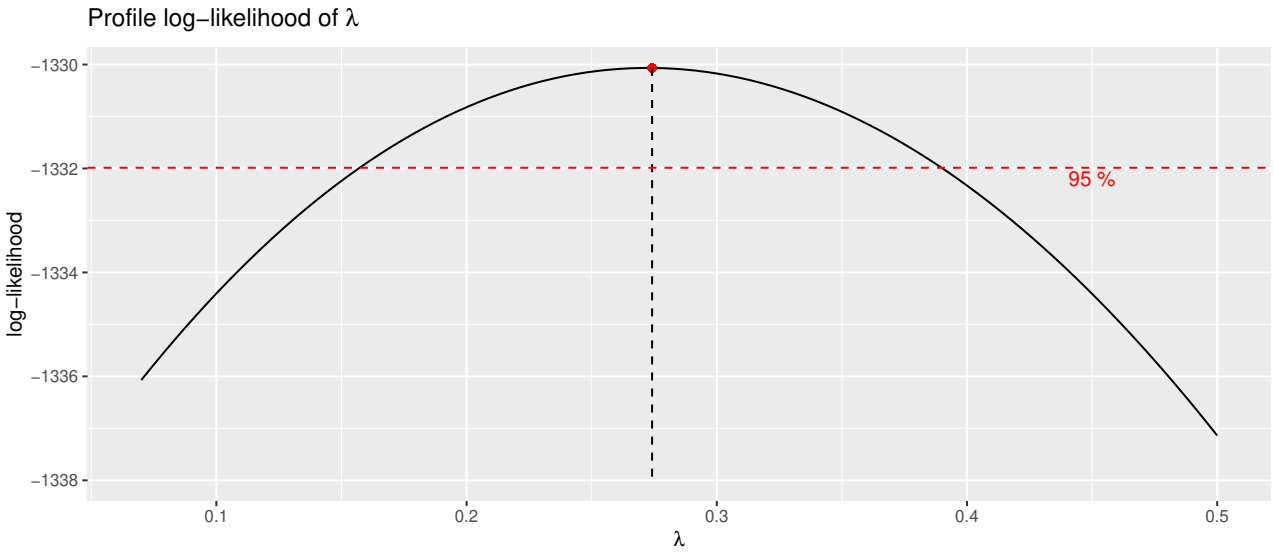

Figure 23: *Profile likelihood of λ given the initial final model, which is a classical GLM.*

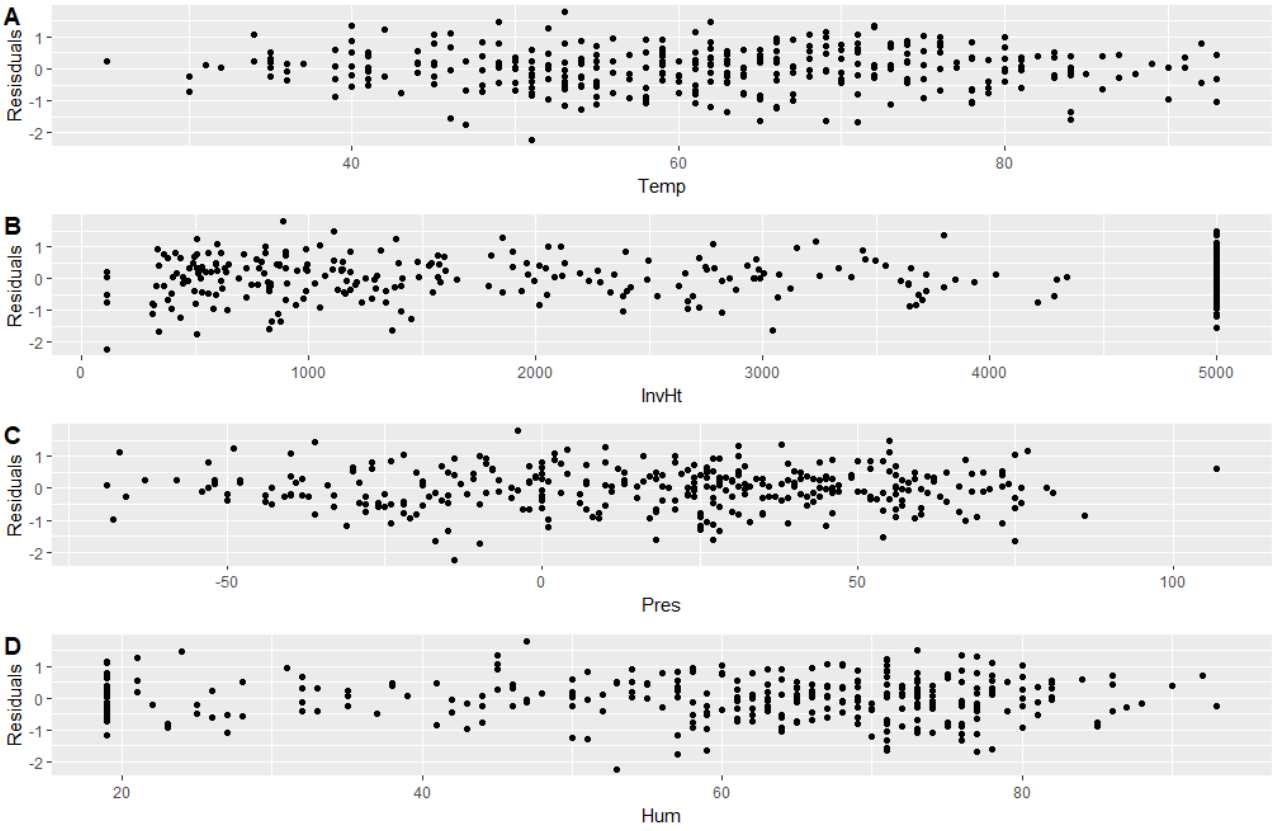

Figure 24: *Residuals of the predictors. One can observe that there is a cutoff for* `InvHt` *around 5000 and for* `Hum` *below 20, it is not clear why these cutoffs occur. Because of the integers in ozone one can see "lines" within the residuals. Besides that no clear heteroscedasticity is observed.*

1.	I(Temp^2)	9.	Hum:Wind
2.	Pres:Vis	10.	Pres:Wind
3.	Temp:Vis	11.	Pres:Hum
4.	Temp:Wind	12.	I(Vis^2)
5.	InvHt:Wind	13.	InvHt:Pres
6.	InvHt:Vis	14.	I(InvHt^2)
7.	Vis:Hum	15.	Temp:Pres
8.	Temp:InvHt		

Table 16: *Order in which parameters are removed from the model*

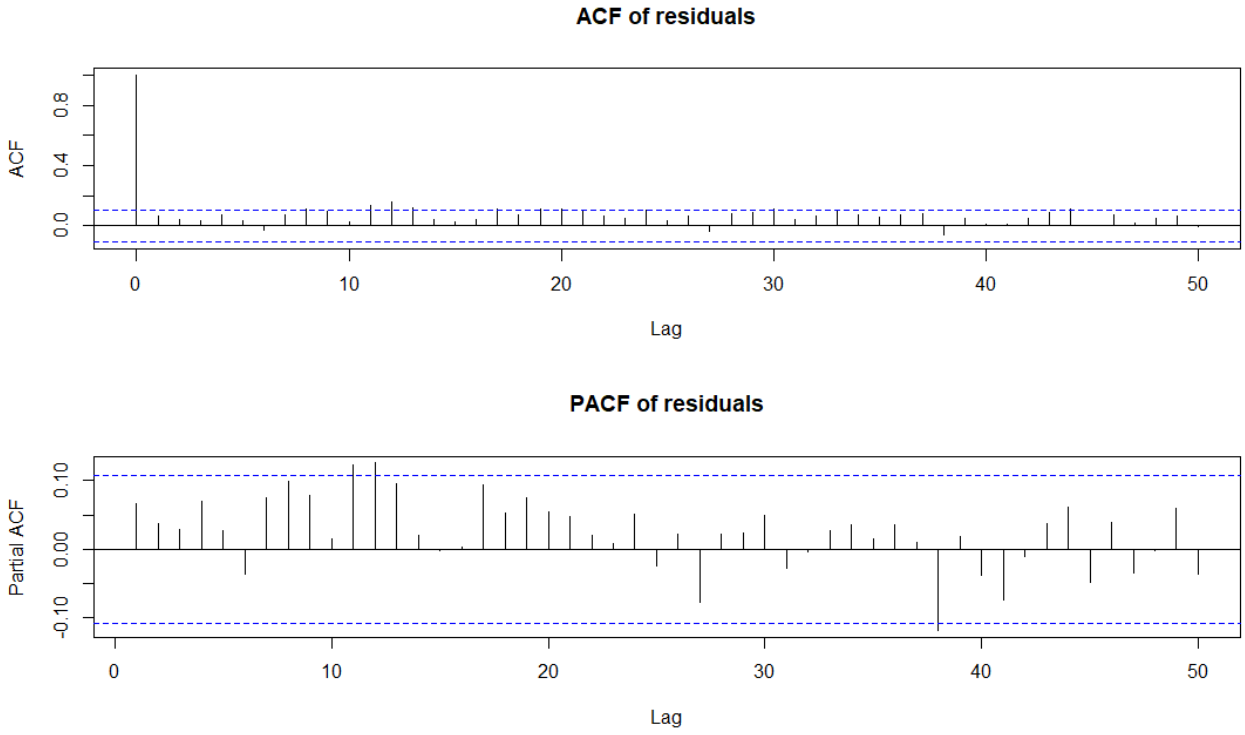

Figure 25: *PACF and ACF of the model with an AR(1) component*

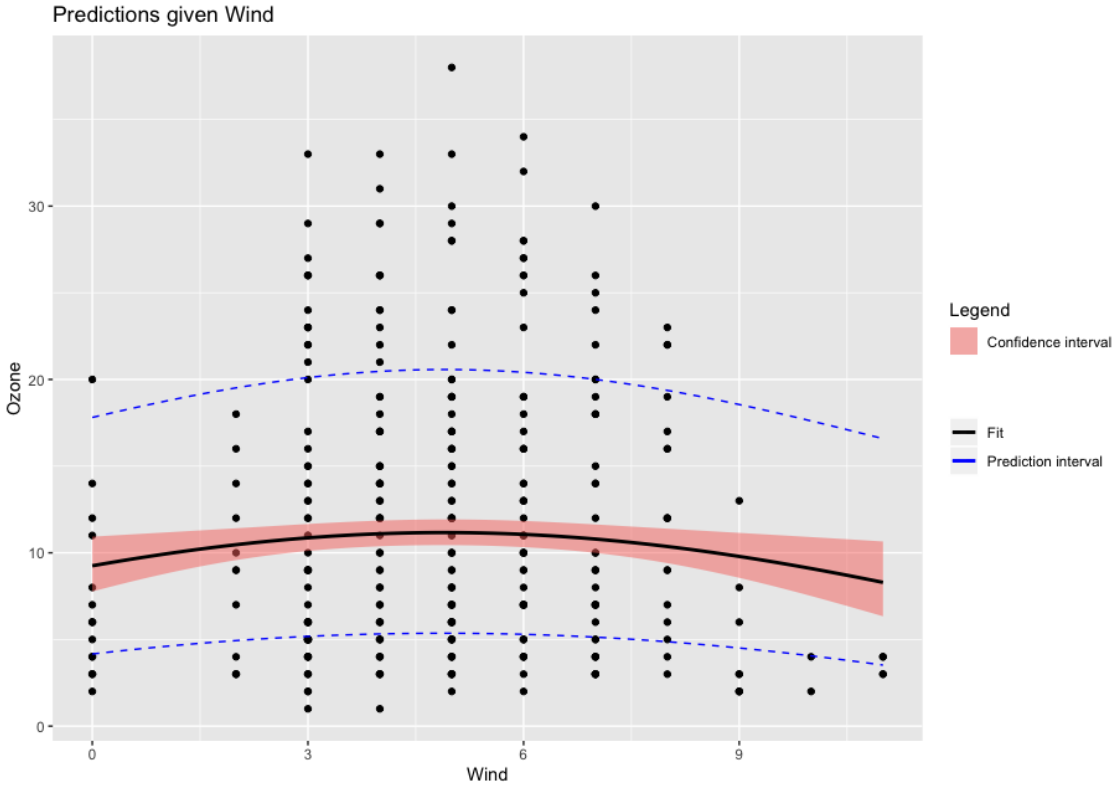

Figure 26: *The of the final model given wind*

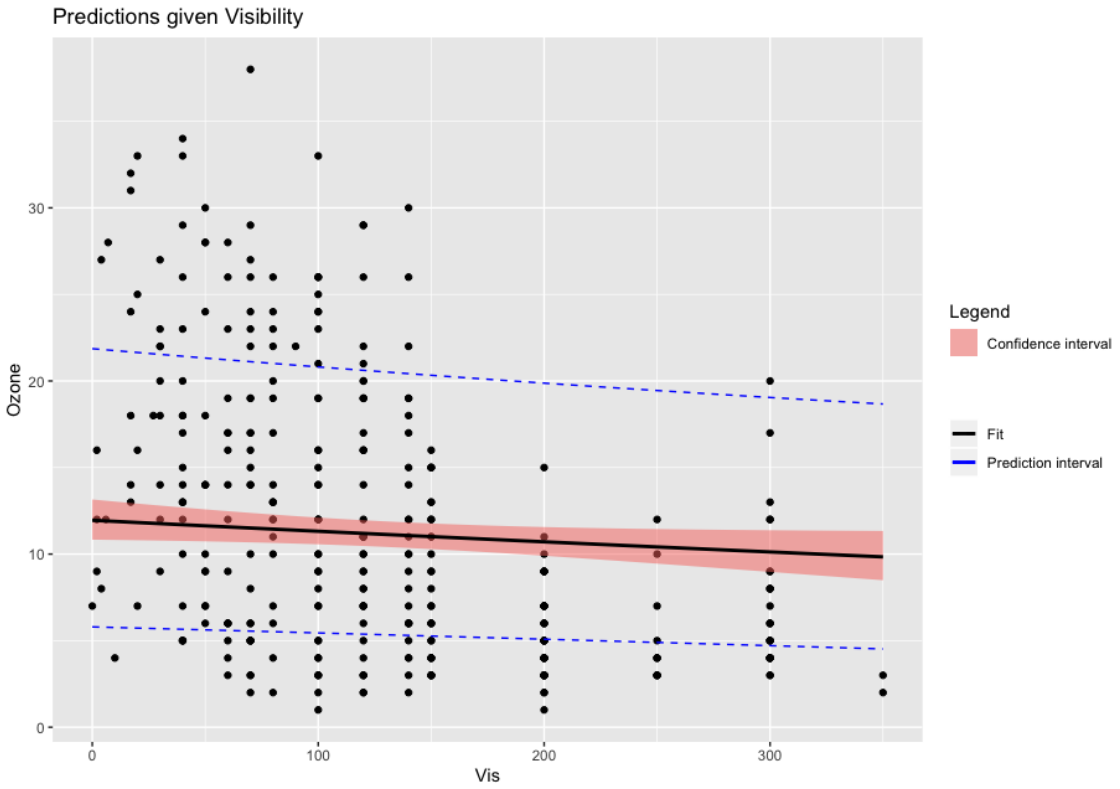

Figure 27: *The of the final model given visibility*

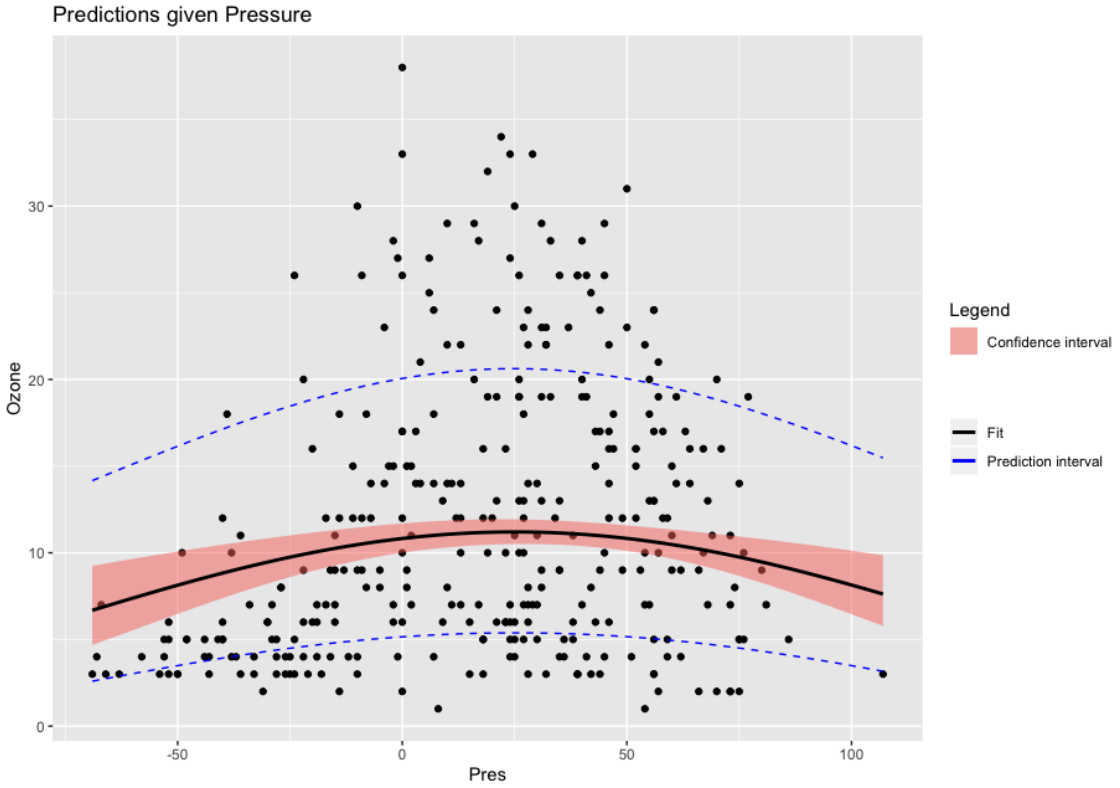

Figure 28: *The of the final model given pressure*

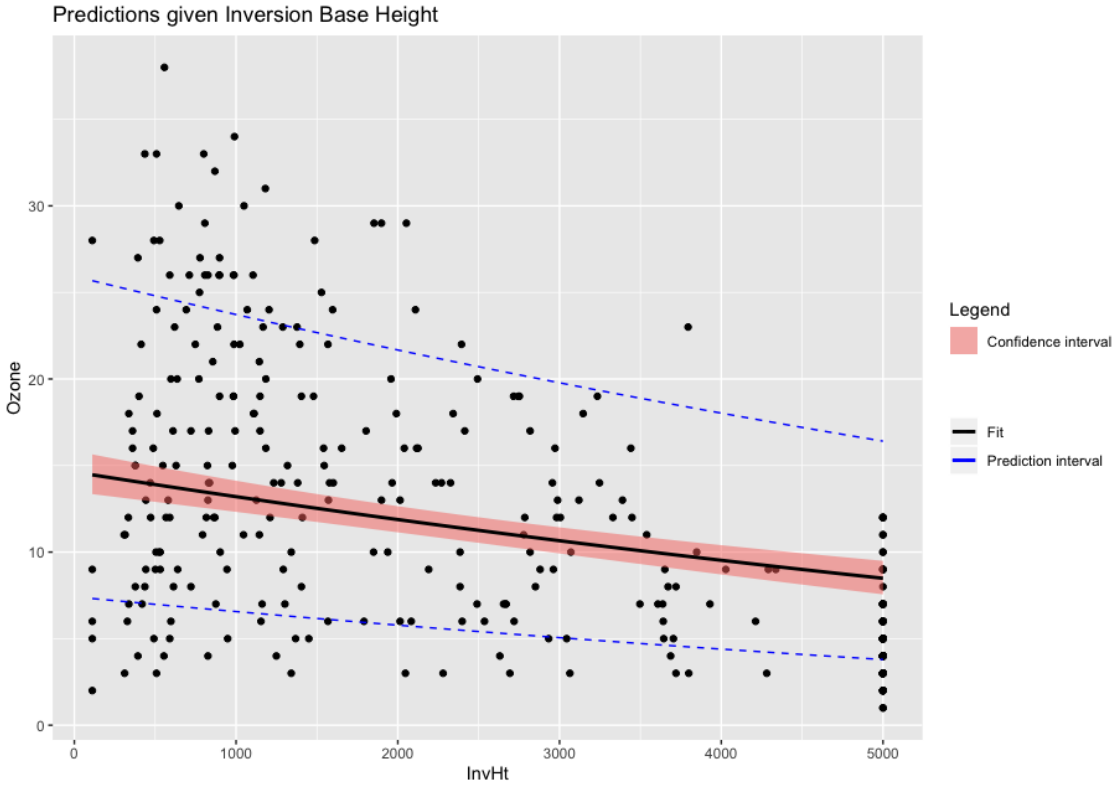

Figure 29: *The of the final model given inversion base height*

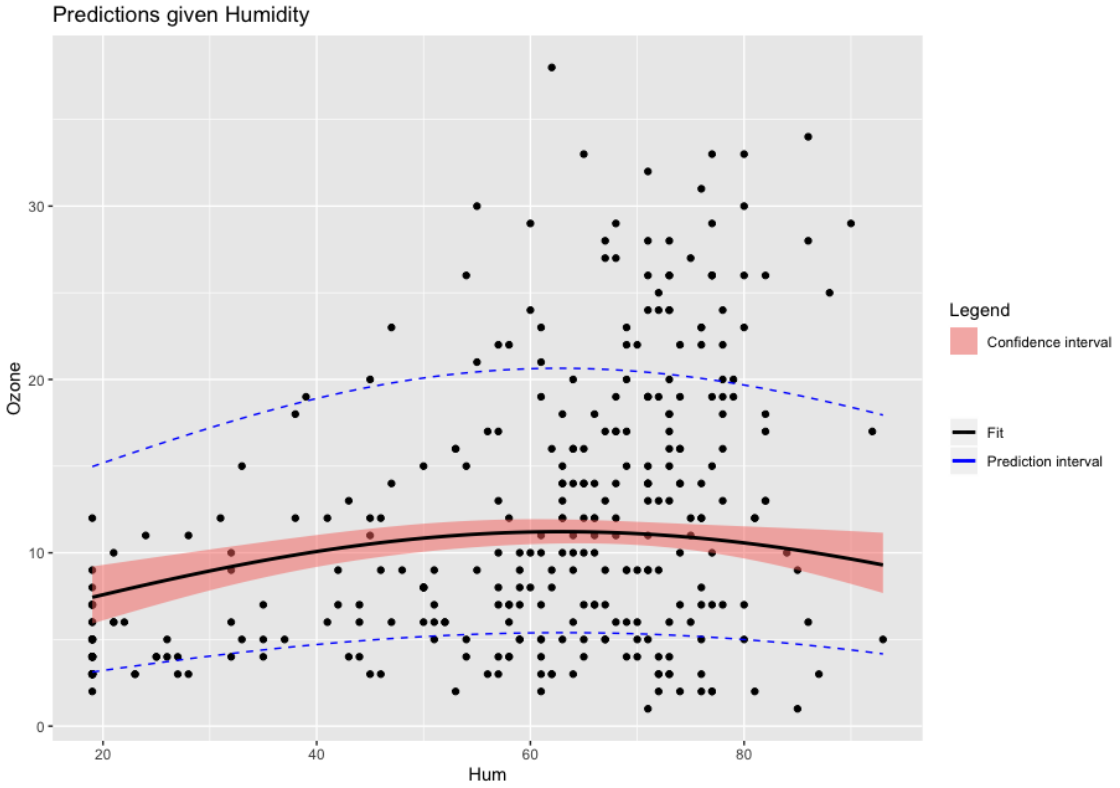

Figure 30: *The of the final model given humidity*